Integrated Bioinformatics Exploration and Preliminary Clinical Verification for the Identification of Crucial Biomarkers in Severe Cases of COVID-19

Zhisheng Huang1,2, Zuowang Cheng3, Xia Deng4, Ying Yang5, Na Sun5, Peibin Hou5, Ruyue Fan5, Shuai Liu6,7

1Department of Pulmonary and Critical Care Medicine, The First Affiliated Hospital of Nanchang University, Nanchang, Jiangxi, People’s Republic of China; 2Department of Pulmonary and Critical Care Medicine, National Regional Center for Respiratory Medicine, Jiangxi Hospital of China-Japan Friendship Hospital, Nanchang, Jiangxi, People’s Republic of China; 3Department of Clinical Laboratory, Zhanjiang District People’s Hospital Affiliated to Jining Medical University, Jining, Shandong, People’s Republic of China; 4School of Public Health, Shandong Second Medical University, Weifang, Shandong, People’s Republic of China; 5Shandong Center for Disease Control and Prevention, Jinan, Shandong, People’s Republic of China; 6Department of Pulmonary and Critical Care Medicine, National Regional Center for Respiratory Medicine, Jiangxi Hospital of China-Japan Friendship Hospital, Nanchang, Jiangxi, People’s Republic of China; 7Department of Pulmonary and Critical Care Medicine, Shandong Provincial Hospital Affiliated to Shandong First Medical University, Jinan, Shandong, People’s Republic of China

Correspondence: Shuai Liu; Ruyue Fan, Email liushuai_sam@163.com; fanruyue2022@163.com

Background: Coronavirus disease 2019 (COVID-19) is a respiratory infectious illness caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The objective of this study is to identify reliable and accurate biomarkers for the early stratification of disease severity, a crucial aspect that is currently lacking for the impending phases of the next COVID-19 pandemic.

Methods: In this study, we identified important module and hub genes related to clinical severe COVID-19 using differentially expressed genes (DEGs) screening combing weighted gene co-expression network analysis (WGCNA) in dataset GSE213313. We further screened and confirmed these hub genes in another two new independent datasets (GSE172114 and GSE157103). In order to evaluate these key genes’ stability and robustness for diagnosing or predicting the progression of illness, we used RT-PCR validation of selected genes in blood samples obtained from hospitalized COVID-19 patients.

Results: A total of 968 and 52 DEGs were identified between COVID-19 patients and normal people, critical and non-critical patients, respectively. Then, using WGCNA, 10 modules were constructed. Among them, the blue module positively associated with clinic disease severity of COVID-19. From overlapped section between DEGs and blue module, 12 intersected common differential genes were obtained. Subsequently, these hub genes were validated in another two new independent datasets as well and 9 genes that overlapped showed a highly correlation with disease severity. Finally, the mRNA expression levels of these hub genes were tested in blood samples from COVID-19 patients. In severe cases, there was increased expression of MCEMP1, ANXA3, CD177, and SCN9A. In particular, MCEMP1 increased with disease severity, which suggested an unfavorable development and a frustrating prognosis.

Conclusion: Using comprehensive bioinformatical analysis and the validation of clinical samples, we identified four major candidate genes, MCEMP1, ANXA3, CD177, and SCN9A, which are essential for diagnosis or development of COVID-19.

Keywords: COVID-19, differentially expressed genes, WGCNA, hub genes, neutrophil

Introduction

Over the past four years, coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2), is a worldwide health concern. As of 8 November 2023, there have been 771 million reported cases in 188 countries and territories, resulting in over 6.97 million deaths.1 The disease can manifest in various degrees of severity, ranging from mild to critical. Moreover, emerging evidence suggests that individuals infected with SARS-CoV-2 can be asymptomatic carriers, contributing to the silent spread of the virus.2,3 Despite extensive investigation, the underlying question of why there is such a marked difference in the disease progression among infected individuals remains unresolved.
Current studies highlight that the illness severity and progression are largely determined by the dysregulated host inflammatory response.\(^4\)\(^-\)\(^6\) When a virus penetrates a cell, it may activate both innate and adaptive host immune responses, which are critical components of viral defense.\(^7\)\(^-\)\(^8\) The exaggerated or dysregulated inflammatory responses and an imbalanced adaptive immunity may account for severe tissue destruction.\(^9\) The preexisting immunity on the response to viral infection include the excessive production of inflammatory cytokines, overactivation and dysregulated of immune cells, profound T lymphopenia, as well as decreased frequencies of dendritic cells and natural killer cells, complement activation.\(^7\)\(^-\)\(^10\)

A comprehensive understanding of the host antiviral immunity in COVID-19 is critical for developing an explanation for why some patients with no comorbidities progress to critical illness, whereas others do not, a phenomenon exacerbated by new viral variants in current epidemic waves all over the country.

Several studies have described the potential immunological characteristics associated with the disease severity at transcriptomic levels, which provide effective avenue for better patient risk stratification, and further predicting hospitalization, intensive care unit (ICU) admission, monitoring patients’ response to treatment and choosing new treatment.\(^11\)\(^-\)\(^14\)

Unfortunately, these novel biomarkers remain purely investigational concern and difficult to replicate and validate, despite their prognostic potential. Because of the continual evolution of virus and the widespread use of vaccination, identifying potential innovative biomarkers and significant genes linked to cytokine storms or poor immune function is crucial in diagnosing and predicting the severity of COVID-19, which is beneficial to early active treatment and reduce mortality.

In this study, we utilize the comprehensive bioinformatic analysis to identify candidate biological markers (hub genes) that might influence COVID-19 progression. Firstly, one microarray datasets GSE213313 from Gene Expression Omnibus database of the National Cancer Institute (NCBI) was analyzed to identify differentially expressed genes (DEGs) between COVID-19 patients and non-pneumonia controls, critical and non-critical patients, respectively. Meanwhile, the dataset was analyzed using weighted gene co-expression network analysis (WGCNA) to find the most important modules related to clinical severity as well. Secondly, using a Venn diagram to find the intersection of DEGs and WGCNA-derived hub genes and 12 overlapping genes were discovered. Subsequently, the top four upregulated genes, $MCEMP1$, $ANXA3$, $CD177$, and $SCN9A$ were then ultimately screened out through another two datasets and clinical sample validation. The primary objective of this study is to offer a comprehensive illustration of the typical characteristics of severe COVID-19, proposing potential biomarkers for the early diagnosis and prognosis of the disease.

**Materials and Methods**

**Microarray Data Information**

We screened and downloaded datasets from Gene Expression Omnibus database of the NCBI (NCBI-GEO). Patients met the following selection criteria. Firstly, patients with COVID-19 were diagnosed by real time reverse transcription polymerase chain reaction (RT-PCR) analysis of respiratory tract samples. Second, gene expression profiles of patients with COVID-19 who had severity categorization, and classification criteria were generally similar. Last, datasets from microarray data or RNA-sequencing data contained a minimum of 10 COVID-19 patients’ whole blood samples or white blood samples. According to the aforementioned standards, three gene expression profiles GSE213313, GSE172114 and GSE157103 were sorted out. The dataset of GSE213313 included 83 COVID-19 patients (50 critical and 33 non-critical patients) and 11 healthy donors. The critical and non-critical were defined as patients being placed on mechanical ventilation or not.\(^15\) Whole blood samples from 83 patients were taken from three sampling points. Samples from acute phase (T1: day 1) rather than convalescent phase (T2: day 3; T3: day 9) were selected to analyze due to disease severity and were assessed at the first timepoint of sampling. The dataset of GSE172114 included 46 critical and 23 non-critical patients. The dataset of GSE157103 included 50 critical and 50 non-critical patients. The above two datasets employed the similar diagnostic criteria and classified the patients as severe status based on the use of invasive mechanical ventilation. The GSE213313 was utilized to identify DEGs and construct WGCNA for the investigation. The GSE172114 and GSE157103 were utilized to validate hub genes.

**Identification of DEGs, Functional Enrichment and Immune Infiltration Analysis**

The matrix files of three datasets from GEO were downloaded for further analysis. The R package “limma” was accounted for identifying DEGs between COVID-19 patients and healthy donors, as well as critical and non-critical patients. DEGs were
screened by setting cut-off value to adjusted p-value <0.05 and/or p-value <0.05 (between critical and non-critical patients), and |log₂FC| ≥ 1. The gene ontology (GO) enrichment and Kyoto encyclopedia of genes and genomes (KEGG) pathway analyses were conducted using the R package “clusterprofiler”. GO terms or KEGG pathways having an adjusted P < 0.05 were deemed statistically significant. We evaluated 28 different immune cell types using the ssGSEA in order to assess the difference in immune cells between COVID-19 patients and healthy donors.

**Weighted Co-Expression Network Analysis (WGCNA)**

Using the R package “WGCNA”, we extracted 5000 genes (based on median absolute deviation) in order to build a weight co-expression network. The adjacency matrix was transformed into topological overlap matrix (TOM) when the power of β = 16 (R² = 0.85). The genes were categorized into modules with a minimum size cutoff of 30. Similar modules having a height cut-off of 0.25 were merged together. The module having the strongest connection with clinical traits was chosen for further investigation into its biological function using GO analysis.

**Hub Genes Detection and Validation**

WGCNA-derived hub genes were screened by setting the criteria that gene significance (GS) >0.5 and module membership (MM) >0.8 in the most important module. Then, using a Venn diagram, we discovered common genes by intersecting DEGs and WGCNA-derived hub genes (http://bioinformatics.psb.ugent.be/webtools/Venn/). Finally, we used GSE172114 and GSE157103 datasets to confirm the difference of hub genes between critical and non-critical COVID-19 patients.

**Clinical Samples Collection and Processing**

Peripheral blood samples were collected from COVID-19 patients hospitalized to Shandong provincial hospital from December 2022 to May 2023. Demographic data were summarized in Table 1. The patients with COVID-19 were diagnosed by RT-PCR test from throat swab samples for the detection of SARS-CoV-2. Their chest CT showed signs of pneumonia. According to the New Coronavirus Pneumonia Diagnosis and Treatment Program published by the National Health Commission of China, patients were considered to have severe pneumonia when they presented the following characteristics or criteria: dyspnea, respiratory rate ≥30/min, pulse oxygen saturation ≤93%, partial pressure of arterial oxygen (PaO₂)-to-fraction of inspired oxygen (FiO₂) ratio ≤300, and/or radiological lung infiltrates >50% within 48 h, or acute respiratory failure, septic shock. Moderate pneumonia had two or more following criteria: flu-like illness such as fever and cough, respiratory rate <30/min, pulse oxygen saturation >93%, and CT suggestive of pneumonia. Controls were asymptomatic individuals or donors who had recovered from mild infection and did not have any signs of pneumonia. White blood cell samples were isolated from peripheral blood samples and stored in liquid nitrogen until use.

**Detection of Hub Genes in Whole Blood Samples**

Total RNA was extracted from white blood cells using RNeasy Mini Kit (QIAGEN). Then, the extracted RNA was reverse transcribed into cDNA using Evo M-MLV RT Mix Kit with gDNA Clean for qPCR (Accurate Biology, Hunan, China) according to the manufacturer’s protocol. RT-PCR was conducted using the SYBR Green Premix pro Taq HS qPCR kit (Accurate Biology, Hunan, China). The experiments were performed according to the manufacturer’s instructions provided in the kit using a 10 µL reaction system. The hub gene expression was normalized to the expression of the glyceraldehyde-3-phosphate dehydrogenase gene (GAPDH). The primers used for the hub genes’ RT-PCR were listed in Supplemental Table 1. The 2-ΔΔCT comparative method was used to estimate relative fold changes in gene expression.

**Statistical Analysis**

Statistical significance was calculated by Student’s t-test, Mann–Whitney U-test, χ² test, or Fisher’s exact test, as appropriate. *P < 0.05, **P < 0.01, and ***P < 0.001 represents significant statistical differences.
Screening of Stable DEGs Linked to Progression of COVID-19

We screened 968 DEGs, including 574 up-regulated and 394 down-regulated genes in the COVID-19 patient group compared with the healthy donor group, according to a threshold of |log$_2$FC| > 1 and adjusted p-value < 0.05, and plotted into the volcano map (Figure 1A). A heatmap depicted the top 25 up-regulated and down-regulated DEGs (Figure 1B).

The interferon-stimulated genes, such as IFI27 and IFI44, were substantially elevated, but immunoglobulin-E receptor α (FcεRIα) gene usually decreased. Further, we used gene ontology (GO) analysis and Kyoto encyclopedia of genes and genomes (KEGG) pathway analysis to evaluate the functional status resulting from these DEGs. These genes mostly functioned in biological processes (BP) associated with cytoplasmic translation, response to virus, and defense response to virus. In terms of cellular components (CC), these DEGs exhibited a significant enrichment in cytosolic ribosome, ribosomal subunit, and specific granule. The molecular function (MF) terms were immune receptor activity, and MHC

| Table 1 Characteristics of All Patients Admitted to the Hospital For COVID-19 |
|-----------------------------|------------------|------------------|------------------|------------------|
|                             | Total (n = 70)   | Moderate (n = 31) | Severe (n = 39)  | P-value          |
| Gender                      |                 |                  |                  |                  |
| Male                        | 44 (62.86%)     | 16 (51.61%)      | 28 (71.79%)      | 0.0826$^a$      |
| Female                      | 26 (37.14%)     | 15 (48.39%)      | 11 (28.21%)      |                  |
| Age/years (mean (SD))       | 70.14 (14.37)   | 66.65 (15.12)    | 72.92 (13.29)    | 0.0691$^b$      |
| Comorbidity                 |                 |                  |                  |                  |
| Pneumonia disease           | 70 (100%)       | 31 (100%)        | 39 (100%)        | –                |
| Cancer                      | 14 (20.00%)     | 8 (25.81%)       | 6 (15.38%)       | 0.2789$^a$      |
| Hypertension                | 31 (44.29%)     | 10 (32.26%)      | 21 (53.85%)      | 0.0709$^a$      |
| Diabetes                    | 18 (25.71%)     | 6 (19.35%)       | 12 (30.77%)      | 0.2778$^a$      |
| Coronary heart disease      | 27 (38.57%)     | 10 (32.26%)      | 17 (43.59%)      | 0.3333$^a$      |
| Cerebrovascular disease     | 13 (18.57%)     | 8 (25.81%)       | 5 (12.82%)       | 0.2200$^a$      |
| Laboratory tests (mean (SD))|                 |                  |                  |                  |
| Total leukocytes ($\times 10^9$/L) | 7.99 (3.70) | 7.12 (3.64) | 8.67 (3.64) | 0.0820$^b$ |
| Neutrophil ($\times 10^9$/L) | 6.35 (3.52) | 5.31 (3.30) | 7.18 (3.51) | 0.0268$^a$ |
| Proportion of Neutrophil (%)| 77.15 (12.48) | 71.93 (12.86) | 81.30 (10.64) | 0.0013$^b$ |
| Lymphocytes ($\times 10^9$/L)| 1.11 (0.76)  | 1.29 (0.87)  | 0.97 (0.64)  | 0.0824$^b$ |
| Proportion of Lymphocytes (%)| 16.82 (12.68) | 22.25 (15.11) | 12.50 (8.27) | 0.0010$^a$ |
| Platelet ($\times 10^9$/L)  | 222.10 (97.69) | 232.39 (82.59) | 213.92 (108.58) | 0.4362$^a$ |
| C–reactive protein (mg/l)   | 33.09 (48.82) | 18.88 (30.75) | 45.32 (37.88) | 0.0259$^b$ |
| Procalcitonin (ng/ml)       | 0.15 (0.17) | 0.10 (0.09) | 0.18 (0.20) | 0.0428$^a$ |
| Interleukin–6 (pg/ml)       | 34.36 (92.03) | 10.34 (19.52) | 51.37 (117.13) | 0.0668$^a$ |
| $\text{PaO}_2/\text{FiO}_2$ grade (mmHg) |            |                  |                  |                  |
| >300                        | 36 (51.43%)     | 31 (100%)        | 5 (12.82%)       | 0.0000$^b$      |
| >200 to ≤300                | 18 (25.71%)     | 0 (0%)           | 18 (46.15%)      | 0.0000$^b$      |
| >100 to ≤200                | 14 (20.00%)     | 0 (0%)           | 14 (35.90%)      | 0.0000$^b$      |
| ≤100                        | 2 (2.86%)       | 0 (0%)           | 2 (5.13%)        | 0.4994$^a$      |
| Respiratory support         |                 |                  |                  |                  |
| Invasive mechanical ventilation | 1 (1.43%)    | 0 (0%)           | 1 (2.56%)        | 1.0000$^b$      |
| Noninvasive mechanical ventilatory support | 3 (4.29%) | 0 (0%)         | 3 (7.69%)        | 0.2491$^b$      |
| High–flow nasal cannula oxygen therapy | 9 (12.86%) | 0 (0%)     | 9 (23.10%)        | 0.0042$^b$      |
| Low–flow oxygen therapy     | 49 (70.00%)     | 23 (74.19%)      | 26 (66.67%)      | 0.4949$^a$      |
| Outcome                     |                 |                  |                  |                  |
| Admission to ICU            | 10 (14.29%)     | 1 (3.23%)        | 9 (23.08%)       | 0.0184$^a$      |
| Death                       | 5 (7.14%)       | 0 (0%)           | 5 (12.82%)       | 0.0616$^a$      |

Notes: All the participants resided in Shandong, China. Data are mean (SD) or n (%). $p$ values were calculated by Student’s t test, chi-squared test, or Fisher’s exact test, as appropriate. $^a$Compared by two-sided chi-squared test. $^b$Compared by two-sided Student’s t-test. $^c$Compared by two-sided Fisher’s exact test.
Figure 1 Identification of DEGs and functional enrichment analyses. (a) A volcano plot was generated to depict the differentially expressed genes (DEGs; $|\log_{2}\text{FC}| > 1$, adjusted $p$-value < 0.05) between healthy donors (n = 11) and COVID-19 patients (n = 34). Up-regulated genes are marked in light red; down-regulated genes are marked in light blue. (b) Heatmap representation of top 25 significant genes, ordered by fold change. Up-regulated genes are shown in red, and down-regulated genes are shown in blue. (c) Bar plot of GO enrichment analysis of DEGs. X-axis represents -log10 adjusted $p$-value. Y-axis represents different functional groups (also named GO terms) and gene counts. The red bars indicate biological process (BP) terms; the green bars indicate cellular component (CC) terms, and the blue bars indicate molecular function (MF) terms. (d) Bubble plot of KEGG pathway analysis of DEGs. The color and size of the dots represent the range of the adjusted $p$-value and the number of genes mapped to the indicated KEGG terms. Gene-Ratio is defined as the ratio of the DEGs number to the total gene number.
class II receptor activity (Figure 1C). Moreover, in terms of KEGG pathway, these genes were mainly enriched in the coronavirus disease-COVID-19 pathway (Figure 1D).

Meanwhile, 52 DEGs from the normalized data were screened out, with the findings displayed as heat map and volcano map (Supplementary Figure 1A and B). There were 46 up-regulated and 6 down-regulated genes in critical patients compared with non-critical patients. The results of the GO analysis revealed that the DEGs were mainly enriched in the acute inflammatory response and regulation of inflammatory in biological process (BP); specific granule and specific granule lumen in cellular component (CC); calcium-dependent protein binding and RAGE receptor binding in molecular function (MF) (Supplementary Figure 1C).

### Immune Landscape Related to Characteristics of COVID-19

Given the pivotal role of immune cells in the progression of COVID-19, we uploaded microarray data of 45 whole-blood samples from dataset GSE213313 to perform immune infiltration analysis for evaluating the immune function. Using ssGSEA and specific gene markers, the enrichment scores of 28 immune cell types between healthy donors and COVID-19 patients were calculated. A heatmap of infiltration was drawn (Figure 2A). The results indicated that a high presence of neutrophils, macrophage, and activated dendritic cell in the COVID-19 group. Moreover, the differential expression analysis was conducted between groups according to Wilcoxon test, and the differences between groups were shown by a violin chart (Figure 2B). According to p-value <0.001, 12 different immune cells were obtained, and activated CD8⁺ T cells, natural killer cells, central memory CD4⁺ T cells, memory B cells, and T follicular helper cells were significantly enriched in healthy donors, while the fraction of neutrophils, macrophages, activated dendritic cells, type 17 T helper cells central, and memory CD8⁺ T cells showed marked elevation in COVID-19 group.

### Key Module Identification Using WGCNA

WGCNA was employed to analyze the gene expression value in 45 samples to find the crucial modules linked to the severity of COVID-19. Clinical features, such as disease status and disease severity were retrieved from raw files (Supplementary Figure S2A). The soft-thresholding power, calculated using a scale-free $R^2$ ($R^2 = 0.85$) analysis, was set to 16 (Supplementary Figure S2B). Ten modules were discerned when the DissThres was initialized to 0.25 subsequent to the fusion of dynamic modules, as shown in the clustering dendrograms (Supplementary Figure S2C). We generated a dendrogram and heatmap to visually present the interrelated eigengenes. The dendrogram clearly showed a substantial association between the blue module and disease severity (Figure 3A). The correlation between the eigengene (ME) values of the modules and samples traits represents the association between the modules and clinical samples traits. The results of this correlation were presented as heatmap profiles. The results indicated that the blue module displayed the most robust link with disease severity (Pearson correlation coefficient = 0.82, $P = 2e^{-11}$), and it was strongly associated with disease status as well (Pearson correlation coefficient = 0.74, $P = 2e^{-8}$) (Figure 3B).

### Validation of Key Module and Functional Enrichment Analysis

The blue module exhibited the strongest positive correlation with severity of COVID-19 and were processed for further analysis (Figure 4A). To identify stable and robust hub genes accurately, a total of 217 critical genes were selected based on their significant correlation in both gene significance (GS) and module membership (MM). This selection was made by setting Gene-Significance >0.5, Module-Membership >0.8, as demonstrated in the scatterplots (Figure 4B). To deeper understand the potential function of those genes, GO enrichment analysis was conducted. These intersection genes were found to be significantly enriched in COVID-19 progression involved in biological processes (BP) such as cytokine mediated signaling pathway and positive regulation of cytokine production; cell components (CC) such as ficolin-1 rich granule and tertiary granule; molecular functions (MF) such as immune receptor activity and inhibitory MHC class I receptor activity (Figure 4C).

### Identification of Hub Genes

A total of 968 and 52 genes were obtained from DEGs (Healthy donors vs COVID-19; Non-critical vs Critical), and 217 genes were obtained from the key modules of WGCNA, respectively (Table 2). To identify co-expression genes between
DEGs and WGCNA genes, we utilized a Venn diagram to determine the intersection of three gene sets. Consequently, we identified 12 overlapping genes that were selected as potential hub genes, indicating their significant involvement in the formation and progression of severe COVID-19 (Figure 5B).

Figure 2 Analysis of immune infiltration associated with COVID-19. (a) Heatmap showing the enrichment score of immune cells between healthy donors and COVID-19 patients from dataset GSE213313. The specific enrichment scores of 28 immune cell types were calculated by ssGSEA based on specific gene marker. (b) Immune cells abundance in healthy donors and COVID-19 patients. The blue indicates samples of healthy donors, and the red indicates the samples of COVID-19 patients. The p-values were obtained using Wilcoxon test.
Validation of Potential Genes Linked to Disease Severity

To confirm the findings mentioned before, the expression levels of the aforementioned 12 hub genes were initially validated in two additional datasets (GSE172114 and GSE157103). The mRNA microarray dataset GSE172114 consists of 46 critical and 23 non-critical hospitalized COVID-19 patients. RNA-seq was performed on 69 whole blood RNA samples. The dataset GSE157103 consists of 50 critical and 50 non-critical COVID-19 patients. RNA-seq was conducted on 100 leukocyte RNA samples. The reconfirmed results showed that, in dataset GSE172114, MCEMP1, SCN9A, ANXA3, CD177, CACNA1E, IL-18R1, BMX, SOCS3, CDK5RAP2, ANKRD22, CD274, and FCGR1A were significantly highly expressed in critical COVID-19 group (Figure 6A and Supplementary Figure S3). However, in dataset GSE157103, the expressions of ANKRD22, CD274, and FCGR1A were not statistically significant between critical and non-critical group (Figure 6B and Supplementary Figure S3).

Candidate Genes Expression in Clinical Specimens from COVID-19 Patients

To further validate the findings of our bioinformatics research above, we conducted RT-PCR analysis on leukocyte samples acquired from 26 healthy donors, 31 patients with moderate COVID-19, and 39 patients with severe COVID-19. The average age of patients was over 70 years old. Therefore, this study is mainly aimed at elderly patients. The RT-PCR results indicated that the relative expression levels of MCEMP1, SCN9A, ANXA3, CD177, CACNA1E, IL-18R1, BMX, SOCS3, CDK5RAP2, ANKRD22, CD274, and FCGR1A were significantly highly expressed in critical COVID-19 group (Figure 6A and Supplementary Figure S3). However, in dataset GSE157103, the expressions of ANKRD22, CD274, and FCGR1A were not statistically significant between critical and non-critical group (Figure 6B and Supplementary Figure S3).
Figure 4 Validation of severity-related modules and functional enrichment analyses. (a) The blue module was the most positively associated with severity of COVID-19. (b) A scatterplot of Gene significance for COVID-19 severity vs Module Membership in blue module (Red line: Module-Membership > 0.8 and Gene-Significance > 0.5 were set to define genes that had the strongest positive correlation with disease severity). (c) Bar plot of GO enrichment analysis of genes from blue module. X-axis represents \(-\log_{10}\) adjusted p-value. Y-axis represents different GO terms and gene counts. The red bars indicate biological process (BP) terms; the green bars indicate cellular component (CC) terms, and the blue bars indicate molecular function (MF) terms.
Discussion

The pathophysiology and severity of COVID-19 exhibit variability among patients and depend partly on underlying host immunological state or risk factors. Although early vaccination and antiviral treatment have been effective in limiting the risk of developing severe or critical COVID-19, many patients still rapidly developed into respiratory failure, even acute respiratory distress syndrome (ARDS) or shock and require ICU management.\(^{18}\) Moreover, some studies indicated that patients who recovered from the severe form of the acute disease, eg, post intensive care syndrome, could also suffer long-term sequelae, which indicated that the return to their former health trajectory is slow and painful.\(^{19–21}\) Hence, identifying potential biomarkers capable of estimating severity and prognosis of COVID-19 may improve predictive accuracy and aid in clinical decision-making during the pandemic. Our research demonstrated that substantial alterations have occurred in COVID-19 patients based on blood samples, which are mainly characterized with excessive neutrophils activation and cytokine production.

Considering that each patient with COVID-19 is captured at different timepoints of infection and the discernible differences in immune response between acute and recovery phases,\(^{22}\) we selectively gathered and analyzed partial transcriptomics information from dataset GSE213313. As a result, we could capture snapshots of patients at early stages of their immune response and analyze the properties of whole-blood RNA during the acute phase. By combining microarray data obtained from healthy donors and patients with COVID-19, we identified 968 reliable DEGs. Notably, certain genes, including \(IFI27, OSA1,\) and \(ISG15,\) held more importance in terms of their impact on virus infection. Indeed, as well-known interferon-stimulated genes (ISGs), they have been reported widely in influenza-related studies.\(^ {23}\) Similarly, we identified 52 robust DEGs from critical and non-critical patients with COVID-19. These hub genes were mostly associated with acute inflammatory response. Overall, it was suggested that inflammatory response after viral infections, especially, interferon signaling response, as a remarkable indicator, was involved in COVID-19 development.

<table>
<thead>
<tr>
<th>List names</th>
<th>Number of elements</th>
<th>Unique elements</th>
</tr>
</thead>
<tbody>
<tr>
<td>COVID-19 vs Healthy</td>
<td>968</td>
<td>968</td>
</tr>
<tr>
<td>Critical vs Non-critical</td>
<td>52</td>
<td>52</td>
</tr>
<tr>
<td>WGCNA</td>
<td>217</td>
<td>217</td>
</tr>
</tbody>
</table>

Notes: Healthy donors vs. COVID-19 (Foldchange > 1 and adjusted p-value < 0.05); blue module of WGCNA (Module-Membership > 0.8 and Gene-significance > 0.5) and Non-critical vs. Critical (Foldchange > 1 and p-value < 0.05).

Figure 5 Key hub gene in severe progression of COVID-19. Venn diagram revealed 12 overlapping candidate hub genes.
Our study preliminarily revealed the alteration of immune cell infiltration subsequent to infection with SARS-COV-2. The fraction of neutrophils and macrophages was significantly higher, but the activated CD8+ T cells, central memory CD4+ T cells, and natural killer cells were all considerably diminished in COVID-19. These findings are predominantly consistent with other recent reports. An elevated absolute neutrophil count has been found to be predictive of ARDS and death in COVID-19.\textsuperscript{24,25} Severe COVID-19 patients had increased levels of neutrophil extracellular traps (NETs), which have been deemed as confirmed contributors to pathological inflammation of pneumonia.\textsuperscript{26} The absolute lymphocyte count is indicative of disease severity and mortality, since severe patients with COVID-19 have been shown to have significantly lower peripheral CD4+ and CD8+ T cells.\textsuperscript{25} The natural killer cells were also depleted and displayed an exhausted phenotype in severe patients.\textsuperscript{27} Overall, these observations indicate that a dysregulated host innate immune response associated with various immune cells could underlie the hyperinflammatory syndrome in COVID-19.\textsuperscript{28} Sepsis is an abnormal systemic inflammatory response of the host to infection.\textsuperscript{29} Some researchers proposed that severe COVID-19 could be defined as a sepsis caused by viral infection, also defined as “viral sepsis”.\textsuperscript{30,31} Therefore, novel biomarkers to diagnose “viral sepsis” might aid in the early identification of patients at risk of developing severe sepsis prior to organ damage, allow early intervention, and further reduce the risk of death. In our study, we observed that the hub genes MCEMP1, ANXA3, and CD177 were strongly up-regulated, which had been previously reported to be associated with sepsis.

Figure 6 Phase I validation: Candidate genes increased with disease severity. Validation of hub genes in the dataset GSE172114 and GSE157103. Nine candidate genes that highly expressed in critical COVID-19 patients from both validation sets were screened. (a) from GSE172114, and (b) from GSE157103.
MCEMP1, a single-pass transmembrane protein, participates in regulating mast cell differentiation activities or immune responses. Mast cells exacerbate septic disease via interfering with phagocyte activity of resident macrophages and increasing the release of inflammatory cytokines, which is consistent with the hyperinflammatory responses noted in COVID-19 patients. The expression levels of nine hub genes in COVID-19 patients were examined, including MCEMP1, SCN9A, ANXA3, CD177, CACNA1E, IL-18R1, BMX, SOCS3, and CDK5RAP2. The expression levels were compared between healthy donors, moderate patients, and severe patients with COVID-19. Statistical significance was determined by unpaired t-test. *P < 0.05, **P < 0.01, ***P < 0.001.
severe COVID-19. Some studies revealed that MCEMP1 could be served as a potential diagnosis marker for septic shock, and elevated MCEMP1 gene during the early phase of disease are predictive of severe COVID-19.\textsuperscript{34,35} As the upregulation of MCEMP1 is characteristic of myeloid-derived suppressor cells (MDSCs) which is a heterogeneous group of immature myeloid cells possessing potent immunosuppressive activities. Besides serving as potential early prognostic biomarkers for severe COVID-19 progression, several studies have also indicated the functional roles of MDSCs in severe COVID-19 pathogenesis and possibly even long COVID-19. Given the potential links between MDSCs and severe COVID-19, MDSCs are capable of suppressing various immune cell types such as B, NK, and in particular, T cells. MDSCs can exert their immunosuppressive activities through the depletion of arginine, tryptophan, and cysteine/ cysteine availability, the induction of Treg cells, RNS/ROS, the induction of T cell apoptosis via Fas/FasL interaction, the production of inhibitory cytokines, and PD-1/PD-L1 interaction. MDSCs can produce reactive nitrogen species (RNS) such as nitric oxide through inducible nitric oxide synthase (iNOS). These NO• released by MDSCs can react with superoxide (O2•−) to form peroxynitrite (ONOO•). MDSCs can also produce reactive oxygen species (ROS) such as superoxide through NADPH oxidase 2 (NOX2).\textsuperscript{36} Consequently, all of the results together suggest that MCEMP1 may be a key molecule in the regulation and maintenance of the bacterial or viral sepsis and it was worthy of additional investigation and development.

Annexin A3
Annexin A3 (ANXA3 encoding protein), also known as lipoprotein 3, is a member of the calcium-binding protein family.\textsuperscript{37} A transcriptomic study of circulating neutrophils reported that ANXA3 abundance increased significantly throughout the course of sepsis.\textsuperscript{38} Considering the fact that ANXA3 expression is almost exclusive to neutrophils among circulating leucocytes, it may have potential effects on the host by facilitating neutrophil microbicidal function and pathogen clearance during sepsis.\textsuperscript{39} ANXA3 could potentially contribute to inflammatory-induced lung damage in COVID-19 through the activation of nuclear factor-kappa B (NF-κB). This is because NF-κB plays a crucial role in inflammation and inhibiting the NF-κB pathway by upregulating IκBα has been observed with ANXA3 knockdown.\textsuperscript{40} NF-κB is also activated by oxidative stress, which is common in COVID-19.\textsuperscript{41}

CD177
CD177 was regarded as a prototypical gene for indicating neutrophil activation state because CD177 is a gene exclusive to neutrophils that encoding a membrane glycoprotein.\textsuperscript{42} Of note, it was also one of the highest DEGs in severe influenza infection.\textsuperscript{43} CD177 protein and mRNA expression increased in circulating neutrophils of patients who had experienced septic shock.\textsuperscript{44} Additionally, the combination of CD177 with genes IL1R2, OLFM4, and RETN has been identified as a possible prognostic predictor in sepsis patients, and it also has more advantages in predicting mortality compared with classical risk scales.\textsuperscript{45} The contribution of CD177 to the pathophysiology of COVID-19 may be attributed to its role as a signature for neutrophil activation. This activation is characterized by the homing of activated neutrophils toward infected lung tissue in acute lung injury.\textsuperscript{46,47} Subsequent to this homing process, aggressive responses are initiated, accompanied by the release of neutrophil extracellular traps (NETs), leading to an oxidative burst and the initiation of thrombus formation.\textsuperscript{48} Coagulopathy is associated with severity of COVID-19 which is characterized by high levels of D-dimer.\textsuperscript{48} Elevated CD177 mRNA expression has also been documented in patients with acute Kawasaki Disease (KD),\textsuperscript{49} and in individuals resistant to intravenous immunoglobulin (IV Ig) therapy.\textsuperscript{50} KD is a syndrome recognized as a potential complication of SARS-CoV-2 infection in children.\textsuperscript{51}

Transcriptomics studies have recently gained detailed gene expression profiles of COVID-19 and reported that ANXA3 and CD177 were associated with COVID-19 severity and mortality.\textsuperscript{52,53} However, these results were solely derived from a single cohort study or lacked further clinical confirmation. In the study, we found that the ANXA3 and CD177 tended to be higher in patients with COVID-19 compared to healthy donors, which indicated that ANXA3 and CD177 were two valid hallmarks of the physiopathology of COVID-19.

In addition, SCN9A gene encoding voltage-gated sodium channels are a critical component in human pain perception.\textsuperscript{54} We found that the expression level of SCN9A gene increased in patients during the acute phase of SARS-COV-2. However, no significant differences between moderate and severe patients were observed. As of yet, the mechanisms account for elevated mRNA levels of SCN9A in COVID-19 are unknown, and the phenomenon and its causes deserve additional investigation.
Some limitations and disadvantages in the study are worth considering. First, the establishment of the generalizability of these potential biomarker genes across diverse populations necessitates the implementation of validation studies. Our limited clinical sample size might increase the standard error of variables. Second, despite the fact that the vaccination rate in the population was almost 90% at the time, the specific vaccination status of the enrolled patients was unclear. This may affect the results of tests. Finally, it is yet uncertain if the modulation of these potential biomarker genes was specific for SARS-COV-2 infection. The possible regulatory mechanisms should be elucidated through further investigation, such as controlled animal trials.

**Conclusion**

In summary, our comprehensive bioinformatics analysis results showed that considerable change occurred in patients with COVID-19 and neutrophils activation was typical characteristics after SARS-COV-2 infection. Several hub genes were related to excessive neutrophil activation, and cytokine-mediated signaling pathway. Moreover, further confirmation with another two datasets and clinical samples revealed *MCEMP1, ANXA3, CD177,* and *SCN9A* were potential diagnostic markers of COVID-19. Particularly, *MCEMP1*, as clinical potential indicator, could reflect disease severity of COVID-19. Overall, understanding the biology and functional significance of the newly discovered hub genes involved in COVID-19 development provides insights into the molecular mechanisms linked to disease severity. They could also facilitate risk stratification and early identification of patients who are more susceptible to developing severe symptoms. These findings will contribute to intensifying monitoring or treatment interventions for patients with COVID-19 at high-risk.

**Data Sharing Statement**


**Ethics Approval and Consent to Participate**

The study was conducted in accordance with the Declaration of Helsinki and approved by the ethics committee of Shandong provincial hospital (approval No. 2023-805). A written informed consent to participate in this study was obtained from all participants.

**Funding**

This research was funded by the Natural Science Foundation of China (Grant NO.82200014) and the Natural Science Foundation of Shandong Province (Grant NO. ZR2022QH216).

**Supplementary Information**

Additional supplementary information: the supplemental data consist of one table and three figures.

**Disclosure**

The authors declare that no potential conflict of interest exists in the submission of this manuscript.

**References**


