International Journal of Medical and Health Sciences
Volume 12 Issue 01 January 2026
ISSN- 2456 - 6063
DOI: https://doi.org/10.69980/w53n3j86

INTEGRATIVE TRANSCRIPTOMIC, PROTEOMIC, AND MACHINE LEARNING ANALYSIS OF COVID-19 SEVERITY AND ITS IMPLICATIONS FOR POST-VIRAL OUTCOMES

Himanshu Shekhar Singh[1],Janani Sp[1], Shaik Mubarak Basha[1], Mahak[1], Dr. Ramachandraprasad La[1],Dr Chemmugil[1]

[1*]Department of life sciences, School of sciences, Garden city university, Bangalore, India.

Corresponding author:-Himanshu Shekhar, Email: singhhimanshuss1003@gmail.com

Received:-16/02/26	Revised:-25/03/26	Accepted:-01/04/26	Published:- 08/04/26

ABSTRACT

COVID-19, caused by SARS-CoV-2, exhibits wide variability in clinical severity, ranging from asymptomatic infection to critical respiratory failure. Understanding the molecular determinants of this severity is essential for biomarker discovery, risk stratification, and improved clinical management. This study presents an integrative multi-omics framework combining transcriptomic analysis, proteomic enrichment profiling, and machine learning-based feature prioritization to identify molecular signatures associated with COVID-19 severity and their potential implications for post-viral outcomes.

Transcriptomic analysis of the publicly available whole-blood dataset GSE213313 (Agilent microarray) identified 1,544 significantly differentially expressed genes (FDR < 0.05) in critical versus healthy samples, with enrichment of innate immune response, leukocyte activation, antigen processing, T-cell differentiation, and cytokine-mediated signalling pathways. Proteomic analysis using a published peer-reviewed dataset comparing mild (n = 3) and severe (n = 5) COVID-19 cases identified 91 severity-associated proteins, with significant enrichment of acute-phase response, complement cascade activation, platelet degranulation, and coagulation-related processes.

Multi-omics integration identified 8 overlapping molecules between transcriptomic and proteomic datasets, with an inverse RNA–protein fold-change correlation ($r = -0.58$), indicative of post-transcriptional regulatory complexity. Machine learning analysis using a Random Forest classifier demonstrated perfect discrimination between mild and severe cases (AUC = 1.0; LOOCV accuracy = 100%), and prioritized proteins including LIPG, CLEC11A, KDR, FAIM3, and BST1 as key severity-associated features.

Collectively, this study demonstrates that severe COVID-19 is characterized by coordinated dysregulation of immune activation, inflammatory signalling, complement amplification, and coagulation mechanisms across multiple molecular layers. The identified molecular signatures provide a foundation for future biomarker validation and therapeutic targeting in COVID-19 and related post-viral conditions.

Keywords - Transcriptomics, Proteomics, Machine Learning, COVID-19 Severity, Multi-Omics Integration, Post-Viral Outcomes

CHAPTER 1: INTRODUCTION

1.1 Global Burden of COVID-19

Coronavirus Disease 2019 (COVID-19), caused by Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2), has emerged as one of the most significant global health challenges in recent history. Since its emergence in late 2019, the disease has resulted in widespread morbidity and mortality, severely impacting healthcare systems worldwide (World Health Organization, 2020). Despite advances in vaccination and therapeutic strategies, COVID-19 continues to pose a threat due to emerging variants and heterogeneous clinical outcomes (Huang et al., 2020).

One of the most striking features of COVID-19 is the broad variability in clinical presentation, ranging from asymptomatic infection to severe respiratory failure and multi-organ dysfunction (Zhou et al., 2020). This variability highlights the importance of host-specific molecular and immunological factors in disease progression.

1.2 Clinical Spectrum and Severity of COVID-19

COVID-19 severity is commonly classified into mild, moderate, severe, and critical stages based on respiratory symptoms, oxygen requirement, and organ involvement (WHO, 2020). While mild cases typically resolve without complications, severe and critical cases often require hospitalization and intensive care support.

Several studies have demonstrated that disease severity is not solely dependent on viral load but is strongly influenced by host immune responses and inflammatory mechanisms (Tay et al., 2020). Patients with severe COVID-19 frequently exhibit excessive immune activation, endothelial dysfunction, and coagulation abnormalities, contributing to poor clinical outcomes (Lucas et al., 2020).

1.3 Host Immune Response to SARS-CoV-2 Infection

The host immune response plays a central role in determining the outcome of SARS-CoV-2 infection. While an early and balanced immune response is essential for viral clearance, dysregulated immune activation can lead to pathological inflammation and tissue damage (Blanco-Melo et al., 2020).

Severe COVID-19 cases are characterized by elevated levels of pro-inflammatory cytokines, impaired interferon responses, and immune cell exhaustion (Hadjadj et al., 2020). This imbalance between protective and pathological immune responses has been identified as a key driver of disease severity and complications.

1.4 COVID-19 Severity and Post-Viral Outcomes

Increasing clinical evidence suggests that individuals who experience severe COVID-19 are at a higher risk of developing persistent symptoms after recovery, including fatigue, respiratory dysfunction, and neurological manifestations (Nalbandian et al., 2021). These post-viral outcomes are believed to be associated with prolonged immune dysregulation and molecular disturbances initiated during the acute phase of infection.

Understanding molecular signatures associated with disease severity may therefore provide valuable insights into biological mechanisms contributing to post-viral outcomes and prolonged recovery (Su et al., 2023).

1.5 Role of Transcriptomics in COVID-19 Research

Transcriptomic profiling has been extensively used to study host responses to SARS-CoV-2 infection. Both microarray-based and RNA-sequencing studies have reported significant changes in the expression of genes involved in immune signalling, interferon pathways, and inflammatory responses (Blanco-Melo et al., 2020).

Microarray platforms, including Agilent microarrays, offer reliable and reproducible gene expression measurements and have been widely applied in clinical transcriptomic studies (Ritchie et al., 2015). However, transcriptomic data represent changes at the RNA level and may not always correlate directly with protein abundance or functional activity due to post-transcriptional regulation (Liu et al., 2016).

1.6 Role of Proteomics in Understanding COVID-19 Severity

Proteomics provides direct insights into protein expression and functional activity, offering a more immediate representation of biological processes than transcriptomics alone (Aebersold & Mann, 2016). Proteomic studies in COVID-19 have identified severity-associated proteins involved in immune regulation, acute-phase response, coagulation, and complement activation (Messner et al., 2020).

Comparative proteomic analyses between mild and severe COVID-19 patients have revealed distinct protein signatures linked to disease progression and immune dysregulation (Shen et al., 2020). These findings highlight the importance of proteomics for functional biomarker discovery.

1.7 Limitations of Single-Omics Approaches

Although transcriptomics and proteomics individually provide valuable insights, each approach has inherent limitations. Transcriptomic analyses do not capture post-translational modifications or protein degradation, while proteomic analyses alone may not explain upstream regulatory mechanisms (Hasin et al., 2017). These limitations emphasize the need for integrative approaches that combine multiple molecular layers to achieve a comprehensive understanding of disease biology.

1.8 Importance of Multi-Omics Integration

Multi-omics integration combines complementary molecular datasets to provide a systems-level view of biological processes (Hasin et al., 2017). In COVID-19 research, integrative analyses have enabled improved characterization of immune dysregulation and inflammatory pathways associated with disease severity (Su et al., 2023). By integrating transcriptomic and proteomic data, researchers can identify consistent molecular signatures across different biological levels, increasing confidence in identified pathways and biomarkers.

1.9 Role of Machine Learning in Omics Data Analysis

The complexity and high dimensionality of omics datasets require advanced computational approaches for effective analysis. Machine learning techniques have been increasingly applied in biomedical research to identify complex patterns and prioritize biologically relevant features (Libbrecht & Noble, 2015). Random Forest, an ensemble learning method, is particularly well-suited for omics data analysis due to its robustness, ability to handle high-dimensional data, and provision of interpretable feature importance measures (Breiman, 2001).

1.10 Rationale and Significance of the Study

Despite extensive research, the molecular determinants underlying COVID-19 severity remain incompletely understood. Single-omics approaches provide partial insights but fail to capture the full complexity of disease biology. Integrating transcriptomics, proteomics, and machine learning offers a comprehensive framework for identifying severity-associated molecular signatures and understanding their biological relevance (Su et al., 2023).

1.11 Molecular Basis of Immune Dysregulation in Severe COVID-19

Severe COVID-19 is characterized by profound alterations in immune cell composition and function. Multiple studies have reported excessive activation of innate immune cells, particularly neutrophils and monocytes, leading to increased production of inflammatory mediators and tissue-damaging enzymes (Lucas et al., 2020). Neutrophil extracellular trap (NET) formation and aberrant macrophage activation have been implicated in lung injury and vascular damage observed in critically ill patients.

In parallel, adaptive immune responses are often impaired in severe cases. Reduced T-cell counts, functional exhaustion of cytotoxic T lymphocytes, and altered B-cell responses have been reported in hospitalized patients (Zhou et al., 2020). These immune abnormalities contribute to ineffective viral clearance and prolonged inflammatory states, further exacerbating disease severity.

1.12 Role of Inflammation, Coagulation, and Endothelial Dysfunction

Inflammation in COVID-19 is closely linked to disturbances in coagulation and endothelial function. Severe cases frequently exhibit elevated levels of D-dimer, fibrinogen, and other coagulation markers, indicating a hypercoagulable state (Tang et al., 2020). Endothelial cell activation and vascular inflammation contribute to microthrombus formation, impaired oxygen exchange, and multi-organ dysfunction.

Proteomic studies have highlighted dysregulation of complement proteins, acute-phase reactants, and coagulation-related factors in severe COVID-19 patients (Messner et al., 2020). These findings suggest that COVID-19 is not solely a respiratory disease but a systemic disorder involving immune, vascular, and metabolic pathways.

1.13 Importance of Biomarker Discovery in COVID-19

Biomarkers play a crucial role in disease diagnosis, prognosis, and therapeutic decision-making. In COVID-19, reliable molecular biomarkers can assist in early identification of patients at risk of severe disease, enabling timely clinical intervention. Several inflammatory markers, including C-reactive protein (CRP), interleukins, and ferritin, have been associated with disease severity; however, these markers lack specificity and mechanistic insight (Huang et al., 2020).

Omics-based biomarker discovery offers an opportunity to identify molecular signatures that reflect underlying disease mechanisms rather than isolated clinical measurements. Integrating transcriptomic and proteomic biomarkers provides a more comprehensive framework for understanding disease progression and heterogeneity.

1.14 Challenges in COVID-19 Omics Research

Despite rapid advancements, COVID-19 omics research faces several challenges. Variability in sample types, patient cohorts, disease stages, and analytical methods can lead to inconsistent findings across studies. Additionally, the high dimensionality of omics data increases the risk of false-positive results if not analyzed carefully (Hasin et al., 2017). These challenges necessitate robust analytical frameworks that combine statistical rigour with computational intelligence.

1.15 Machine Learning for Biomarker Prioritization

Machine learning methods have emerged as powerful tools for extracting meaningful patterns from complex biological data. Unlike traditional statistical methods, machine learning algorithms can model non-linear relationships and interactions among features, which are common in biological systems (Libbrecht & Noble, 2015). Random Forest algorithms are particularly advantageous for biomarker prioritization, as they can rank features based on their contribution to classification performance (Patel et al., 2023).

1.16 Justification for Using Publicly Available Omics Datasets

Publicly available omics datasets offer a valuable resource for biomedical research by enabling data reuse, reproducibility, and transparency. Reanalysing such datasets allows researchers to apply novel analytical frameworks and extract additional insights without the need for resource-intensive experimental generation of data (Rung & Brazma, 2013). In the context of COVID-19, the rapid sharing of omics datasets has accelerated scientific discovery and facilitated comparative analyses across studies.

CHAPTER 2: REVIEW OF LITERATURE

2.1 Overview of COVID-19 Pathophysiology

COVID-19, caused by SARS-CoV-2, is a complex infectious disease characterized by diverse clinical manifestations and outcomes. While many infected individuals experience mild symptoms, a significant proportion develop severe or critical illness marked by respiratory failure, systemic inflammation, and multi-organ involvement (Huang et al., 2020). Early studies highlighted that disease severity is influenced not only by viral factors but also by host immune responses and underlying biological conditions (Zhou et al., 2020).
Understanding the molecular mechanisms underlying this variability is essential for identifying biomarkers of disease severity and improving clinical management strategies.

2.2 Immune Dysregulation and COVID-19 Severity

Severe COVID-19 is strongly associated with dysregulated immune responses. Multiple studies have reported excessive activation of innate immune pathways, leading to elevated levels of pro-inflammatory cytokines and chemokines, a phenomenon often described as a hyperinflammatory state (Tay et al., 2020). This excessive inflammation contributes to tissue damage, vascular dysfunction, and poor clinical outcomes.
At the same time, impairment of adaptive immune responses has been observed in severe cases. Reduced T-cell counts, T-cell exhaustion, and altered B-cell responses have been linked to ineffective viral clearance and prolonged inflammation (Lucas et al., 2020). These findings suggest that a delicate balance between protective immunity and pathological inflammation is crucial in determining COVID-19 severity.

2.3 Transcriptomic Studies in COVID-19

Transcriptomic analysis has been widely applied to study host responses to SARS-CoV-2 infection. Both microarray-based and RNA-sequencing studies have revealed extensive changes in gene expression related to immune signalling, interferon responses, inflammatory pathways, and cellular stress mechanisms (Blanco-Melo et al., 2020).
Several transcriptomic studies have demonstrated that severe COVID-19 cases exhibit upregulation of genes associated with innate immune activation, neutrophil responses, and cytokine signalling, along with downregulation of genes involved in adaptive immunity (Hadjadj et al., 2020). However, transcriptomic data represent changes at the RNA level and may not always correlate directly with protein abundance or activity due to post-transcriptional regulation and protein turnover (Liu et al., 2016).

2.4 Proteomic Studies in COVID-19

Proteomics enables direct measurement of protein abundance and offers functional insights into disease biology. Proteomic studies in COVID-19 have identified numerous severity-associated proteins involved in immune regulation, complement activation, coagulation pathways, and acute-phase responses (Messner et al., 2020).

Comparative proteomic analyses between mild and severe COVID-19 patients have consistently reported increased expression of inflammatory and immune-related proteins in severe cases (Shen et al., 2020). Despite its strengths, proteomics alone may not fully capture upstream regulatory events that drive protein expression changes. Therefore, integrating proteomic data with transcriptomic insights can enhance biological interpretation and biomarker discovery.

2.5 Multi-Omics Approaches in COVID-19 Research

Multi-omics approaches integrate multiple molecular datasets to provide a systems-level understanding of disease biology. In COVID-19 research, integrative analyses combining transcriptomics and proteomics have been increasingly used to investigate complex immune and inflammatory responses (Hasin et al., 2017). By integrating multiple omics layers, researchers can identify consistent molecular patterns across different biological levels, improving confidence in identified pathways and reducing false discoveries (Su et al., 2023).

2.6 Application of Machine Learning in Omics Studies

The high dimensionality and complexity of omics datasets pose significant analytical challenges. Machine learning techniques have emerged as powerful tools for extracting meaningful patterns from large biological datasets and prioritizing features associated with disease phenotypes (Libbrecht & Noble, 2015).

In COVID-19 research, machine learning has been applied to transcriptomic and proteomic data to classify disease severity, predict clinical outcomes, and identify important molecular features (Patel et al., 2023). Ensemble learning methods, particularly Random Forest, have shown strong performance due to their robustness, ability to handle high-dimensional data, and provision of interpretable feature importance measures (Breiman, 2001).

2.7 COVID-19 Severity and Post-Viral Outcomes

Clinical studies have shown that disease severity during acute SARS-CoV-2 infection is a major risk factor for prolonged recovery and post-viral complications. Individuals who experience severe COVID-19 are more likely to report persistent symptoms, including fatigue, respiratory dysfunction, and neurological impairments (Nalbandian et al., 2021). Although the molecular mechanisms linking acute disease severity to post-viral outcomes are not fully understood, sustained immune activation and molecular dysregulation are believed to play a key role.

2.8 Research Gap

Despite extensive research on COVID-19, several gaps remain. Many studies have independently investigated transcriptomic or proteomic alterations associated with COVID-19 severity, but integrative analyses combining these molecular layers with machine learning-based feature prioritization are still limited. Furthermore, while machine learning has been applied in COVID-19 research, its integration with proteomic data for severity classification and biomarker prioritization alongside transcriptomic insights has not been sufficiently explored. Addressing this gap requires an integrative multi-omics approach that combines transcriptomics, proteomics, and machine learning to identify robust molecular signatures associated with COVID-19 severity and their potential implications for post-viral outcomes.

CHAPTER 3: MATERIALS AND METHODS

3.1 Study Design and Overall Workflow

This study employed an integrative multi-omics framework to identify molecular signatures associated with COVID-19 severity. Publicly available transcriptomic and proteomic datasets were independently analysed and subsequently integrated to capture gene-level and protein-level alterations.

The overall workflow consisted of the following stages:

1. Selection of publicly available transcriptomic and proteomic datasets.

2. Transcriptomic preprocessing, normalization, differential expression analysis, and visualization through Principal Component Analysis (PCA), hierarchical clustering, and gene expression heatmaps.
3. Proteomic statistical filtering to identify severity-associated proteins (n = 91), followed by functional enrichment analysis.
4. Multi-omics integration through overlap analysis and RNA–protein fold-change correlation assessment.
5. Machine learning–based feature prioritization using Random Forest, evaluated through ROC analysis and Leave-One-Out Cross Validation (LOOCV).

This integrative approach enabled multi-layer validation of severity-associated molecular signatures and strengthened biological interpretation.

3.2 Transcriptomics Analysis

3.2.1 Transcriptomic Dataset Selection

Transcriptomic analysis was performed using a publicly available dataset obtained from the Gene Expression Omnibus (GEO) database. The dataset GSE213313 was selected for this study as it provides whole-blood gene expression profiles from COVID-19 patients with well-defined clinical classifications. The dataset includes samples from critical COVID-19 patients, non-critical COVID-19 patients, and healthy controls, making it suitable for severity-based differential expression analysis. The platform used for gene expression profiling was an Agilent microarray, which offers reliable and reproducible measurement of transcript-level expression across a large number of genes.

3.2.2 Data Preprocessing and Normalization

Raw microarray data files were imported into the R statistical environment for preprocessing and analysis. Background correction was performed to reduce technical noise and improve signal accuracy. Following background correction, data normalization was carried out using quantile normalization to ensure comparability of expression values across all samples. Log2 transformation of expression values was applied to stabilize variance and improve the statistical properties of the data. Quality control checks were conducted to confirm successful normalization and to identify any potential outlier samples before proceeding with differential expression analysis.

3.2.3 Differential Gene Expression Analysis

Differential gene expression analysis was performed to identify genes showing significant expression changes between COVID-19 patient groups. The analysis was carried out using the limma package in R, which is widely used for microarray-based differential expression studies. Linear models were fitted to the normalized expression data, and statistical significance was assessed using moderated t-statistics. Multiple testing correction was applied using the Benjamini–Hochberg false discovery rate (FDR) method to control for false positives. Genes with an adjusted p-value below the defined threshold (FDR < 0.05) and a biologically meaningful fold-change were considered significantly differentially expressed.

3.2.4 Functional Enrichment Analysis

To interpret the biological significance of the differentially expressed genes, functional enrichment analysis was performed. Gene Ontology (GO) enrichment analysis was used to identify overrepresented biological processes associated with the identified gene sets. Enrichment analysis enabled characterization of immune-related, inflammatory, and cellular pathways associated with COVID-19 severity. These results provided biological context to the transcriptomic findings and supported downstream integration with proteomic and machine learning analyses.

3.3 Proteomics Analysis

3.3.1 Proteomics Dataset Source

Proteomics data used in this study were obtained from a published, peer-reviewed COVID-19 proteomics study available through a public proteomics repository. The selected dataset compared protein expression profiles between mild and severe COVID-19 cases, making it appropriate for investigating severity-associated molecular signatures. Use of a published dataset ensured high data quality, validated experimental protocols, and reliable statistical processing (Aebersold & Mann, 2016). The dataset provided a curated list of proteins reported as significantly altered between clinical groups, along with associated statistical information (Shen et al., 2020).

3.3.2 Selection of Severity-Associated Proteins

Severity-associated proteins were selected based on statistical significance (Student's t-test, raw p-value < 0.05) between mild and severe COVID-19 groups. This filtering approach identified a total of 91 proteins significantly altered between clinical groups. Both upregulated and downregulated proteins were included to capture the full spectrum of severity-associated protein-level changes.

3.3.3 Functional Enrichment Analysis of Proteins

Functional enrichment analysis of severity-associated proteins was carried out using the clusterProfiler package in R. Gene Ontology (GO) enrichment analysis was conducted to identify overrepresented biological processes associated with the selected protein set. Enriched pathways related to immune response, inflammation, and host defense mechanisms were examined to understand molecular processes contributing to COVID-19 severity. Enrichment results were visualized using bar plots and dot plots to facilitate interpretation (Yu et al., 2012).

3.3.4 Rationale for Proteomics Analysis Strategy

Proteomics provides direct insight into functional molecular changes occurring during disease progression, as proteins represent the primary effectors of cellular processes (Aebersold & Mann, 2016). Focusing on enrichment analysis of statistically significant proteins allows meaningful biological interpretation even when raw mass spectrometry data are not available for reprocessing. This strategy has been widely used in COVID-19 proteomics studies to identify immune-related pathways, inflammatory responses, and host defense mechanisms associated with severe disease (Shen et al., 2020; Messner et al., 2020).

3.4 Machine Learning Analysis

3.4.1 Feature Matrix Preparation

Machine learning analysis was performed using the proteomics data to identify proteins most strongly associated with COVID-19 severity. The selected severity-associated proteins obtained from the proteomics analysis were used as input features. Each protein represented a feature, while the clinical severity status (mild or severe COVID-19) served as the target class label. Prior to model training, the dataset was checked for consistency and suitability for machine learning analysis.

3.4.2 Random Forest Model Selection

The Random Forest algorithm was selected for machine learning analysis due to its robustness, ability to handle high-dimensional biological data, and resistance to overfitting. Random Forest is an ensemble learning method that constructs multiple decision trees and combines their predictions to improve classification performance (Breiman, 2001). Random Forest has been widely applied in biomedical and omics studies for disease classification and biomarker discovery (Chen & Ishwaran, 2012).

3.4.3 Model Training and Evaluation

The Random Forest model was trained to classify samples into mild and severe COVID-19 groups based on proteomic features. The dataset was divided into training and testing sets to evaluate model performance. Model evaluation was carried out using standard classification metrics, including accuracy, confusion matrix-based assessment, and Receiver Operating Characteristic (ROC) analysis. Leave-One-Out Cross Validation (LOOCV) was additionally applied given the limited sample size (n = 8) to assess model robustness and reduce the risk of overfitting.

3.4.4 Feature Importance Analysis

One of the key advantages of the Random Forest algorithm is its ability to estimate feature importance based on each feature's contribution to classification performance. Feature importance scores were calculated using the Mean Decrease Gini index to identify proteins that played the most significant role in distinguishing between mild and severe COVID-19 cases. Proteins with higher importance scores were considered potential severity-associated biomarkers for further biological interpretation.

3.4.5 Integration with Omics Findings

The machine learning results were interpreted in conjunction with transcriptomic and proteomic enrichment analyses. Proteins identified as important by the Random Forest model were examined in the context of enriched biological pathways and immune-related processes. This integrative interpretation strengthened confidence in the identified molecular signatures and provided a multi-layered understanding of COVID-19 severity.

3.5 Software, Tools, and Statistical Criteria

All computational analyses in this study were performed using established bioinformatics and statistical software tools to ensure reproducibility and reliability. Data preprocessing, statistical analysis, functional enrichment, and machine learning were carried out primarily in the R statistical environment. Transcriptomic data analysis, including normalization and differential gene expression analysis, was performed using the limma package. Proteomic functional enrichment analysis was performed using the clusterProfiler package in R (Yu et al., 2012). Machine learning analysis was carried out using the Random Forest algorithm implemented through standard R packages.

For all statistical analyses, significance thresholds were applied as appropriate to each method. In transcriptomic analysis, multiple testing correction was performed using the Benjamini–Hochberg false discovery rate (FDR) method to control for false-positive results. In enrichment analyses, adjusted p-values were used to determine statistically significant pathways. A significance threshold of $p < 0.05$ was considered statistically meaningful unless otherwise specified.

CHAPTER 4: RESULTS

4.1 Transcriptomics Results

4.1.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) was performed to evaluate global transcriptional variation across control, critical, and non-critical COVID-19 samples. The PCA plot demonstrated clear separation of critical COVID-19 samples from healthy controls along the first principal component (PC1), which explained 35.3% of the total variance. The second principal component (PC2) accounted for 11.5% of the variance. Critical samples clustered distinctly from controls, indicating substantial transcriptional reprogramming associated with severe disease.

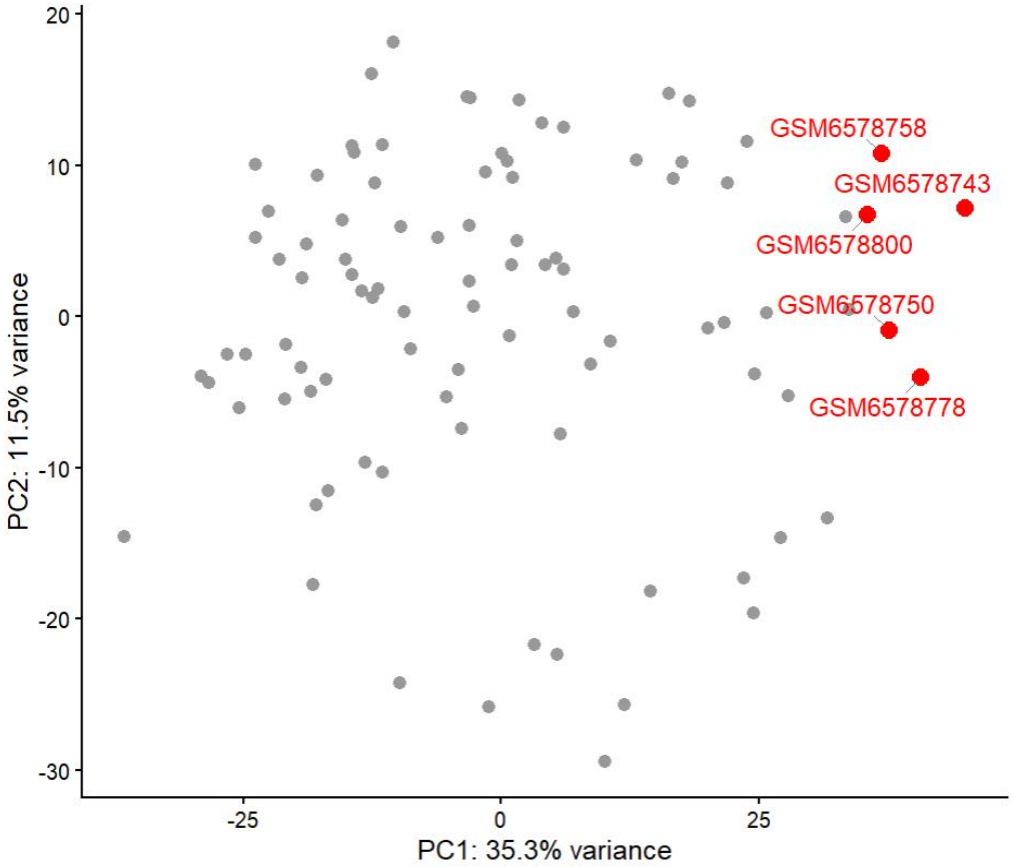

Figure 4.1: Principal Component Analysis (PCA) plot showing separation of critical COVID-19 samples from healthy controls based on global gene expression profiles.

4.1.2 Hierarchical Clustering Analysis

Hierarchical clustering was performed using expression profiles of significantly differentially expressed genes to examine similarity patterns among samples. The dendrogram revealed distinct clustering of critical and healthy samples, further confirming robust separation between disease groups. Samples within the same clinical category clustered together, reflecting coordinated regulation of immune-related genes in critical COVID-19 cases.

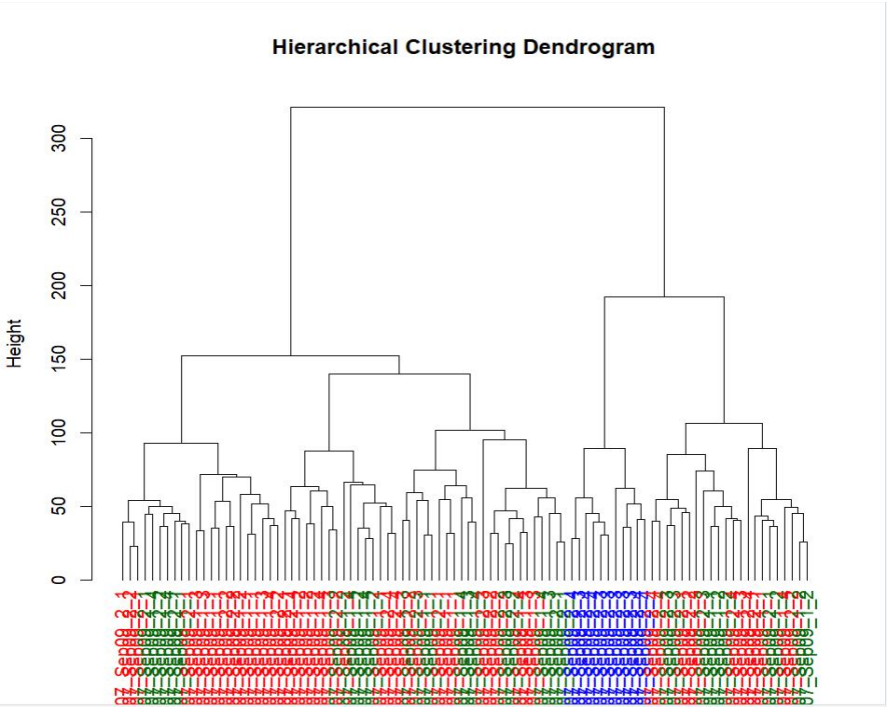

Figure 4.2: Hierarchical clustering dendrogram based on significantly differentially expressed genes demonstrating distinct grouping of critical and healthy samples.

4.1.3 Differential Gene Expression Analysis

Differential gene expression analysis was conducted using the limma package in R. Genes were considered significantly differentially expressed based on an adjusted p-value threshold (FDR < 0.05) combined with log2 fold-change criteria. A total of 1,544 genes were identified as significantly differentially expressed in critical COVID-19 samples compared to healthy controls. The volcano plot illustrates significantly upregulated and downregulated genes, with a large number exhibiting significant upregulation in critical samples, indicating strong activation of immune and inflammatory pathways.

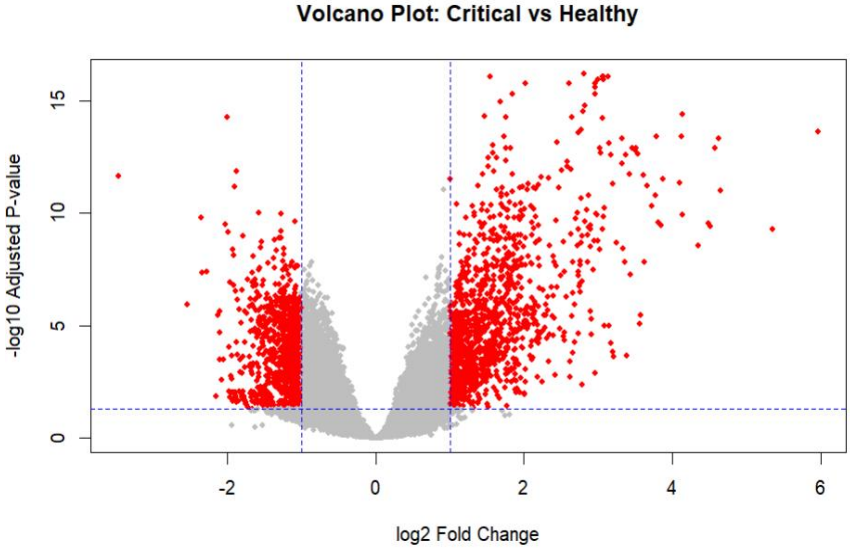

Figure 4.3: Volcano plot showing significantly upregulated and downregulated genes (Critical vs. Healthy) based on FDR-adjusted p-values and log2 fold-change thresholds.

4.1.4 Heatmap of Top Differentially Expressed Genes

To further visualize expression patterns, a heatmap was generated using the top 50 differentially expressed genes. The heatmap demonstrates clear separation between control, critical, and non-critical groups, with distinct clustering of critical samples. Coordinated upregulation and downregulation patterns are evident, particularly among immune-related genes, supporting the robustness of the differential expression findings.

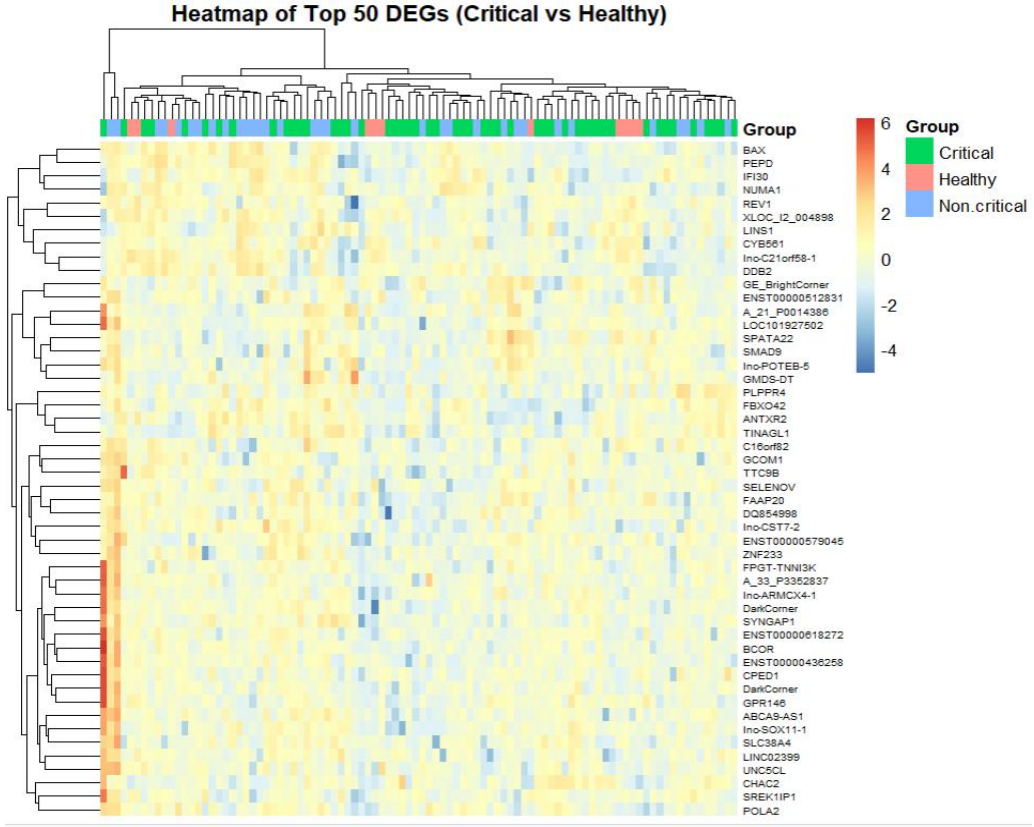

Figure 4.4: Heatmap of top 50 differentially expressed genes illustrating distinct clustering of critical and control samples.

4.1.5 Expression Visualization of Selected Top Genes

Boxplots were generated for selected top differentially expressed genes to validate expression differences across groups. The plots demonstrate consistent expression variation between control, critical, and non-critical samples. Several genes exhibited markedly increased expression in critical COVID-19 cases, supporting their potential role in disease severity and immune activation.

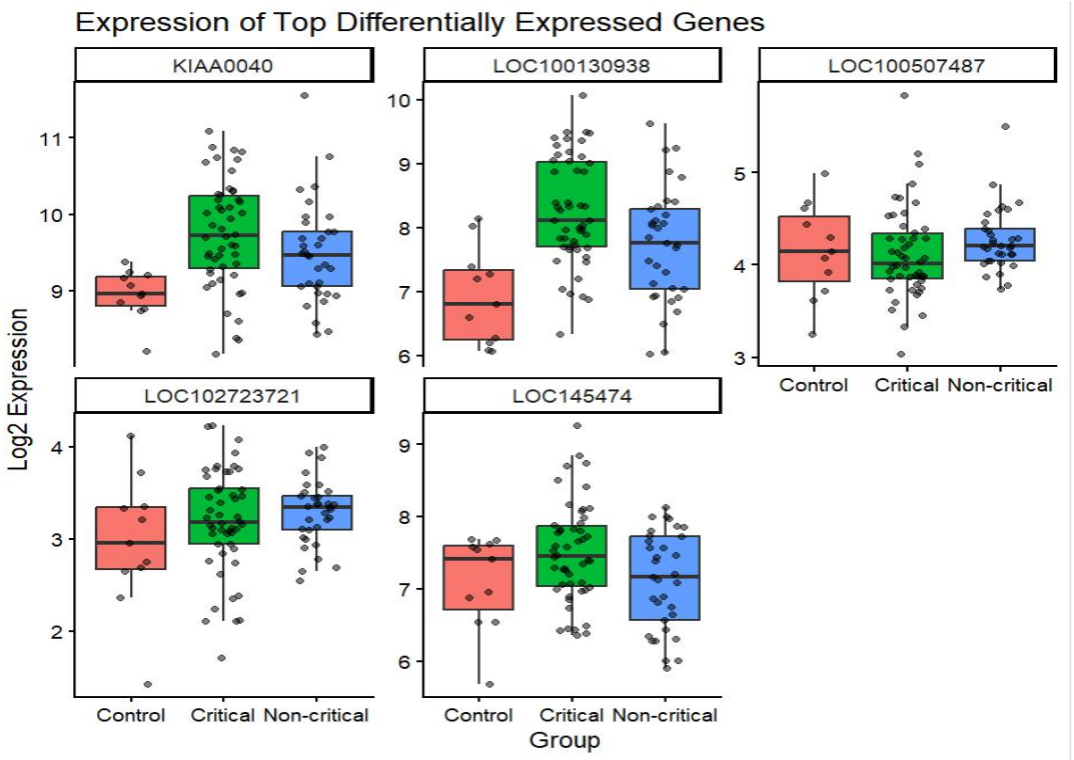

Figure 4.5: Boxplots showing expression levels of selected top differentially expressed genes across control, critical, and non-critical groups.

4.1.6 Functional Enrichment Analysis of Differentially Expressed Genes

Functional enrichment analysis was performed to interpret the biological significance of the differentially expressed genes identified between critical COVID-19 patients and healthy controls. Gene Ontology (GO) Biological Process enrichment revealed significant overrepresentation of immune-related processes, including positive regulation of innate immune response, leukocyte cell–cell adhesion, regulation of immune effector processes, T-cell differentiation, and cell killing. Enrichment of antigen processing and presentation pathways further indicates activation of adaptive immune mechanisms in severe disease.

In addition to GO analysis, KEGG pathway enrichment analysis identified significantly enriched signalling pathways including Coronavirus disease (COVID-19), hematopoietic cell lineage, Th17 cell differentiation, cell adhesion molecule (CAM) interactions, antigen processing and presentation, and inflammatory bowel disease. The enrichment of these immune-related and inflammatory pathways highlights coordinated dysregulation of both innate and adaptive immune responses in critically ill patients.

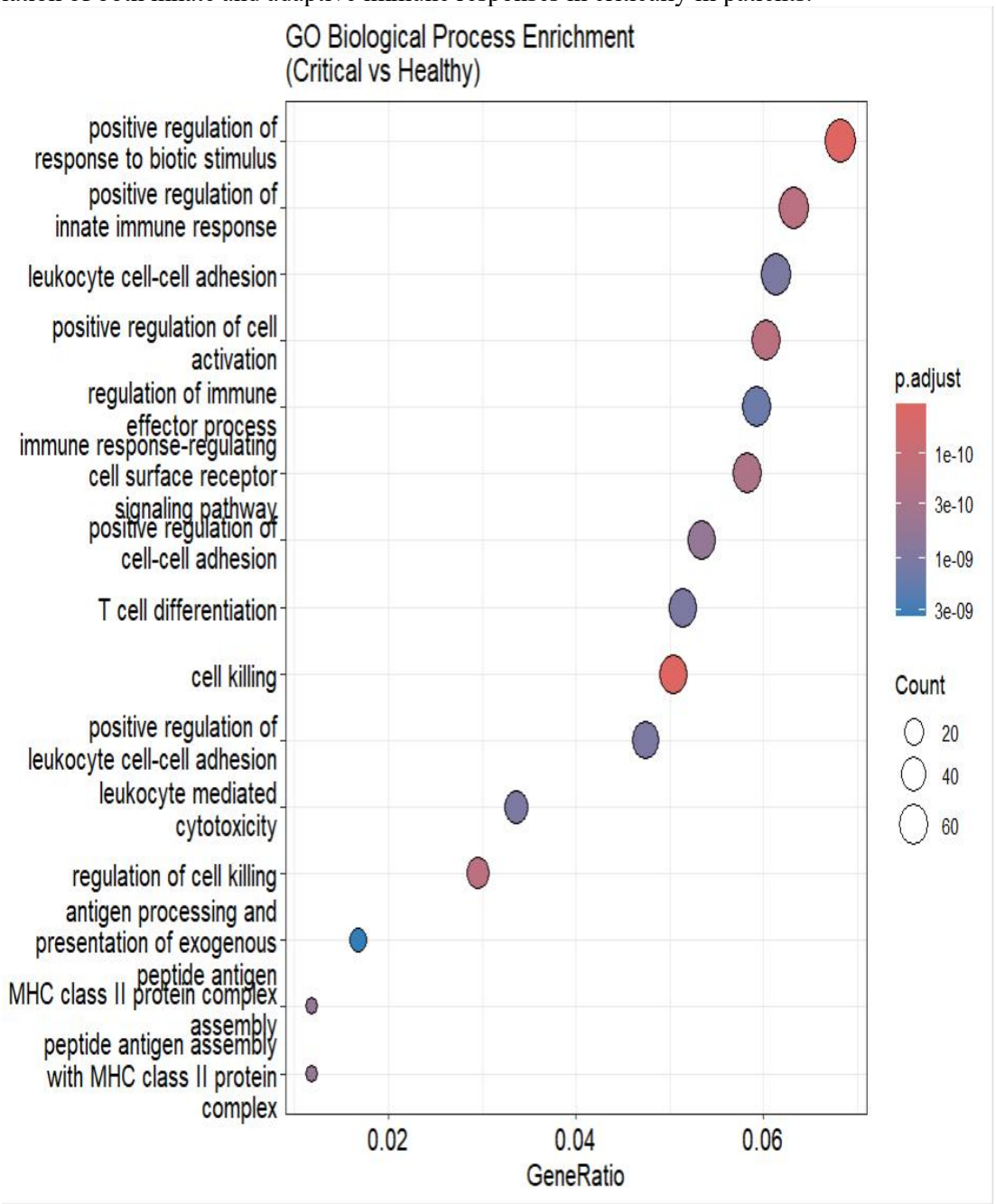

Figure 4.6: Gene Ontology (Biological Process) enrichment analysis of significantly differentially expressed genes (Critical vs. Healthy).

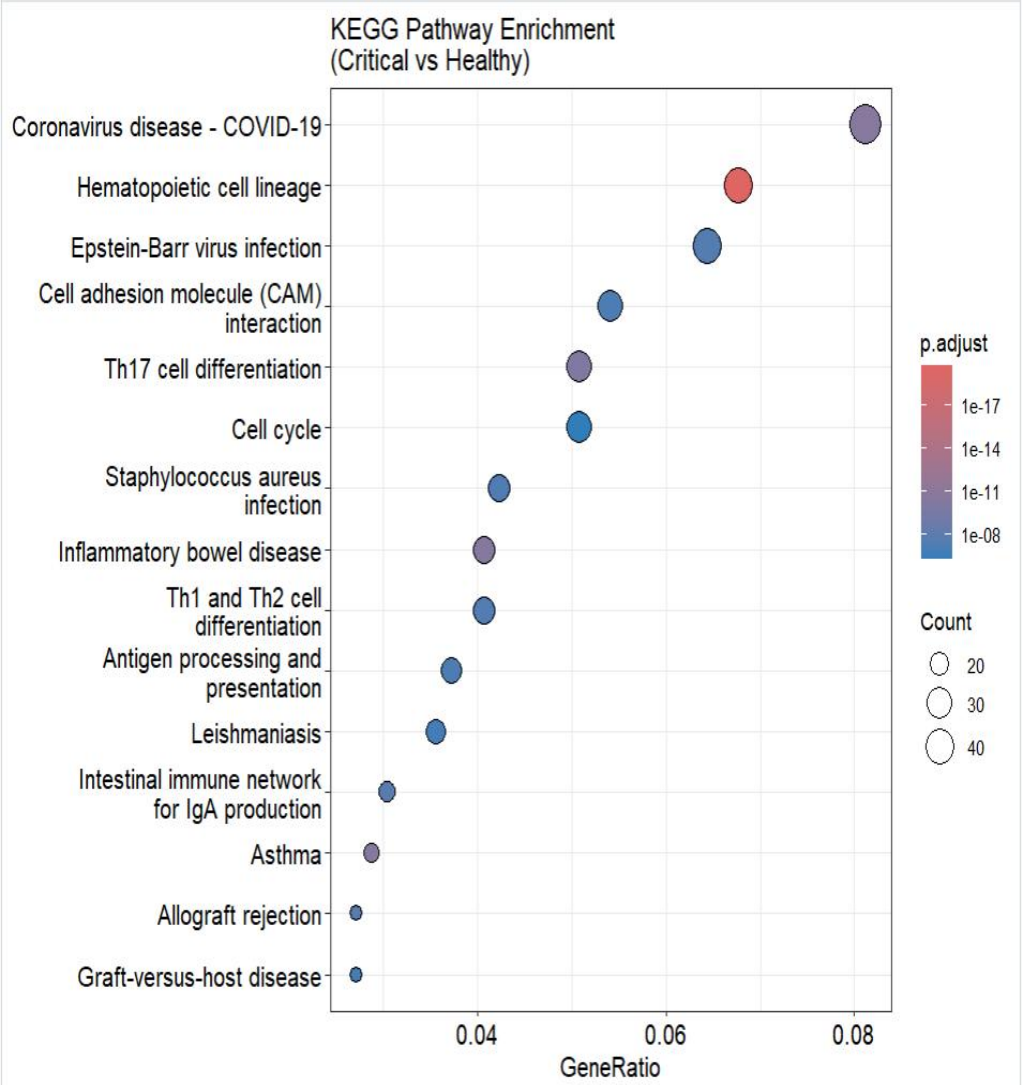

Figure 4.7: KEGG pathway enrichment analysis of significantly differentially expressed genes (Critical vs. Healthy).

4.2 Proteomics Results

4.2.1 Identification of Severity-Associated Proteins

Proteomic analysis was conducted using a publicly available dataset comparing mild (n = 3) and severe (n = 5) COVID-19 cases. Statistical filtering based on Student's t-test (raw p-value < 0.05) identified a total of 91 proteins significantly altered between severity groups. Both upregulated and downregulated proteins were observed in severe COVID-19 cases, indicating widespread protein-level dysregulation associated with disease progression.

Upregulated proteins in severe cases included VAPA, CA2, CD48, GSTO1, ATIC, FCGR3A, and others, many of which are associated with immune activation and inflammatory responses. Conversely, downregulated proteins such as PDE4DIP, CLEC11A, ASPN, FAIM3, LIPG, KDR, and BST1 suggest altered immune regulation and structural remodelling in severe disease. Expression values of the top 15 upregulated and top 15 downregulated proteins are presented in Table 4.1 and Table 4.2, respectively.

Table 4.1: Expression values of top 15 upregulated proteins in mild (n = 3) and severe (n = 5) COVID-19 samples.

Sl.No	Protein	Mild 1	Mild 2	Mild 3	Severe 1	Severe 2	Severe 3	Severe 4	Severe 5
1)	VAPA	9.29	10.79	10.93	16.37	15.68	16.23	16.55	20.63
2)	CA2	11.39	9.685	18.96	20.65	19.09	20.27	17.59	19.81
3)	CD48	10.13	9.11	9.86	18.47	13.09	18.45	17.73	10.53
4)	GSTO1	17.57	10.36	16.34	19.18	19.87	19.44	19.62	20.65
5)	ATIC	14.48	10.37	17.57	18.96	17.30	19.39	19.38	19.24

Sl.No	Protein	Mild 1	Mild 2	Mild 3	Severe 1	Severe 2	Severe 3	Severe 4	Severe 5
6)	FAM20C	13.53	10.93	10.89	16.59	16.19	17.68	17.50	13.57
7)	SEMA3F	11.94	14.06	10.06	17.15	15.24	17.63	14.08	18.27
8)	SCLY	14.09	11.40	9.66	16.64	17.46	17.44	12.90	16.41
9)	FCGR3A	13.06	19.65	18.71	22.37	20.96	21.51	20.18	22.85
10)	ALDH7A1	11.60	11.08	12.78	16.33	18.39	14.18	17.00	14.61
11)	LSAMP	12.56	11.57	12.32	14.99	18.49	13.81	17.28	17.46
12)	EPS15L1	10.15	12.46	9.16	17.08	14.00	14.63	15.63	12.09
13)	PCSK1N	12.57	12.27	11.89	17.75	17.95	15.53	13.92	15.62
14)	IQGAP2	14.56	11.38	12.34	13.89	17.43	17.27	16.11	18.63
15)	MAT2B	10.40	13.34	11.94	16.90	16.74	12.96	16.81	18.63

Table 4.2: Expression values of top 15 downregulated proteins in mild (n = 3) and severe (n = 5) COVID-19 samples.

Sl.No	Protein	Mild 1	Mild 2	Mild 3	Severe 1	Severe 2	Severe 3	Severe 4	Severe 5
1)	PDE4DIP	20.69	21.52	23.54	10.83	8.059	12.91	7.57	11.32
2)	CLEC11A	17.67	16.22	17.01	10.22	10.26	12.30	9.51	8.656
3)	ASPN	20.58	16.17	16.32	10.33	13.38	14.53	8.388	10.94
4)	BASP1	18.29	21.31	19.71	13.32	14.32	13.40	13.16	14.21
5)	FAIM3	18.61	17.90	18.66	12.82	14.04	11.61	12.76	14.37
6)	DEFA3	23.55	22.36	22.44	19.86	14.69	18.19	16.71	19.04
7)	ENTPD5	17.69	18.08	17.79	13.50	11.32	13.52	13.12	15.16
8)	LIPG	15.58	15.61	16.01	12.01	12.46	9.58	10.10	13.37
9)	CCS	13.05	17.62	17.30	12.00	12.88	10.85	10.93	14.01
10)	KDR	16.48	16.65	16.60	16.19	13.00	10.25	11.57	13.45
11)	SEMA3A	14.18	16.29	13.62	11.48	9.49	11.07	11.77	11.70
12)	BST1	18.40	16.81	18.05	15.70	13.15	15.98	12.37	16.28
13)	PGAM1	19.70	19.03	20.79	17.79	15.65	17.18	18.64	17.20
14)	SIRPB1	20.88	22.07	21.12	19.99	17.80	19.31	19.79	17.38
15)	HMCN2	15.04	14.03	15.41	10.88	13.95	10.94	13.75	12.15

4.2.2 Functional Enrichment Analysis of Severity-Associated Proteins

Functional enrichment analysis was performed separately for upregulated and downregulated proteins using the clusterProfiler package in R. Enrichment results were visualized using multiple annotation databases, including Gene Ontology (GO), KEGG, and Reactome.

GO Biological Process enrichment analysis revealed significant enrichment of acute-phase response, acute inflammatory response, blood coagulation, hemostasis, and response to stimulus pathways. These results indicate heightened inflammatory and coagulation activity in severe COVID-19 cases.

Figure 4.8: GO Biological Process enrichment analysis of severity-associated proteins showing significant enrichment of acute inflammatory response and coagulation-related processes.

GO Cellular Component enrichment demonstrated significant overrepresentation of extracellular region, extracellular space, blood microparticle, endoplasmic reticulum lumen, and endomembrane system components. This suggests enhanced extracellular secretion and systemic immune signalling in severe disease.

Figure 4.9: GO Cellular Component enrichment analysis of severity-associated proteins highlighting enrichment of extracellular and secretory components.

4.2.3 Biological Interpretation of Proteomic Findings

Collectively, proteomic results indicate that severe COVID-19 is characterized by activation of acute inflammatory pathways, complement cascade amplification, platelet activation and coagulation dysregulation, and alterations in extracellular and secretory protein networks. These findings align with clinical manifestations of severe COVID-19, including hyperinflammation, endothelial dysfunction, and thrombotic complications. Importantly, the protein-level alterations complement transcriptomic findings, supporting multi-layer validation of severity-associated molecular mechanisms.

4.3 Multi-Omics Integration

4.3.1 Overlap Between Transcriptomic and Proteomic Signatures

To identify shared molecular signatures across omics layers, overlap analysis was performed between significantly differentially expressed genes (transcriptomics) and severity-associated proteins (proteomics). A total of 8 overlapping molecules were identified between the transcriptomic (n = 1,544) and proteomic (n = 78) datasets. These overlapping molecules represent cross-platform validated candidates potentially involved in COVID-19 severity mechanisms.

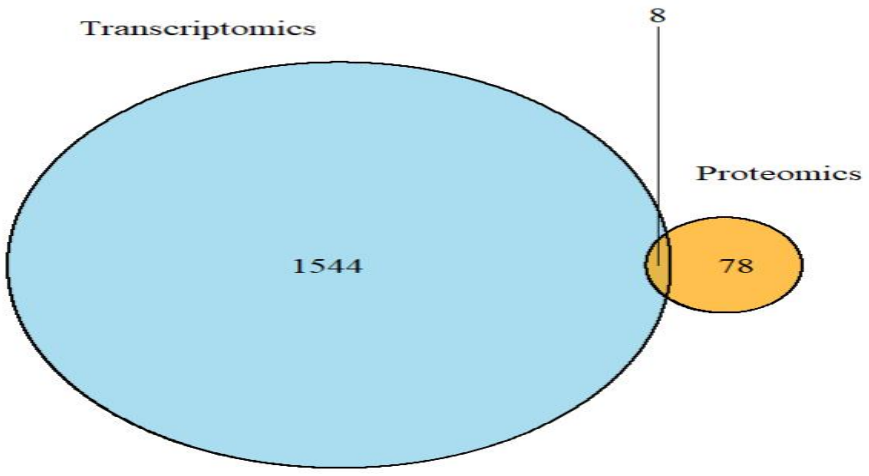

Figure 4.10: Venn diagram illustrating overlap between transcriptomic (n = 1,544) and proteomic (n = 78) severity-associated molecules, identifying 8 shared candidates.

4.3.2 Correlation Between Transcriptomic and Proteomic Changes

To evaluate concordance between RNA-level and protein-level changes, correlation analysis was performed using log2 fold-change values of overlapping molecules. An inverse correlation trend was observed between transcriptomic and proteomic fold changes ($r = -0.58$), suggesting possible post-transcriptional regulatory mechanisms influencing protein abundance. This moderate inverse relationship highlights the complexity of multi-layer molecular regulation and underscores the importance of integrative analysis.

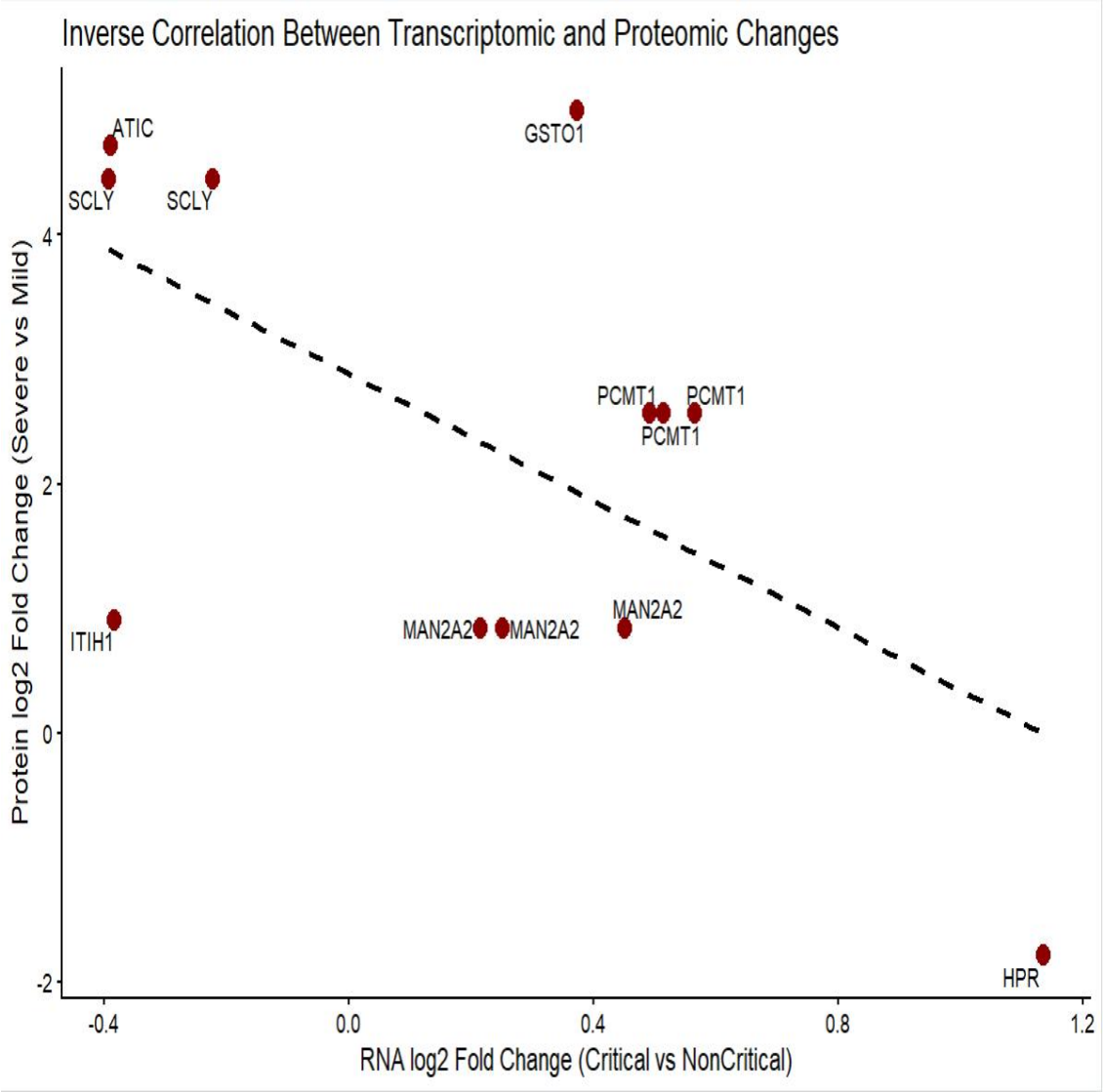

Figure 4.11: Scatter plot illustrating inverse correlation between RNA log2 fold change and protein log2 fold change for overlapping molecules.

4.4 Machine Learning Results

4.4.1 Random Forest Classification Performance

To identify proteomic features most strongly associated with COVID-19 severity, a Random Forest classifier was developed using severity-associated protein expression values as predictor variables and clinical severity (mild vs. severe) as the response variable. The model demonstrated perfect discrimination between mild and severe COVID-19 samples. Receiver Operating Characteristic (ROC) analysis yielded an Area Under the Curve (AUC) of 1.0, indicating complete separation of the two severity groups.

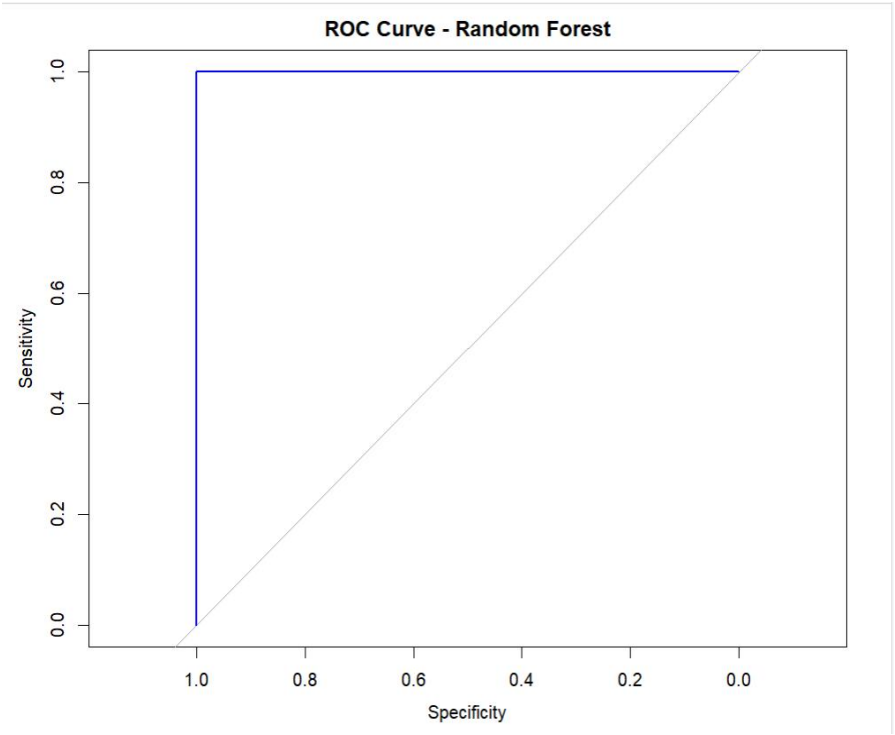

Figure 4.12: Receiver Operating Characteristic (ROC) curve of the Random Forest model distinguishing mild and severe COVID-19 cases (AUC = 1.0).

4.4.2 Leave-One-Out Cross Validation (LOOCV)

Given the limited sample size (n = 8), Leave-One-Out Cross Validation (LOOCV) was applied to assess model robustness and reduce the risk of overfitting. LOOCV produced a classification accuracy of 100% (Accuracy = 1.0) with a Kappa statistic of 1.0, indicating perfect agreement between predicted and actual class labels across all validation iterations. These results suggest strong separability of mild and severe cases within the proteomic dataset; however, results should be interpreted with caution given the small sample size.

4.4.3 Feature Importance Analysis

Feature importance analysis was performed using the Mean Decrease Gini index to identify proteins contributing most strongly to severity classification. The Random Forest model prioritized the following proteins as key discriminative features distinguishing mild and severe COVID-19 cases:

- LIPG (Lipase G, Endothelial Type)
- CLEC11A (C-Type Lectin Domain Family 11 Member A)
- KDR (Kinase Insert Domain Receptor / VEGFR2)
- FAIM3 (Fas Apoptotic Inhibitory Molecule 3)
- BST1 (Bone Marrow Stromal Cell Antigen 1)

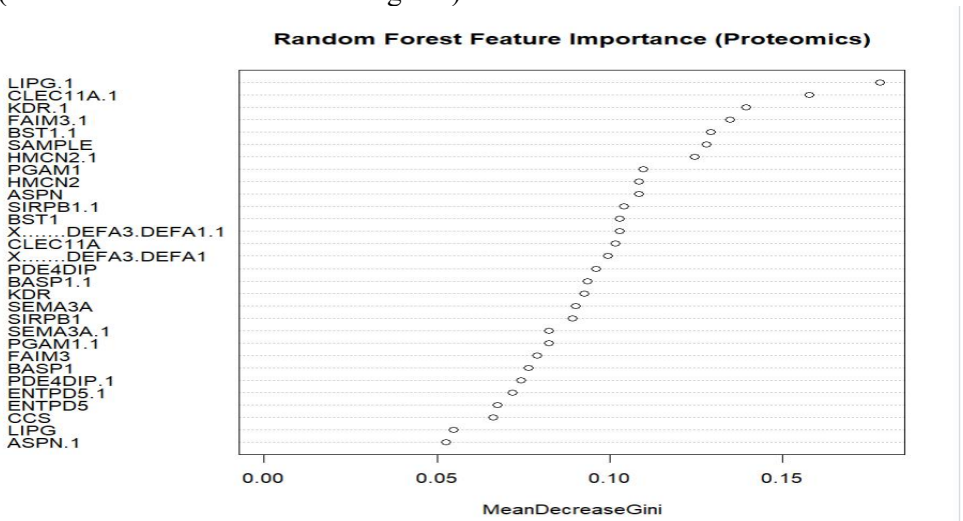

Figure 4.12: Random Forest feature importance ranking based on Mean Decrease Gini values for severity-associated proteomic features.

CHAPTER 5: DISCUSSION

5.1 Transcriptomic Alterations in Severe COVID-19

The transcriptomic analysis of whole blood samples revealed substantial gene expression changes associated with COVID-19 severity. Principal Component Analysis demonstrated clear separation between critical and healthy samples, indicating global transcriptional reprogramming in severe disease. Differential expression analysis identified 1,544 significantly altered genes (FDR < 0.05), highlighting extensive immune activation.

Functional enrichment analysis revealed strong overrepresentation of innate immune response, leukocyte activation, antigen processing, T-cell differentiation, and cytokine-mediated pathways. These findings are consistent with the well-documented hyperinflammatory state observed in severe COVID-19 and support the presence of systemic immune dysregulation. The enrichment of interferon signalling and immune effector pathways suggests that critical cases are characterized by heightened antiviral and inflammatory responses, which may contribute to immunopathology and disease progression.

5.2 Proteomic Signatures of Disease Severity

Proteomic analysis identified 91 severity-associated proteins, with significant enrichment of inflammatory and coagulation-related processes. The overrepresentation of acute-phase response, complement cascade activation, platelet degranulation, and hemostasis pathways highlights the protein-level manifestation of hyperinflammation and thrombotic complications in severe COVID-19.

Enrichment within extracellular region and blood microparticle components suggests enhanced secretion of inflammatory mediators and systemic immune activation. These findings align with clinical observations of cytokine storm, endothelial dysfunction, and coagulopathy in critically ill patients. Importantly, proteomic alterations complement transcriptomic findings by demonstrating that immune and inflammatory dysregulation observed at the RNA level is reflected at the protein level, reinforcing the biological relevance of identified molecular signatures.

5.3 Multi-Omics Integration and Cross-Platform Validation

Integration of transcriptomic (n = 1,544 genes) and proteomic (n = 78 mapped proteins) datasets identified 8 overlapping molecules. These shared candidates represent cross-platform validated severity-associated markers. The limited overlap between RNA and protein datasets is biologically expected due to post-transcriptional regulation, differences in protein stability, translation efficiency, and technical detection limitations.

The observed inverse correlation trend ($r = -0.58$) between RNA and protein fold changes further underscores the complexity of gene–protein regulation in severe COVID-19. These findings highlight the importance of integrative multi-omics approaches in capturing complementary molecular information across regulatory layers.

5.4 Machine Learning-Based Feature Prioritization

Random Forest classification analysis demonstrated perfect discrimination between mild and severe cases (AUC = 1.0), with Leave-One-Out Cross Validation yielding 100% accuracy (Kappa = 1.0). These results indicate strong separability within the proteomic dataset. It is important to note, however, that perfect classification performance (AUC = 1.0) with a sample size of only n = 8 is likely influenced by the small cohort size and may represent overfitting. LOOCV at this scale does not reliably prevent this, and results should be considered preliminary until validated in larger, independent cohorts. The Random Forest model was therefore primarily applied as a feature prioritization framework rather than a predictive clinical tool.

Feature importance analysis identified proteins such as LIPG, CLEC11A, KDR, FAIM3, and BST1 as key contributors to severity classification. Many of these proteins are functionally linked to immune regulation, vascular biology, and inflammatory signalling pathways identified in enrichment analyses, providing concordance between statistical prioritization and biological interpretation.

5.5 Biological Implications and Clinical Relevance

The combined transcriptomic, proteomic, and machine learning analyses collectively demonstrate that severe COVID-19 is characterized by coordinated immune activation, complement cascade amplification, platelet activation, and coagulation dysregulation. The identification of cross-platform validated molecules and machine learning-prioritized features provides potential candidates for biomarker development and therapeutic targeting. Integrative multi-omics analysis enhances understanding of disease mechanisms and highlights the value of combining molecular layers to improve severity stratification.

5.6 Limitations of the Study

Several limitations should be acknowledged. First, the proteomic dataset used for machine learning had a small sample size (n = 8), which limits the generalizability of classifier performance. Second, the transcriptomic and proteomic datasets were derived from different cohorts, which may introduce biological variability in integration analyses. Third, the overlap analysis identified only 8 shared molecules, which may partly reflect differences in detection sensitivity between platforms. Finally, all findings are computational and require experimental validation in larger, well-characterized patient cohorts before clinical application.

CHAPTER 6: CONCLUSION

6.1 Summary of Findings

The present study employed an integrative bioinformatics framework combining transcriptomic analysis, proteomic enrichment profiling, and machine learning-based feature prioritization to investigate molecular signatures associated with COVID-19 severity.

Transcriptomic analysis of the GSE213313 whole-blood dataset identified 1,544 significantly differentially expressed genes (FDR < 0.05), revealing substantial transcriptional reprogramming in severe COVID-19 cases. Enrichment analysis highlighted immune response, cytokine-mediated signalling, leukocyte activation, antigen processing, and inflammatory pathways, indicating pronounced immune dysregulation in critically ill patients.

Proteomic analysis identified 91 severity-associated proteins, with enrichment of acute-phase response, complement cascade activation, platelet degranulation, extracellular region components, and coagulation-related pathways. These findings reflect protein-level manifestations of systemic inflammation and thrombotic complications commonly observed in severe COVID-19.

Machine learning analysis using a Random Forest classifier enabled prioritization of proteomic features contributing to severity discrimination. Proteins including LIPG, CLEC11A, KDR, FAIM3, and BST1 demonstrated higher importance scores, supporting their potential relevance as severity-associated molecular candidates.

6.2 Overall Conclusion

The integrative multi-omics analysis demonstrates that severe COVID-19 is characterized by coordinated dysregulation of immune activation, inflammatory signalling, complement cascade amplification, and coagulation mechanisms across multiple molecular layers. The convergence of transcriptomic enrichment results, proteomic pathway analysis, and machine learning-based feature ranking strengthens confidence in the identified severity-associated molecular signatures.

6.3 Study Significance

This study highlights the value of integrating publicly available transcriptomic and proteomic datasets with machine learning methodologies to identify biologically meaningful molecular signatures associated with disease severity. The systematic multi-omics framework adopted in this work demonstrates how computational approaches can be used to prioritize candidate biomarkers and uncover mechanistic pathways in infectious diseases. Although experimental validation was beyond the scope of this study, the identified proteins and pathways provide promising targets for future investigation in larger and independent clinical cohorts.

6.4 Final Statement

In conclusion, the integrative transcriptomic, proteomic, and machine learning approach applied in this study successfully identified key molecular alterations associated with COVID-19 severity. The findings emphasize the central role of immune dysregulation, inflammatory responses, and coagulation pathway activation in severe disease and establish a computational foundation for future translational and clinical research aimed at biomarker development and therapeutic targeting.

REFERENCES

1. Aebersold, R., & Mann, M. (2016). Mass-spectrometric exploration of proteome structure and function. Nature, 537(7620), 347–355.
2. Blanco-Melo, D., Nilsson-Payant, B. E., Liu, W. C., Uhl, S., Hoagland, D., Møller, R., ... & tenOever, B. R. (2020). Imbalanced host response to SARS-CoV-2 drives development of COVID-19. Cell, 181(5), 1036–1045.
3. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
4. Chen, X., & Ishwaran, H. (2012). Random forests for genomic data analysis. Genomics, 99(6), 323–329.

5. Hadjadj, J., Yatim, N., Barnabei, L., Corneau, A., Boussier, J., Smith, N., ... & Terrier, B. (2020). Impaired type I interferon activity and inflammatory responses in severe COVID-19 patients. Science, 369(6504), 718–724.
6. Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. Genome Biology, 18(1), 1–15.
7. Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., ... & Cao, B. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The Lancet, 395(10223), 497–506.
8. Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. Nature Reviews Genetics, 16(6), 321–332.
9. Liu, Y., Beyer, A., & Aebersold, R. (2016). On the dependency of cellular protein levels on mRNA abundance. Cell, 165(3), 535–550.
10. Lucas, C., Wong, P., Klein, J., Castro, T. B., Silva, J., Sundaram, M., ... & Iwasaki, A. (2020). Longitudinal analyses reveal immunological misfiring in severe COVID-19. Nature, 584(7821), 463–469.
11. Messner, C. B., Demichev, V., Wendisch, D., Michalick, L., White, M., Freiwald, A., ... & Ralser, M. (2020). Ultra-high-throughput clinical proteomics reveals classifiers of COVID-19 infection. Cell Systems, 11(1), 11–24.
12. Nalbandian, A., Sehgal, K., Gupta, A., Madhavan, M. V., McGroder, C., Stevens, J. S., ... & Wan, E. Y. (2021). Post-acute COVID-19 syndrome. Nature Medicine, 27(4), 601–615.
13. Patel, M. A., Knauer, M. J., Nicholson, M., Daley, M., Van Nynatten, L., Martin, C., ... & Cepinskas, G. (2023). Elevated anti-inflammatory proteins and reduced alarmin-related proteins are associated with ICU COVID-19 deaths. Critical Care Medicine, 51(5), 582–594.
14. Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research, 43(7), e47.
15. Rung, J., & Brazma, A. (2013). Reuse of public genome-wide gene expression data. Nature Reviews Genetics, 14(2), 89–99.
16. Shen, B., Yi, X., Sun, Y., Bi, X., Du, J., Zhang, C., ... & Hou, Y. (2020). Proteomic and metabolomic characterization of COVID-19 patient sera. Cell, 182(1), 59–72.
17. Su, Y., Yuan, D., Chen, D. G., Ng, R. H., Wang, K., Choi, J., ... & Heath, J. R. (2023). Multiple early factors anticipate post-acute COVID-19 sequelae. Cell, 185(5), 881–895.
18. Tang, N., Li, D., Wang, X., & Sun, Z. (2020). Abnormal coagulation parameters are associated with poor prognosis in patients with novel coronavirus pneumonia. Journal of Thrombosis and Haemostasis, 18(4), 844–847.
19. Tay, M. Z., Poh, C. M., Rénia, L., MacAry, P. A., & Ng, L. F. (2020). The trinity of COVID-19: immunity, inflammation and intervention. Nature Reviews Immunology, 20(6), 363–374.
20. World Health Organization. (2020). COVID-19: Case definitions. WHO Reference Number: WHO/2019-nCoV/Surveillance_Case_Definition/2020.2.
21. Yu, G., Wang, L. G., Han, Y., & He, Q. Y. (2012). clusterProfiler: an R Package for comparing biological themes among gene clusters. OMICS: A Journal of Integrative Biology, 16(5), 284–287.
22. Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., ... & Cao, B. (2020). Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. The Lancet, 395(10229), 1054–1062.