Transcriptomic Analysis of GSE152418 (COVID-19)

1.0 Introduction

The coronavirus disease 2019 (COVID-19), caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), triggered a global health crisis since its emergence in late 2019. While the majority of infected individuals experienced mild to moderate symptoms, a substantial proportion of people developed severe respiratory complications, systemic inflammation, or long-term sequelae, collectively known as post-acute sequelae of SARS-CoV-2 infection (PASC) or "Long COVID." The heterogeneous clinical manifestations of COVID-19 reflect a complex interplay between viral replication and host immune responses, necessitating the need for a deeper molecular understanding of how the host reacts to infection and recovery.

Transcriptomic profiling offers a powerful means of exploring host gene expression dynamics in response to viral infections. By analyzing RNA sequencing (RNA-seq) data, we can identify differentially expressed genes (DEGs) and disrupted signaling pathways, thereby elucidating the molecular signatures that distinguish infected individuals from those who have recovered. Such insights are critical for discovering biomarkers, informing therapeutic strategies, and improving disease classification.

The present study focuses on GSE152418, a publicly available RNA-seq dataset from the NCBI Gene Expression Omnibus (GEO). This dataset profiles whole blood transcriptomes of individuals with active COVID-19 infection and convalescent individuals who have recovered. Using a rigorous bioinformatics workflow, this analysis aims to:

- Identify genes that are significantly differentially expressed between COVID-19 and convalescent individuals.
- Perform functional enrichment analyses to uncover biological processes and signaling pathways associated with infection and immune recovery.
- Explore immune-specific transcriptional signatures using curated gene sets.

By leveraging robust statistical tools and visualization techniques, this study provides a comprehensive snapshot of the host transcriptomic landscape during COVID-19 and highlights key immunological processes that may underpin disease progression and resolution.

2. Materials and Methods

2.1 Dataset Description

The RNA-Seq dataset used in this study, GSE152418, was obtained from the NCBI Gene Expression Omnibus (GEO). It comprises whole blood transcriptomes collected from 34 human subjects, including 17 patients with active COVID-19 infection and 17 convalescent individuals who had recovered from the disease. The data were generated using high-throughput sequencing to profile global gene expression and characterize host immune responses at different disease stages.

2.2 Data Import and Preprocessing

The raw count matrix was downloaded in compressed format (GSE152418_p20047_Study1_RawCounts.txt.gz) and imported into R. Gene expression data were organized into a matrix where rows represent genes and columns represent samples. Sample identifiers were cleaned and matched with phenotype information.

A phenotype metadata table was created to define two experimental groups:

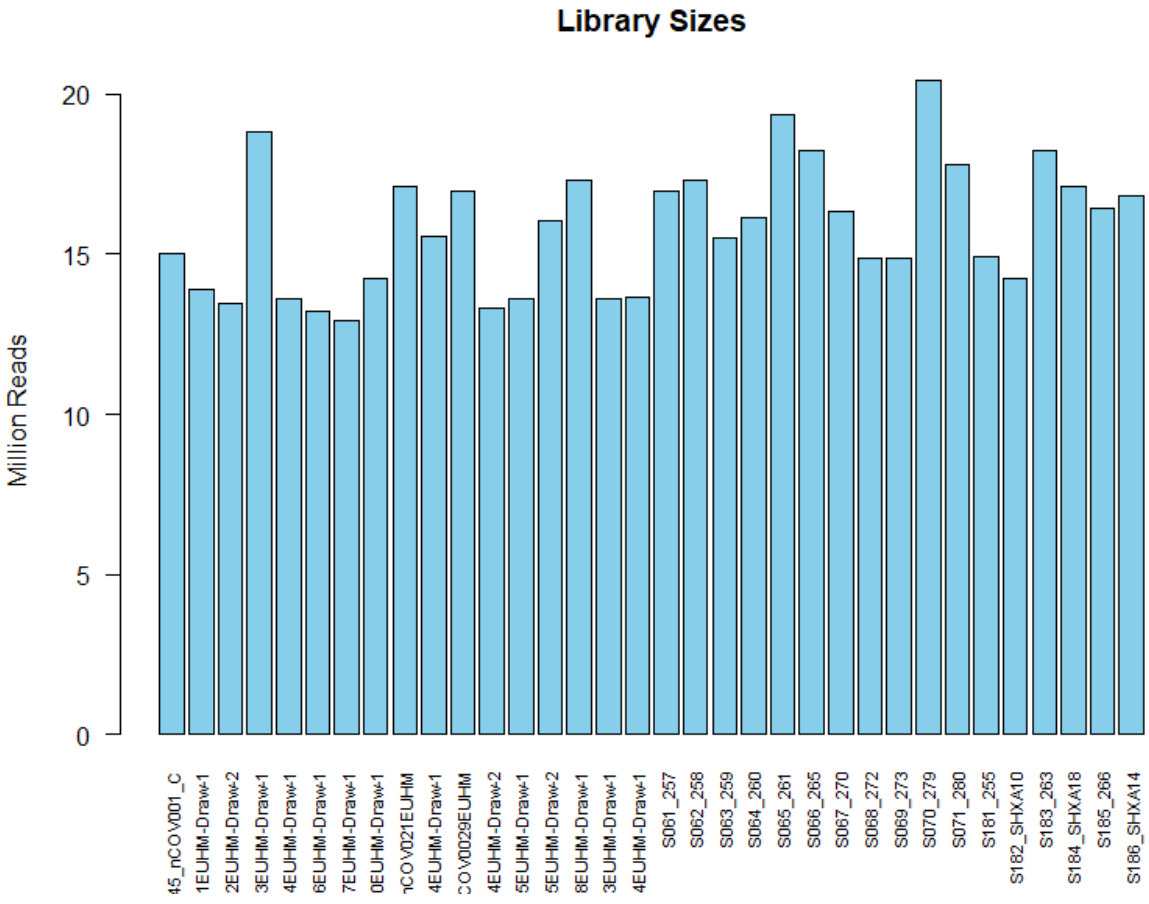
- COVID-19 (n = 17)
- Convalescent (n = 17)

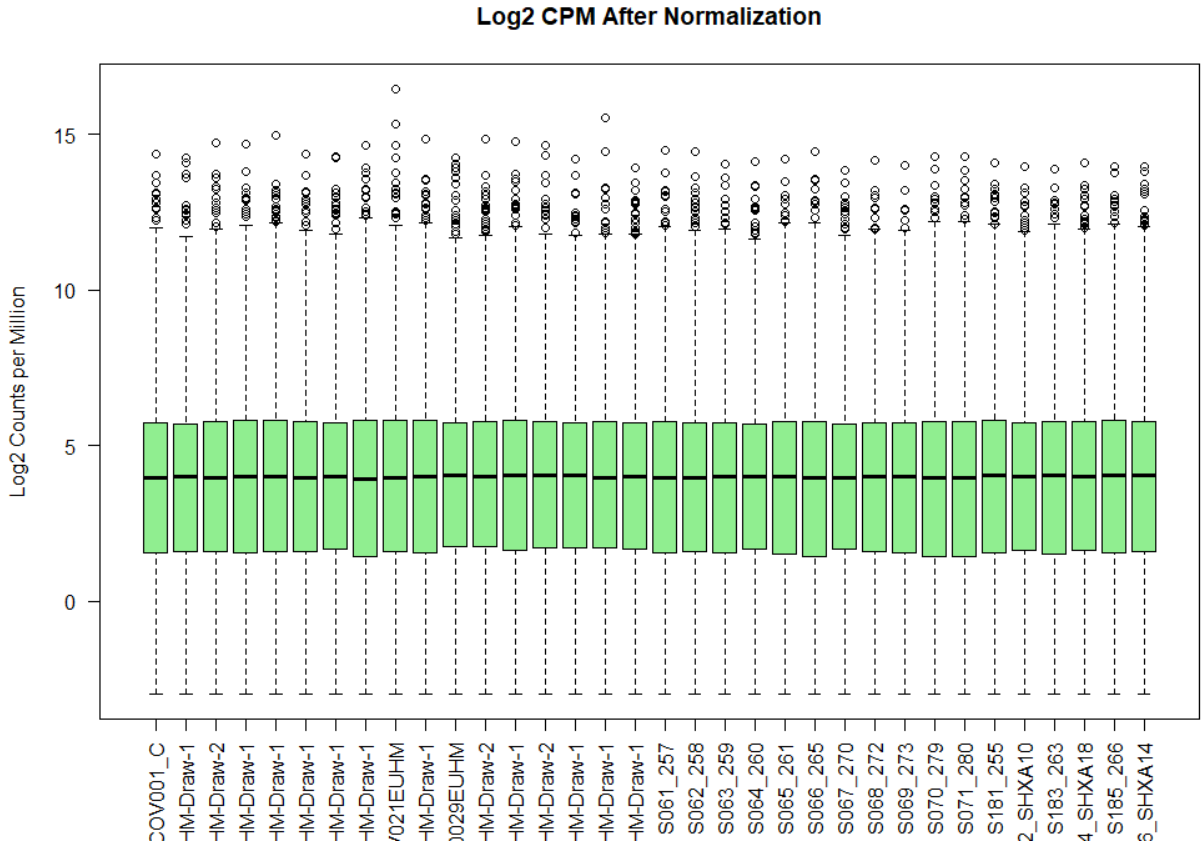
Lowly expressed genes were filtered out using the filterByExpr() function from the edgeR package to retain only informative transcripts for downstream analysis.

2.3 Quality Control and Normalization

To ensure accurate quantification, raw library sizes were examined using bar plots. The RNA-seq count data were then normalized using the Trimmed Mean of M-values (TMM) method implemented in calcNormFactors() to account for composition bias across libraries.

Normalized expression values were log-transformed to log2 counts per million (logCPM) for better visualization and variance stabilization. Sample clustering was assessed through Multidimensional Scaling (MDS) plots to check for outliers and group separation.


Library Sizes
Million Reads
0
5
10
15
20
45_nCOVD01_C
1EUHM-Draw-1
2EUHM-Draw-2
3EUHM-Draw-1
4EUHM-Draw-1
6EUHM-Draw-1
7EUHM-Draw-1
0EUHM-Draw-1
nCOVD21EUHM
4EUHM-Draw-1
COVD029EUHM
4EUHM-Draw-2
5EUHM-Draw-1
5EUHM-Draw-2
8EUHM-Draw-1
3EUHM-Draw-1
4EUHM-Draw-1
S061_257
S062_258
S063_259
S064_260
S065_261
S066_265
S067_270
S068_272
S069_273
S070_279
S071_280
S181_255
S182_SHXA10
S183_263
S184_SHXA18
S185_266
S186_SHXA14



2.4 Differential Gene Expression Analysis

Differential expression analysis was performed using the **edgeR** quasi-likelihood pipeline. A design matrix was constructed to model the group effect (COVID-19 vs. Convalescent). Dispersion estimates were obtained using estimateDisp() and a generalized linear model (GLM) was fitted with glmQLFit().

Statistical testing was conducted using glmQLFTest() to obtain log2 fold changes, p-values, and false discovery rate (FDR) adjusted significance values. Genes were considered significantly differentially expressed if:

- FDR < 0.05
- |log2 fold change| > 1

Results were visualized using a volcano plot and a heatmap of the top 50 most significantly regulated genes.

2.5 Functional Enrichment Analysis

To interpret the biological significance of the DEGs, gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analyses were conducted using the clusterProfiler package.

2.5.1 GO Enrichment

Gene IDs were converted from ENSEMBL to Entrez format using the biomaRt and bitr() functions. GO enrichment focused on the Biological Process (BP) category using enrichGO(), with cutoff thresholds of:

- p-value < 0.05
- q-value < 0.2

Results were visualized using barplots, dotplots, enrichment maps, and circular network plots (cnetplot).

2.5.2 KEGG Pathway Enrichment

KEGG pathway analysis was performed using enrichKEGG() for human (organism = 'hsa'). Significant pathways were visualized via barplots and dotplots.

2.6 Gene Set Enrichment Analysis (GSEA)

To evaluate the enrichment of ranked gene lists, Gene Set Enrichment Analysis (GSEA) was performed using the gseGO() and GSEA() functions.

2.6.1 GO-Based GSEA

Genes were ranked by log2 fold change. Enrichment was tested against GO-Biological Process terms. Visualizations included dotplots, ridgeplots, enrichment maps, and GSEA running score plots for specific terms.

2.6.2 Immune Signature GSEA

To assess immune-specific transcriptional programs, immune-related gene sets were retrieved from the MSigDB C7 (immunologic signatures) collection using msigdbr(). These gene sets were used to run GSEA and uncover enriched immune signatures in COVID-19.

2.7 Exporting and Saving Results

All key results were exported as .csv files for transparency and reuse, including:

- Differential expression table (DEGs_GSE152418.csv)
- GO enrichment results
- KEGG pathway results

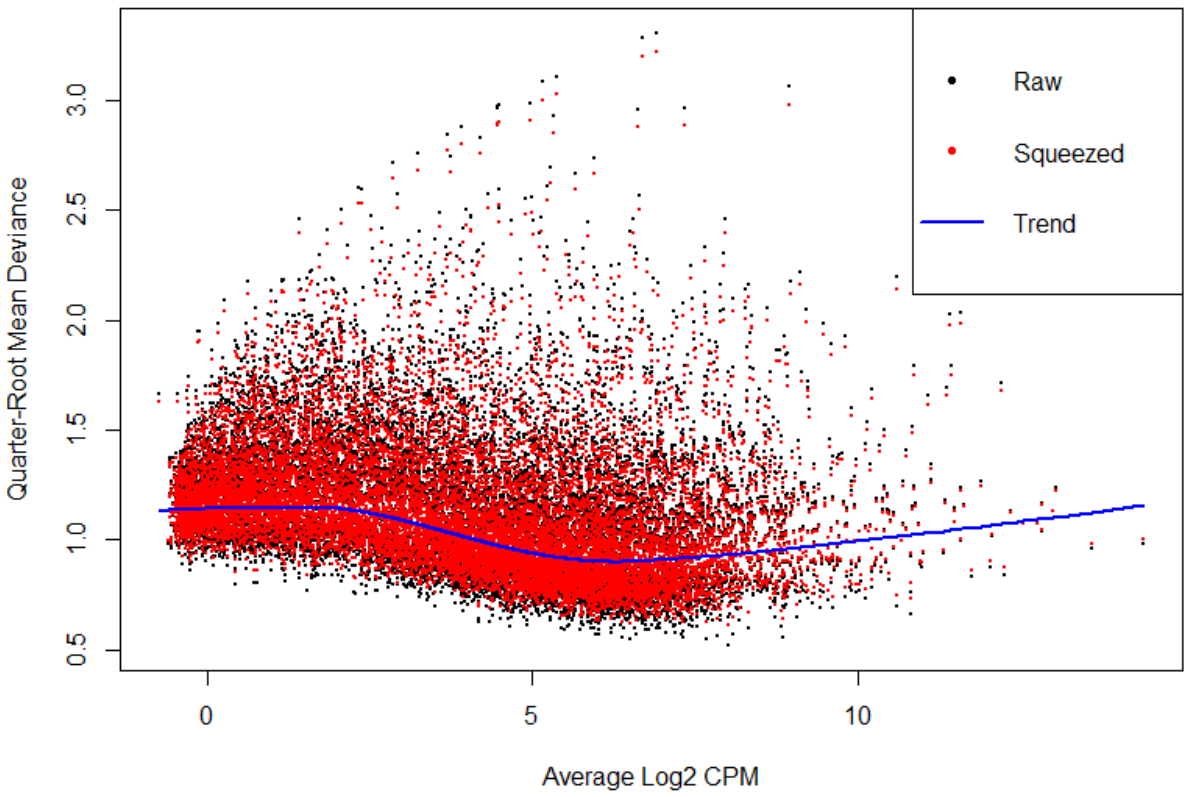
Plots were generated using ggplot2, pheatmap, and enrichplot, and saved for reporting.

3. Differential Gene Expression Analysis

To identify genes that are significantly modulated during active COVID-19 infection compared to convalescent recovery, a differential gene expression (DGE) analysis was conducted using the edgeR package, which is tailored for RNA-seq count data and implements a robust quasi-likelihood (QL) framework.

3.1 Model Design and Dispersion Estimation

A design matrix was constructed to model the effect of the experimental group (COVID-19 vs. Convalescent). The DGEList object, previously normalized and filtered, was used to estimate both common and tagwise dispersions with estimateDisp(), accounting for biological variation among replicates. The dispersion estimates were visualized using plotQLDisp(fit) to confirm appropriate model fit.



3.2 Fitting the Model and Statistical Testing

Using the dispersion estimates, a generalized linear model was fitted with the glmQLFit() function, followed by a quasi-likelihood F-test (glmQLFTest()) to compare gene expression levels between the two groups.

The test results were extracted using topTags() and converted into a data frame for further analysis and visualization. Each gene was characterized by its log2 fold change (logFC), raw p-value, and adjusted p-value (FDR) to account for multiple testing.

3.3 Summary of Results

Out of the total tested genes:

- 2,753 genes were significantly downregulated in COVID-19 patients compared to convalescent individuals.
- 3,240 genes were significantly upregulated in COVID-19 patients.
- 9,681 genes showed no significant change (FDR ≥ 0.05 or $|\log2FC| \leq 1$).

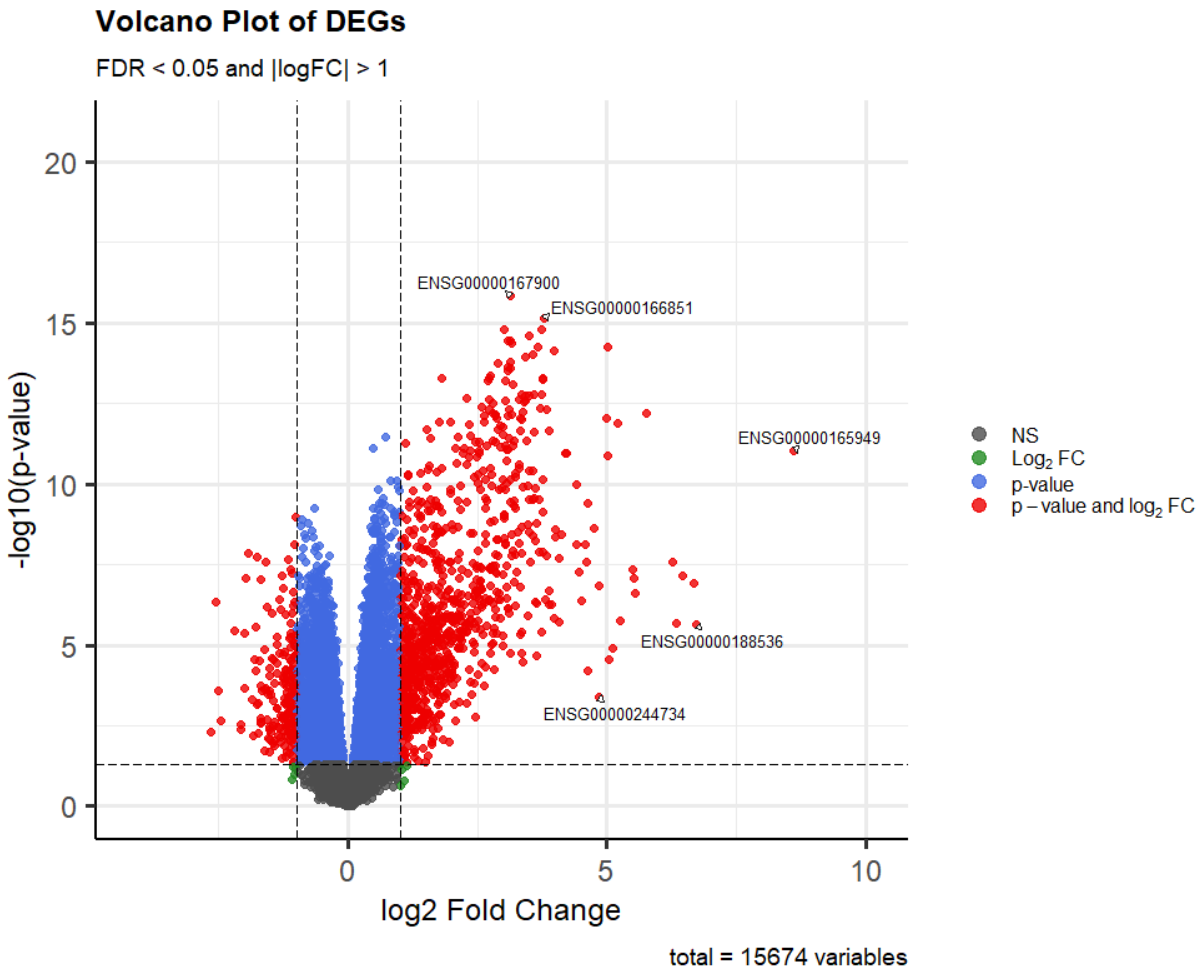
Top 10 Differentially Expressed Genes

	logFC	logCPM	F	PValue	FDR
ENSG00000167900	3.131654	4.894398	211.5524	0	0
ENSG00000166851	3.792626	3.580563	190.9849	0	0
ENSG00000171848	3.742333	5.764055	180.1725	0	0
ENSG00000111206	3.001165	2.552239	181.0904	0	0
ENSG00000175063	3.483723	3.194802	175.5504	0	0
ENSG00000178999	3.135028	2.944081	171.3765	0	0
ENSG00000088325	3.085792	3.861850	170.7789	0	0
ENSG00000123485	3.150995	2.495765	169.7789	0	0
ENSG00000154277	5.017562	1.001783	116.8335	0	0
ENSG00000145386	3.665545	3.538345	166.1403	0	0

3.4 Visualization of Differential Expression

Volcano Plot

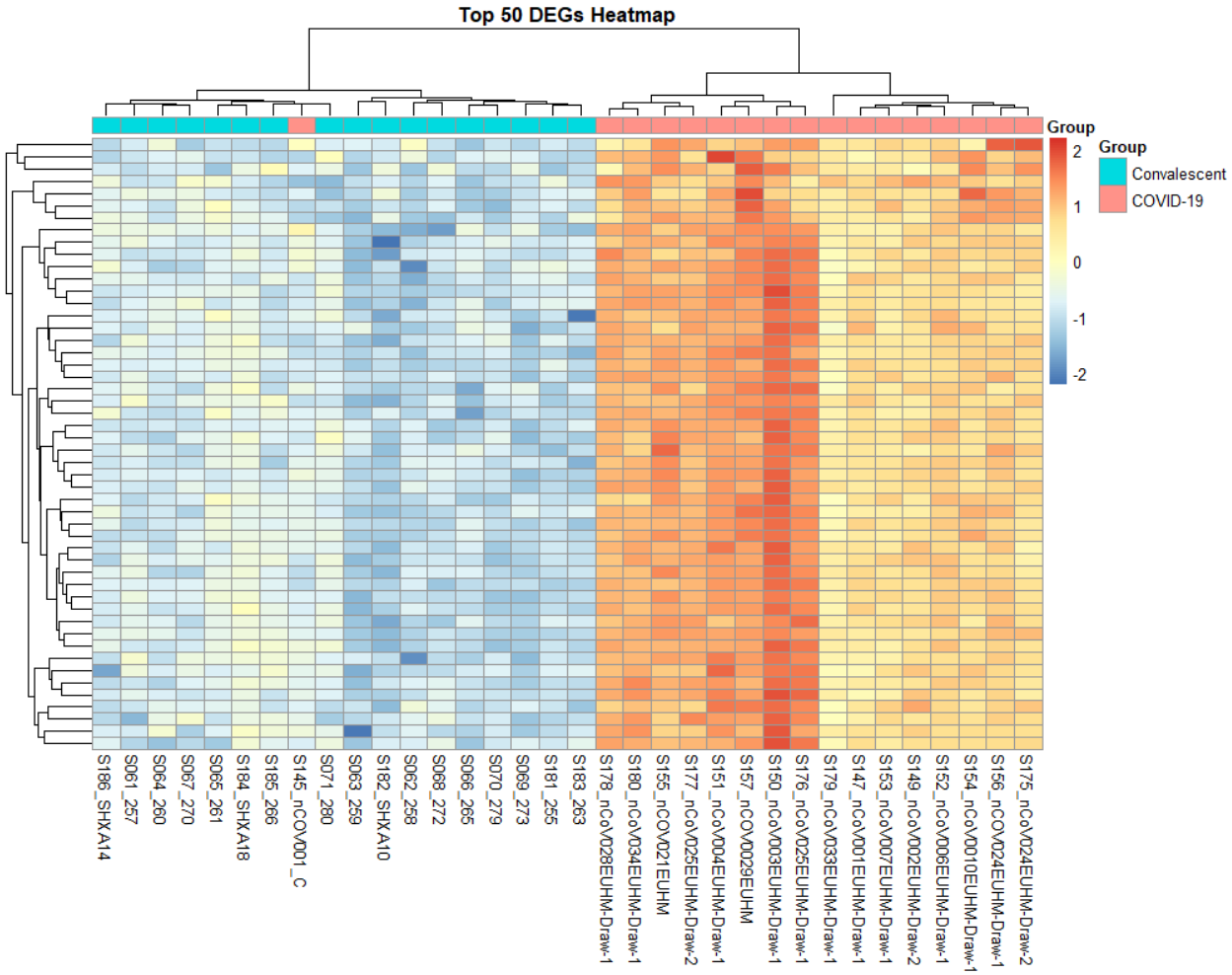
A volcano plot was generated to visually represent the global distribution of gene expression changes. Genes with |log2 fold change| > 1 and FDR < 0.05 were highlighted in red to denote statistical significance:



This plot highlights the prominent upregulation and downregulation patterns, offering a quick visual reference for significant genes.

Heatmap of Top 50 DEGs

To explore the expression dynamics of the most differentially regulated genes, a heatmap of the top 50 DEGs (ranked by statistical significance) was plotted using pheatmap. Expression values were scaled across genes to emphasize relative changes, and samples were annotated by group:



The heatmap reveals clear segregation between COVID-19 and convalescent groups, suggesting distinct transcriptional profiles and validating the robustness of the DGE analysis.

4. Functional Enrichment Analysis

To elucidate the biological implications of the differentially expressed genes (DEGs) identified between COVID-19 and convalescent individuals, functional enrichment analysis was performed. This step aimed to uncover overrepresented biological processes and signaling pathways, thereby translating the observed gene-level changes into system-level insights. Analyses focused on Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways using the clusterProfiler package.

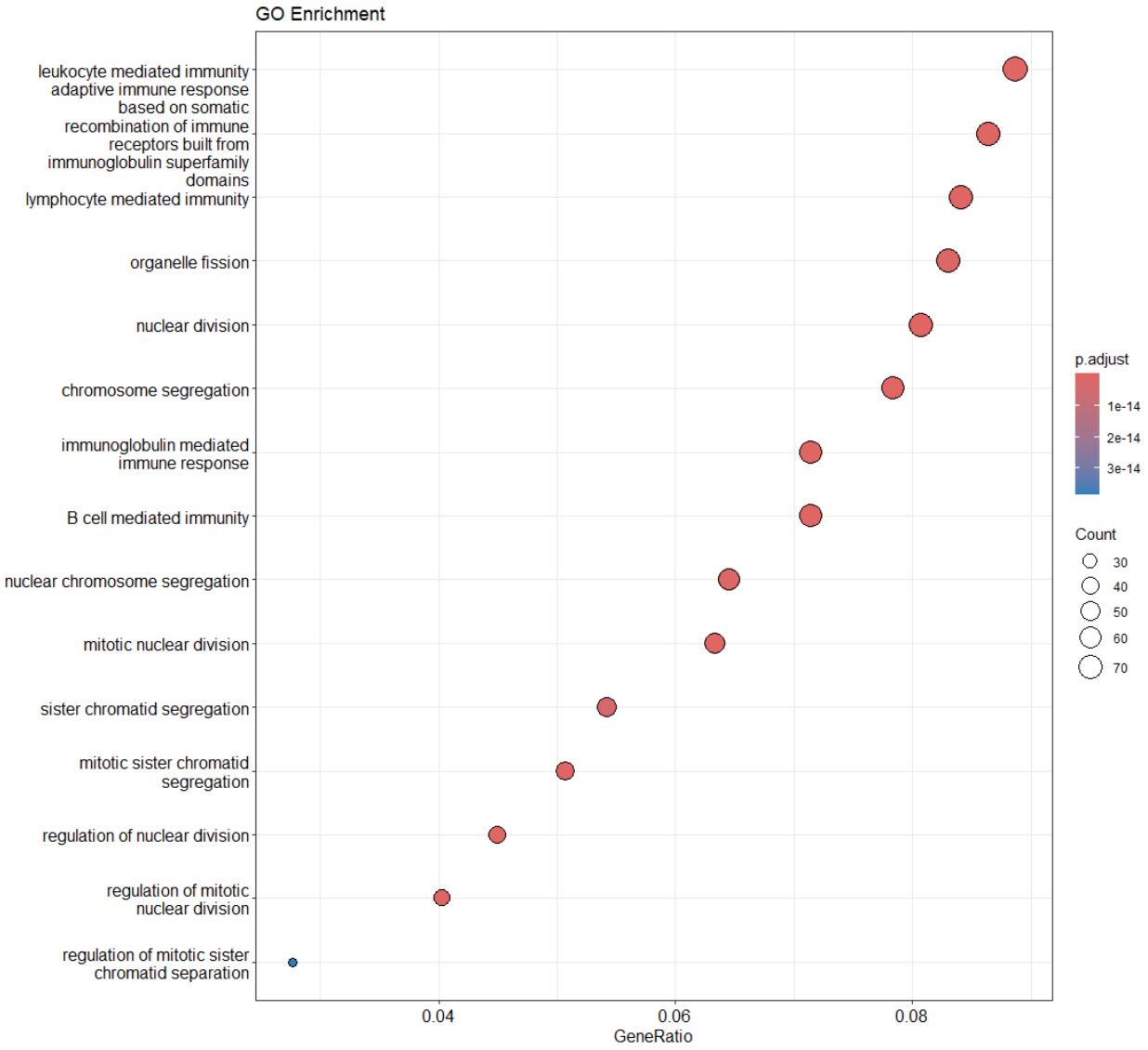
4.1 Gene Ontology (GO) Enrichment

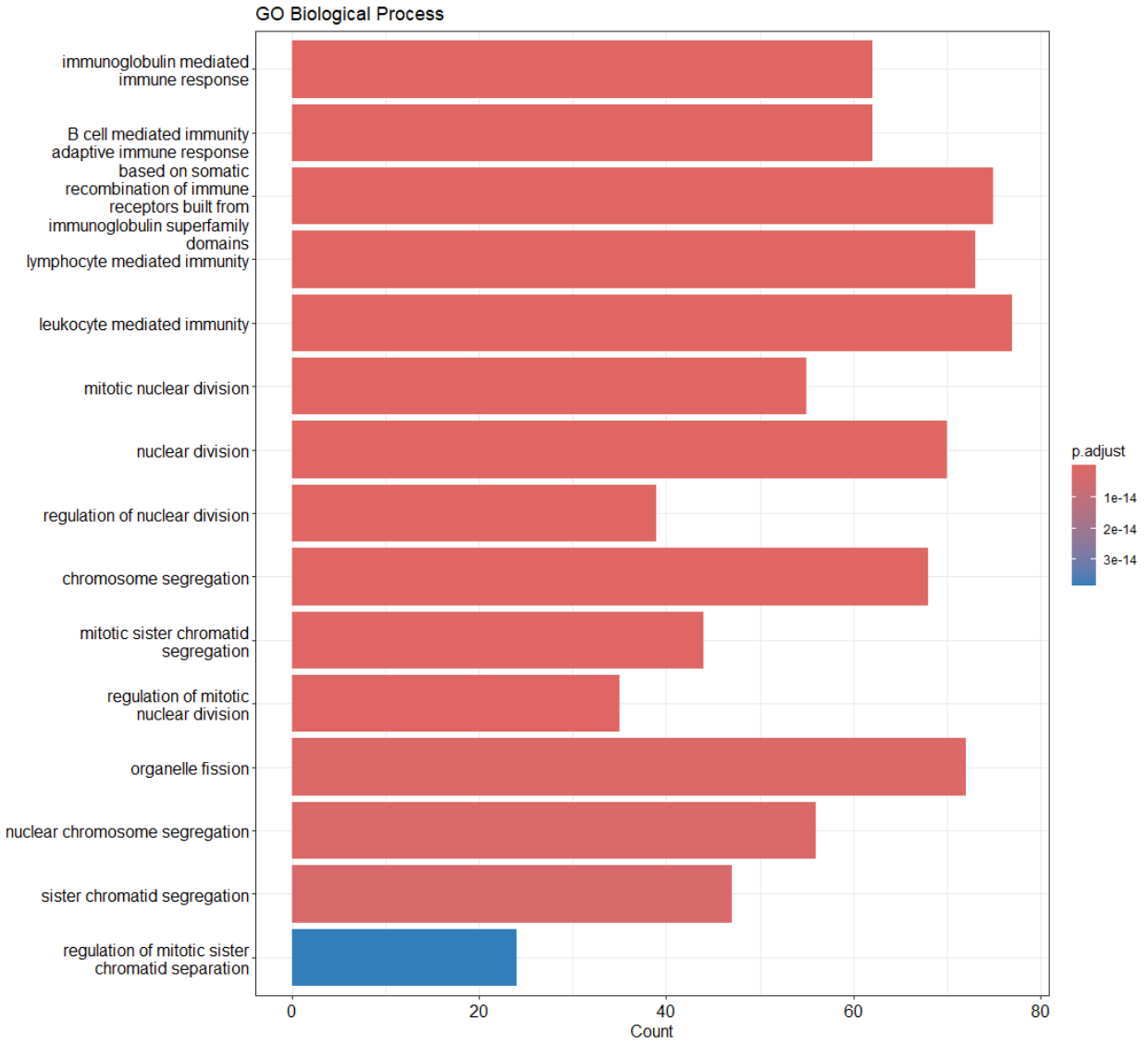
4.1.1 Gene ID Conversion

Prior to enrichment, gene symbols were converted to Entrez Gene IDs using the biomaRt and bitr() functions. Only DEGs with $|\log2FC| > 1$ and $FDR < 0.05$ were retained for enrichment:

4.1.2 GO Biological Process Analysis

Enrichment was performed using the enrichGO() function, targeting the Biological Process (BP) ontology:


GO Enrichment
leukocyte mediated immunity
adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains
lymphocyte mediated immunity
organelle fission
nuclear division
chromosome segregation
immunoglobulin mediated immune response
B cell mediated immunity
nuclear chromosome segregation
mitotic nuclear division
sister chromatid segregation
mitotic sister chromatid segregation
regulation of nuclear division
regulation of mitotic nuclear division
regulation of mitotic sister chromatid separation
0.04
0.06
0.08
GeneRatio
p.adjust
1e-14
2e-14
3e-14
Count
30
40
50
60
70



Significantly enriched GO terms included processes related to:

- Lymphocyte activation
- Cytokine-mediated signaling
- Antigen processing and presentation
- T cell differentiation
- Inflammatory response regulation

These results suggest a robust reprogramming of immune-related pathways in COVID-19 patients compared to recovered individuals.

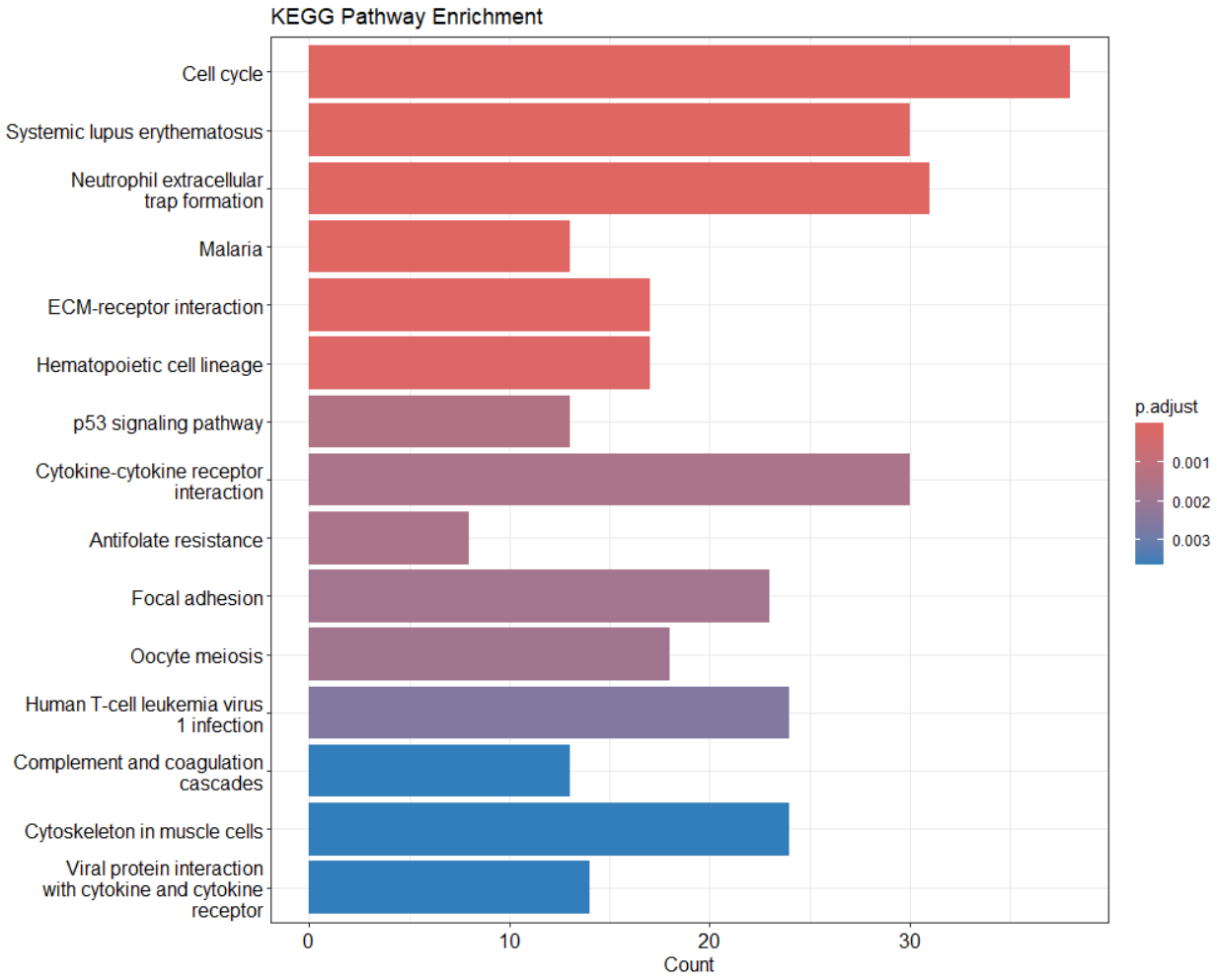
4.2 KEGG Pathway Enrichment

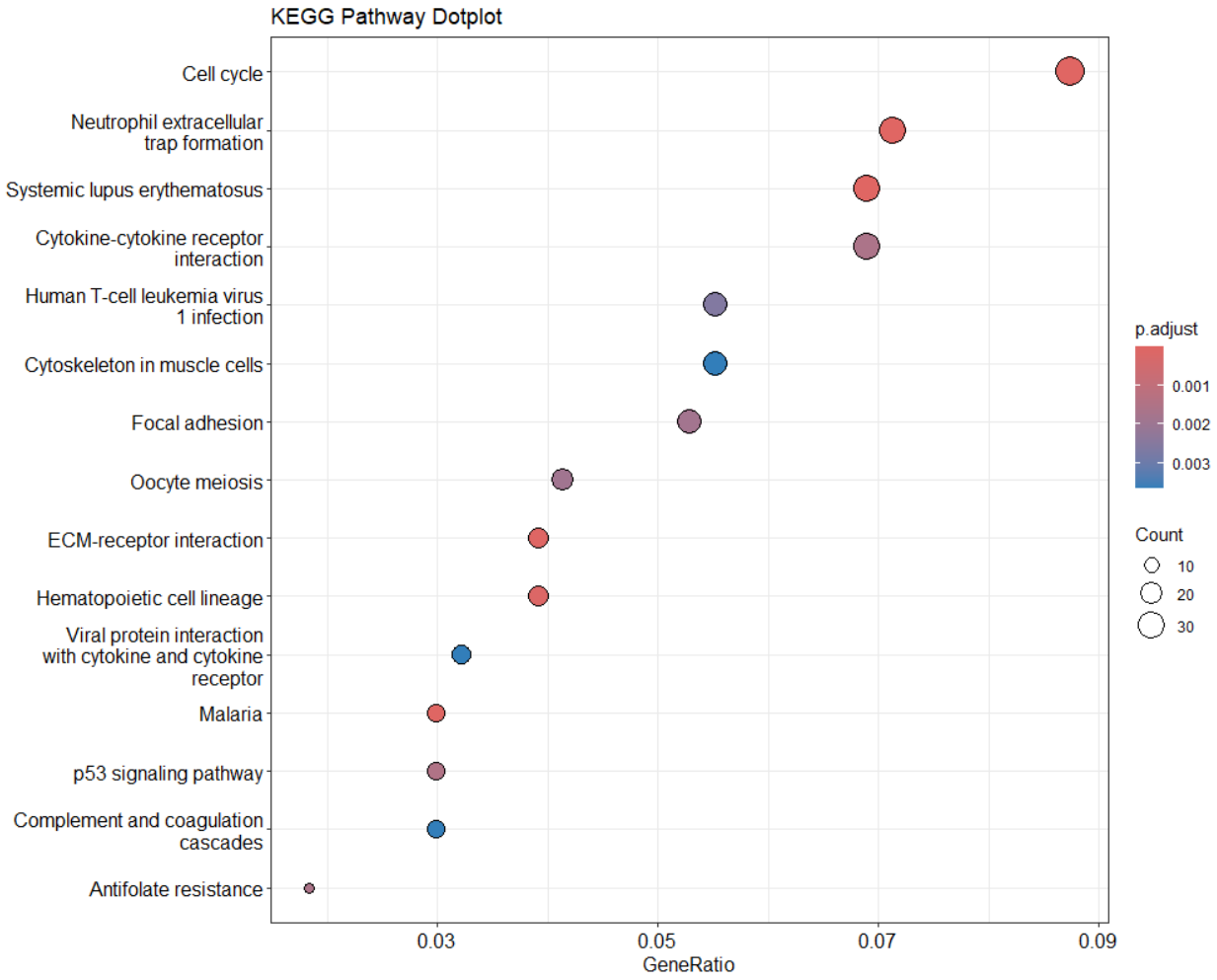
The enrichKEGG() function was applied to discover enriched molecular pathways from the KEGG database:

Significant KEGG pathways included:

- Cytokine-cytokine receptor interaction
- NF-kappa B signaling pathway
- Antigen processing and presentation
- Toll-like receptor signaling
- Natural killer cell-mediated cytotoxicity

These pathways are consistent with known mechanisms of viral response, including innate immune activation, inflammation, and immune cell recruitment.


KEGG Pathway Enrichment
Cell cycle
Systemic lupus erythematosus
Neutrophil extracellular trap formation
Malaria
ECM-receptor interaction
Hematopoietic cell lineage
p53 signaling pathway
Cytokine-cytokine receptor interaction
Antifolate resistance
Focal adhesion
Oocyte meiosis
Human T-cell leukemia virus 1 infection
Complement and coagulation cascades
Cytoskeleton in muscle cells
Viral protein interaction with cytokine and cytokine receptor
0
10
20
30
Count
p.adjust
0.001
0.002
0.003



These visuals highlight how critical immune signaling cascades are differentially engaged in COVID-19, supporting their potential relevance as biomarkers or therapeutic targets.

5. Gene Set Enrichment Analysis (GSEA)

While traditional enrichment analysis focuses on discrete sets of differentially expressed genes (DEGs), Gene Set Enrichment Analysis (GSEA) evaluates ranked gene lists to identify biologically meaningful trends that might be missed by arbitrary statistical cutoffs. In this study, GSEA was used to further investigate coordinated gene expression programs relevant to COVID-19 pathogenesis, particularly focusing on Gene Ontology (GO) terms and immune-specific pathways.

5.1 Gene Ranking and Preparation

The full list of genes from the differential expression analysis was ranked based on log2 fold change (logFC), representing the direction and magnitude of expression change. To ensure compatibility with enrichment databases, ENSEMBL gene IDs were mapped to Entrez IDs using the biomaRt package. Duplicates and unmapped entries were removed to clean the gene list.

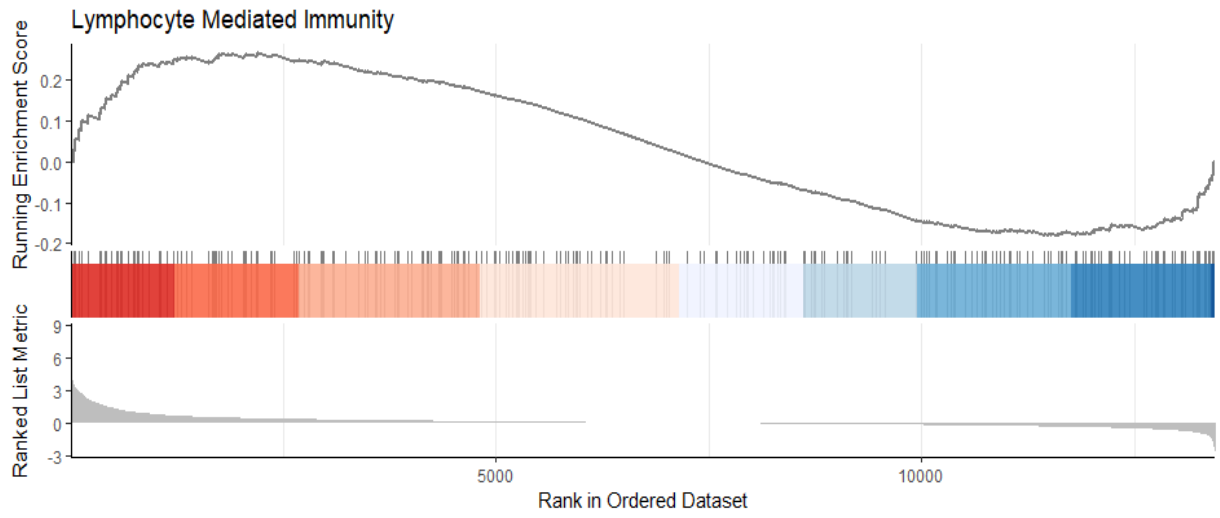
5.2 GSEA: GO Biological Process (BP)

GSEA was performed on the ranked gene list using the gseGO() function to detect enriched GO Biological Processes across the expression spectrum:

5.2.1 GO GSEA Findings

Top enriched GO terms included:

- Regulation of lymphocyte-mediated immunity
- Interferon-gamma production
- T cell activation and proliferation
- Myeloid leukocyte activation



These results suggest that immune signaling and inflammatory responses remain globally altered in COVID-19, not just limited to DEGs but across the gene expression distribution.

5.3 GSEA: Immune-Related Gene Sets (MSigDB C7)

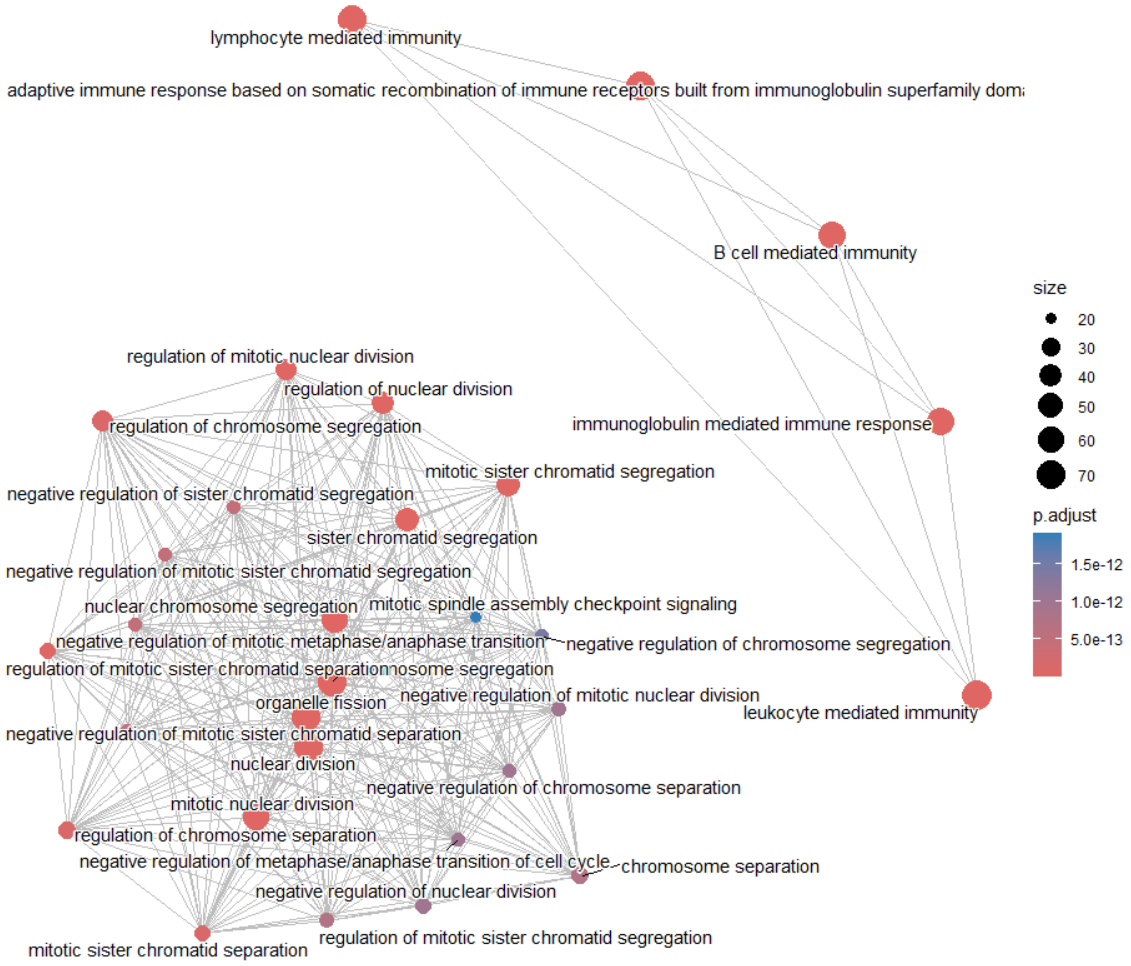
To deepen the immune interpretation, gene set enrichment was extended to MSigDB C7 collections—curated immunologic signatures from experimental datasets—using the msigdbr package.

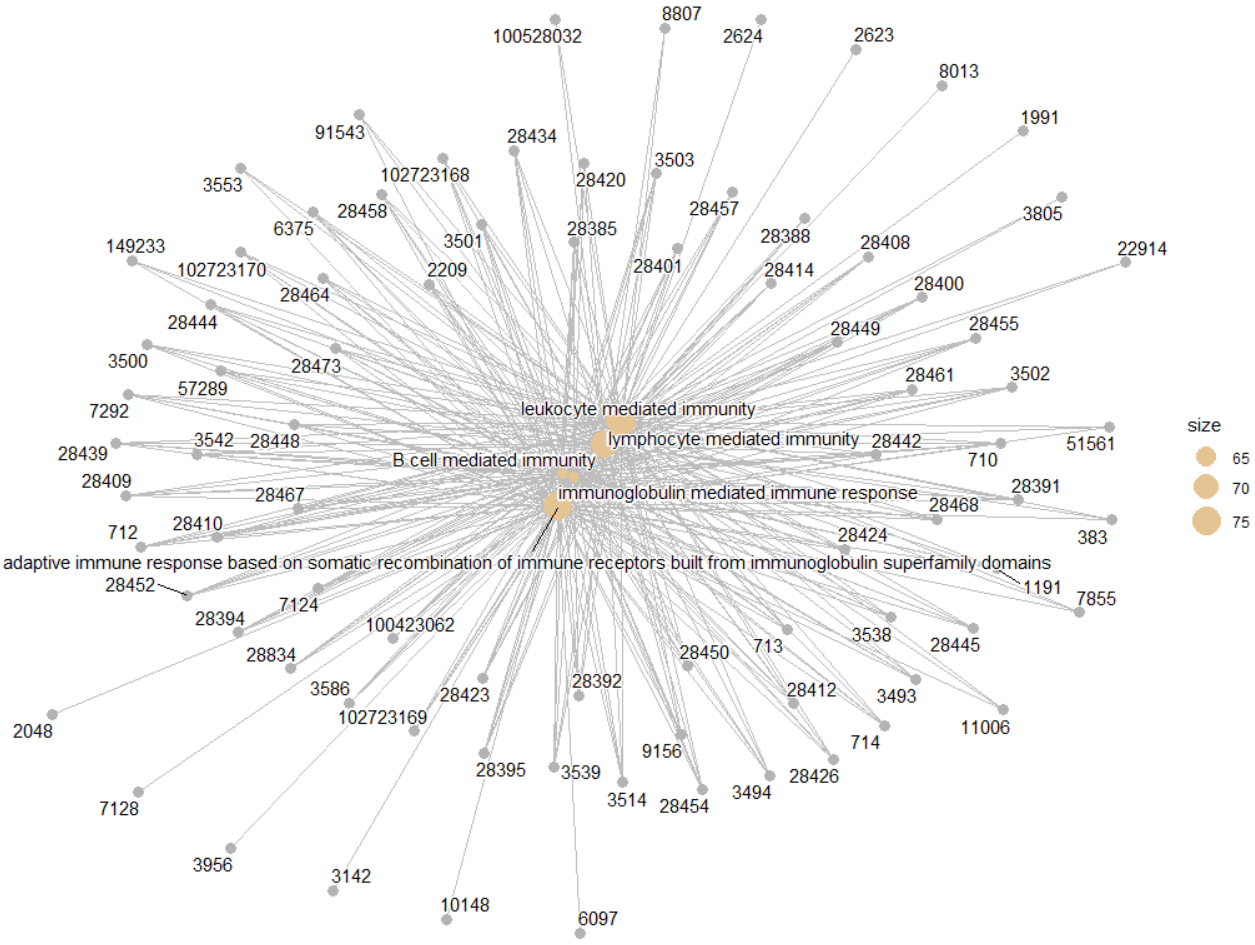
5.3.1 Immune GSEA Results

The analysis revealed strong enrichment in pathways involving:

- CD4+ and CD8+ T cell activation
- NK cell cytotoxicity
- Monocyte inflammatory responses
- Type I and II interferon responses

This highlights that even genes with modest fold changes contribute to broader immune regulatory programs.


lymphocyte mediated immunity
adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily dom
B cell mediated immunity
size
20
30
40
50
60
70
p.adjust
1.5e-12
1.0e-12
5.0e-13
regulation of mitotic nuclear division
regulation of nuclear division
regulation of chromosome segregation
immunoglobulin mediated immune response
mitotic sister chromatid segregation
negative regulation of sister chromatid segregation
sister chromatid segregation
negative regulation of mitotic sister chromatid segregation
nuclear chromosome segregation
mitotic spindle assembly checkpoint signaling
negative regulation of mitotic metaphase/anaphase transition
negative regulation of chromosome segregation
regulation of mitotic sister chromatid separationnosome segregation
organelle fission
negative regulation of mitotic nuclear division
leukocyte mediated immunity
negative regulation of mitotic sister chromatid separation
nuclear division
negative regulation of chromosome separation
mitotic nuclear division
regulation of chromosome separation
negative regulation of metaphase/anaphase transition of cell cycle
chromosome separation
negative regulation of nuclear division
regulation of mitotic sister chromatid segregation
mitotic sister chromatid separation



These findings affirm the dominance of immune signaling in the COVID-19 transcriptome and provide finer resolution than classical DEG-based enrichment alone.

6. Results Summary

This study presents a comprehensive transcriptomic analysis comparing whole blood gene expression profiles of COVID-19 patients and convalescent individuals using the GSE152418 dataset. The results provide strong evidence of immune system activation and dysregulation during active infection, supported by differential gene expression patterns and pathway enrichment.

6.1 Differential Gene Expression

- A total of 5,993 genes were found to be significantly differentially expressed (FDR < 0.05, |log2FC| > 1):
    - 3,240 genes were upregulated
    - 2,753 genes were downregulated

6.2 Functional Enrichment (GO and KEGG)

Gene Ontology (GO) Biological Processes:

- Enriched terms were predominantly immune-related, including:
    - Lymphocyte activation
    - Cytokine-mediated signaling
    - T cell differentiation
    - Antigen processing and presentation

KEGG Pathways:

- Significantly enriched pathways included:
    - Cytokine-cytokine receptor interaction
    - NF-kappa B signaling pathway
    - Toll-like receptor signaling
    - Natural killer cell-mediated cytotoxicity

6.3 Gene Set Enrichment Analysis (GSEA)

GO-Based GSEA:

- Even genes with modest expression changes were collectively enriched for immune-related biological processes.
- Top enriched terms included:
    - Regulation of lymphocyte-mediated immunity
    - Interferon-gamma production
    - Myeloid leukocyte activation

Immune Signature GSEA (MSigDB C7):

- Enriched gene sets were associated with:
    - CD4+/CD8+ T cell activation
    - Natural killer cell responses
    - Inflammatory monocyte signatures
    - Interferon-stimulated pathways

These results indicate a globally heightened and coordinated immune response signature in COVID-19 patients, beyond the subset of DEGs.

7. Discussion

This study provides a transcriptome-wide view of host immune responses during active COVID-19 infection compared to the convalescent phase. By leveraging differential expression analysis and both classical and rank-based functional enrichment approaches, we identified key genes and pathways that reflect the immunological dynamics of SARS-CoV-2 infection in peripheral blood.

7.1 Immune Activation in Active Infection

The analysis revealed a significant number of differentially expressed genes (DEGs), with more than 3,000 genes upregulated in COVID-19 patients relative to convalescent individuals. Many of these genes were enriched in biological processes and pathways central to innate immunity, T cell activation, and inflammatory signaling. This is consistent with previous findings that COVID-19 triggers a systemic immune response characterized by heightened interferon signaling, pro-inflammatory cytokine production, and lymphocyte activation.

The enrichment of pathways such as NF-kappa B signaling, cytokine-cytokine receptor interaction, and Toll-like receptor signaling further emphasizes the role of innate immune mechanisms in viral recognition and response. These pathways contribute to both antiviral defense and the cytokine storm phenomenon associated with severe disease outcomes.

7.2 Adaptive Immune Signatures and Convalescence

Interestingly, the gene expression profiles of convalescent individuals exhibited a relative downregulation of these immune pathways, indicating a return toward homeostasis following viral clearance. GSEA revealed that T cell–mediated immunity, while still detectable, showed lower activity in the convalescent group, suggesting resolution of acute inflammation and restoration of immune balance.

Furthermore, enrichment of gene sets related to memory T cells, natural killer cell cytotoxicity, and antigen processing points to the ongoing adaptation of the immune system post-infection—potentially reflecting memory formation and long-term immune surveillance.

7.3 Implications for Disease Monitoring and Therapeutics

The ability to distinguish COVID-19 patients from convalescent individuals based on blood transcriptomic signatures has several important implications:

- **Biomarker discovery**: The identified DEGs and pathways could serve as candidate biomarkers for disease severity, prognosis, or treatment response.
- **Therapeutic targeting**: Dysregulated pathways such as **interferon signaling** and **NF-kappa B activation** could be explored for targeted modulation in severe COVID-19 cases.
- **Immune monitoring**: Understanding how immune signatures evolve from infection to recovery may support vaccine evaluation or long-COVID diagnostics.

8. Conclusion

This study presents a comprehensive transcriptomic analysis of whole blood samples from individuals with active COVID-19 infection and those in the convalescent phase, utilizing the publicly available GSE152418 dataset. Through rigorous differential expression and enrichment analyses, we identified thousands of genes significantly altered in expression during active infection, many of which are involved in immune system regulation, inflammation, and antiviral defense.

Key findings include:

- Upregulation of immune-related pathways such as cytokine signaling, lymphocyte activation, and interferon responses during active infection.
- Downregulation or normalization of these immune signatures in convalescent individuals, suggesting resolution of acute immune activation.
- Enrichment of immune-specific gene sets through GSEA, further highlighting the complex interplay between innate and adaptive immunity during SARS-CoV-2 infection and recovery.

These results contribute to our understanding of the systemic immune landscape of COVID-19 and provide valuable leads for identifying biomarkers, therapeutic targets, and monitoring tools. Despite limitations such as moderate sample size and lack of clinical annotation, the analysis demonstrates the power of transcriptomic profiling in uncovering host-pathogen interactions and immune dynamics.

Future studies leveraging longitudinal sampling, single-cell sequencing, and integrative multi-omics approaches will be essential to fully elucidate the trajectory of immune responses in COVID-19 and to inform both acute and long-term management strategies.