European Review for Medical and Pharmacological Sciences 2021; 25: 6411-6424

Identification of hub genes and molecular subtypes in COVID-19 based on WGCNA

R.-W. HU[1,2], C. LIU[3], Y.-Y. YAN[1,2], D. LI[1,2]

[1]Department of Gastrointestinal Surgery, Henan Provincial People's Hospital, Zhengzhou, Henan, China
[2]Department of Gastrointestinal Surgery, Zhengzhou University People's Hospital, Zhengzhou, Henan, China
[3]Department of Radiology, The First Affiliated Hospital of Anhui Medical University, Hefei, Anhui, China

Abstract. – OBJECTIVE: The heterogeneity of clinical manifestations and mortality rates in Coronavirus disease 2019 (COVID-19) patients may be related to the existence of molecular subtypes in COVID-19. To improve current management, it is essential to find the hub genes and pathways associated with different COVID-19 subtypes.

MATERIALS AND METHODS: The whole-genome sequencing information (GSE156063, GSE163151) of nasopharyngeal swabs from normal subjects and COVID-19 patients were downloaded from the Gene Expression Omnibus (GEO) database. The molecular subtypes of patients with COVID-19 were classified using the "consistent clustering" method, and the specific genes associated with each subtype were found. Differentially expressed genes (DEGs) were screened between normal subjects and COVID-19 patients; the Weighted gene co-expression network analysis (WGCNA) method was used to find the key module genes of COVID-19 patients. Subtype-specific, differentially expressed and module-related genes were collected and intersected. Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis were carried out and protein-protein interaction (PPI) networks were generated. The pathways enriched in COVID-19 subtypes were analyzed by gene set variation analysis (GSVA).

RESULTS: Patients with COVID-19 were divided into three subtypes, and there was no significant difference in gender and age distribution between subtypes. 82 differential gene pathways were screened between Subtypes I and II, 131 differential gene pathways were screened between Subtypes I and III, and 107 differential gene pathways were screened between Subtypes II and III. Finally, 44 differentially expressed key genes were screened, including 11 hub genes (RSAD2, IFIT1, MX1, OAS1, OAS2, BST2, IFI27, IFI35, IFI6, IFITM3, STAT2).

CONCLUSIONS: There are significant differences in gene activation and pathway enrichment among different molecular subtypes of COVID-19, which may account for the heterogeneity in clinical presentation and the prognosis of patients.

Key Words:
COVID-19, WGCNA, GEO, Genes cluster, Molecular subtypes.

Abbreviations

COVID-19: Corona Virus Disease 2019; SARS-COV-2: Severe acute respiratory syndrome coronavirus 2; GEO: Gene Expression Omnibus; GSVA: Gene set variation analysis; GSEA: Gene set enrichment analysis; DEGs: Differentially expressed genes; WGCNA: Weighted gene co-expression network analysis; GO: Gene ontology; KEGG: Kyoto Encyclopedia of Genes and Genomes; PCA: Principal component analysis; BP: Biological process; CC: Cellular Component; MF: Molecular Function; PPI: protein-protein interaction. (The abbreviations for the Supplementary Figures are placed in the **Supplementary Materials**).

Introduction

The outbreak of Coronavirus disease 2019 (COVID-19) has become a severe threat to people's lives worldwide, and so far, the pandemic has not been effectively controlled[1,2]. The clinical manifestations and the mortality rates of these patients are significantly different. Advanced age, male gender, and comorbidities have been documented as mortality risk factors[3-7]. Nowadays, nucleic acid detection by collecting nasopharyngeal swabs is the most commonly used method for diagnosing COVID-19 patients. The Gene Expression Omnibus (GEO) database[8,9] is one of the largest genome sequencing databases at pres-

Corresponding Author: Dan Li, MD; e-mail: sry_lidan@sina.com

ent and contains detailed genome sequencing and clinical information of nasopharyngeal swabs from COVID-19 patients. Weighted gene co-expression network analysis (WGCNA) is a widely used bioinformatics tool to identify gene sets with high synergistic variation[10] and has been acknowledged by most scholars[11-13]. Unlike the traditional screening methods for Differentially expressed genes (DEGs), WGCNA focuses more on identifying genes with similar functions in the whole module rather than the differential expression of individual genes.

Gene activation is significantly different in COVID-19 patients, which may be determining factor affecting the prognosis of patients. Based on the whole genome sequencing data, COVID-19 patients were clustered by a consistent clustering method to find the differentially expressed genes among different subtypes of patients[14-16]. Differentially expressed genes of each subtype were intersected with the related genes screened by DEGs and WGCNA, and the hub genes expressed in COVID-19 patients were screened. Finally, Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses were carried out[17,18]. Constructing a protein-protein interaction (PPI) network based on differential expression genes and searching for hub genes can help us narrow the research field and conduct studies more efficiently[19,20]. Gene set variation analysis (GSVA) is a functional enrichment analysis method similar to gene set enrichment analysis (GSEA)[21]. It is widely used in clinical research and can screen the differentially expressed pathways between molecular subtypes[22,23]. GSVA analysis of COVID-19 patients found different signaling pathways between different subtypes, thus explaining the differences in prognosis of COVID-19 patients.

Materials and Methods

Data Collection

The patients' high-throughput genomic sequences, clinical information, and annotated genetic information (GSE156063, GSE163151) were downloaded from the GEO database[9], and the whole-genome sequencing information from nasopharyngeal swabs of normal subjects and COVID-19 patients were screened and standardized.

DEG Analysis

R software's "limma" package[24] was used to analyze the differential gene expression between normal subjects and COVID-19 patients. Gene expression values were log2 transformed. Genes with |log2 fold change (FC)| > 1 and $p < 0.05$ were considered as significant. A higher-ranked gene was associated with a smaller p-value. Volcano plots and heatmaps of the DEGs were generated.

Co-Expression Network Construction by WGCNA

The gene co-expression network was constructed by using the R software "WGCNA" package[10], and the optimal soft thresholding power was calculated using the "PickSoftThreshold" function. When the scale independence was set to 0.9, the best power value was selected, and gene correlation modules were constructed. The gene modules with the largest correlation coefficient and statistical significance were screened.

Consensus Clustering for COVID-19

The datasets GSE156063 and GSE163151 were log2 transformed, and the batch correction of the two datasets was carried out using the R software "sva" package[25] to eliminate the batch effect between different platforms. Principal component analysis (PCA) was used to evaluate the effect of data processing. Then, the "ConsensusClusterPlus" package[14] was used to perform consistent clustering among the 231 standardized COVID-19 patients, and the most stable K value among different molecular subtypes was selected to determine the number of classifications.

GSVA in Different Molecular Subgroups of COVID-19

In order to find specific activation pathways in each subtype, the R software packages "GSVA" and "GSEABase"[21] were used to analyze the gene sets variation among COVID-19 patients with different molecular subtypes.

Identification of the Key Genes in Functional Modules and Differentially Expressed Genes

In datasets GSE156063, GSE163151, the differentially expressed genes were screened by DEGs, the gene modules with the strongest correlation were screened by WGCNA, and the specific genes related to COVID-19 patient subtypes. These genes were intersected to screen the key genes.

Functional and Pathway Enrichment

GO, and KEGG enrichment analyses of the selected key genes were performed by R software

packages "org.Hs.eg.db" and "enrichplot", and the statistically significant gene enrichment pathways were screened[26-28]. GO enrichment analysis can be divided into three sub-ontologies: Biological process (BP), Cellular Component (CC), and Molecular Function (MF)[17,29].

Analysis of PPI of Crucial Genes

The STRING database was used to build a PPI network for the representative genes, and then to screen the hub genes with the most related nodes[30,31].

Correlation Analysis of Hub Gene Expression

To study the differential expression of hub genes between normal subjects and COVID-19 patients, and to study the ability of each gene to diagnose COVID-19 patients by ROC curve, the Pearson correlation coefficient method was used to analyze the correlation between hub gene expression and age, gender and subtypes of COVID-19 patients.

Analysis of Pan-Cancer Association of Hub Gene Expression

The whole-genome sequencing information, clinical information, immunophenotype data and tumor stem cell index data of 33 types of tumors in the The Cancer Genome Atlas (TCGA) database were downloaded from the UCSC Xena website (https://xenabrowser.net/datapages/) to study the correlation between them and hub gene expression[32]. Gene-related sensitive drugs were screened by CellMiner Drug Database[33](https://discover.nci.nih.gov/cellminer/home.do)

Results

Included Patient's Data and Analysis Flow-Chart

Patients with nasopharyngeal swab RNA sequencing data were selected, including 100 normal subjects and 93 COVID-19 patients in the GSE156063 data set, and 93 normal subjects and 138 COVID-19 patients in the GSE163151 data set. The detailed flow chart is shown in Figure 1.

Differentially Expressed Genes Screening and Co-Expression Network Construction

As shown in Figure 2, in the GSE156063, 175 upregulated genes and 75 downregulated genes were screened by the DEGs method (Figure 2A: volcano plot, Figure 2B: heatmap, Figure 2C: histogram). The "PickSoft Threshold" function in the "WGCNA" package was used to calculate the best soft thresholding power. When the Power value was set to 5, and the scale independence was set to 0.9, the mean connectivity was relatively high (Figure 2D). The 11 co-expressed gene modules were clustered (Figure 2E and Figure 2F), among them, the pink module (419 genes) was the most correlated with COVID-19 patients (R=0.68, $p < 0.05$). Meanwhile, a significant correlation was found between the pink module group and module-related genes (R=0.86, $p < 0.05$, Figure 2G and Figure 2H). In the GSE163151 dataset, the same method was used for DEGs and WGCNA analysis. 675 upregulated genes and 137 downregulated genes were screened using the DEGs method (Figure 3A: volcano plot, Figure 3B: heatmap, Figure 3C: histogram). When the Power value was set to 6, and the scale independence was set to 0.9, the mean connectivity was relatively high (Figure 3D). The 14 co-expressed gene modules were clustered (Figure 3E and Figure 3F), among which the red module (422 genes) was the most correlated with COVID-19 patients (R=0.71, $p < 0.05$). There was also a significant correlation between the red module group and module-related genes (R=0.89, $p < 0.05$, Figure 3G and Figure 3H). Therefore, the related genes from the pink module in GSE156063 and the red module from GSE163151 were selected for follow-up analysis.

Consensus Clustering of COVID-19 Cases

GSE156063 and GSE163151 datasets were merged and included a total of 231 COVID-19 patients. Figure 4A shows the clustering of these two datasets before batch correction, while Figure 4B shows the clustering of two datasets after batch correction. After batch correction, the results showed that the batch effect between datasets of different platforms was successfully eliminated, and the subsequent analysis results were reliable. After batch correction, 231 COVID-19 patients were clustered into different molecular subtypes. Clustering results were most stable when the number was set to three (k=3). The patients were divided into three molecular subtypes (Figure 4C). There were 87, 108 and 36 COVID-19 patients in subtypes I, II and III, respectively. Differences in gene expression patterns were significant among the three groups, which exhibited highly similar gene expression patterns (Figure 4C). When K

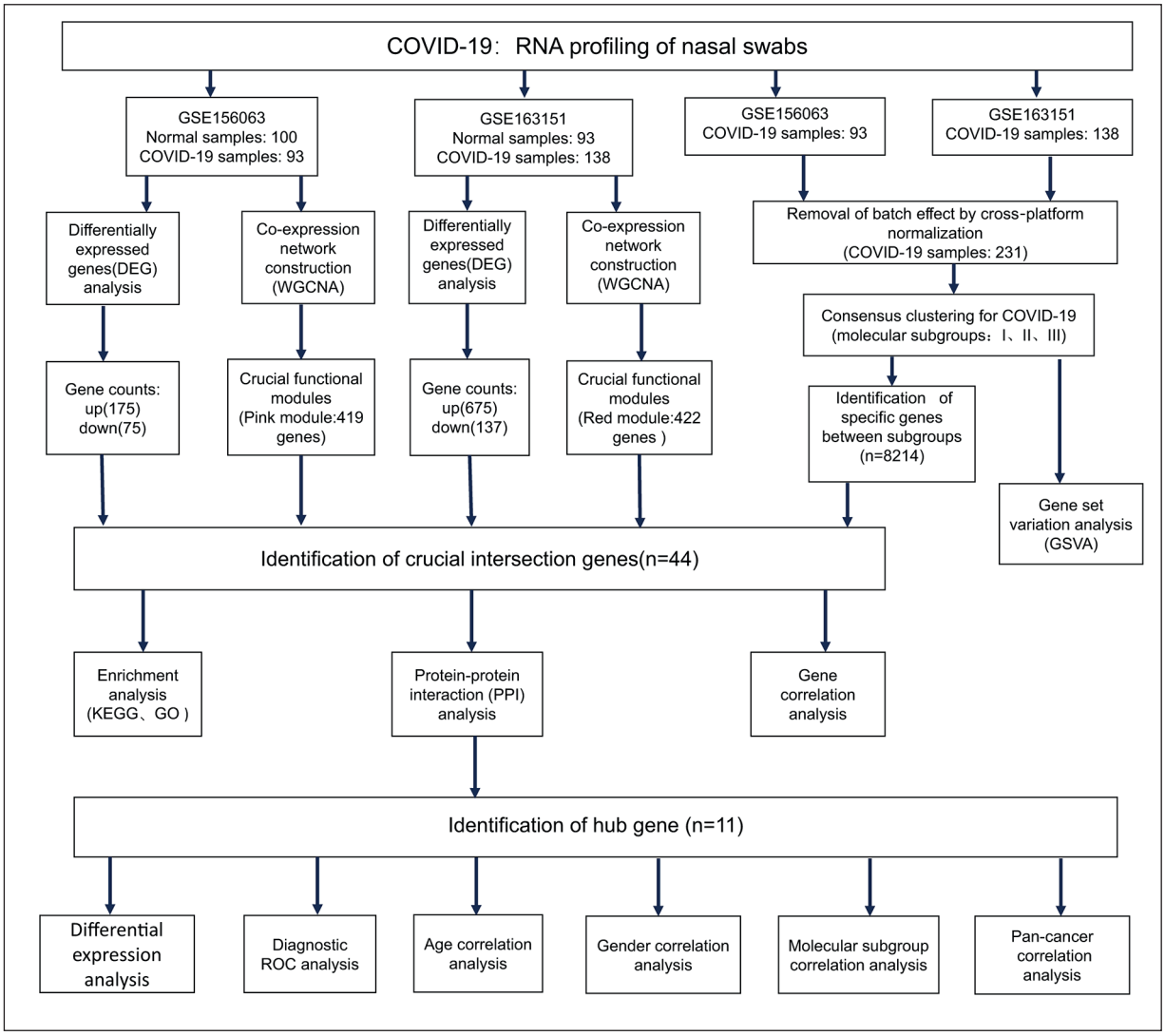

Figure 1. Research flow chart. The RNA sequencing data of nasopharyngeal swabs from normal subjects and COVID-19 patients with GSE156063 and GSE163151 were analyzed by DEGs and WGCNA methods, respectively, and differentially expressed genes and module related genes were screened. Consistent cluster analysis was used to screen subgroup specific genes in 231 patients with COVID-19. The differentially expressed genes, module-related genes and molecular subtype-related genes were selected and intersected for KEGG, GO enrichment analysis and PPI network construction. Hub gene was screened from PPI network and analyzed by differential expression analysis, gene correlation analysis, ROC analysis and pan-cancer correlation analysis. The gene pathways activated by different molecular subsets in patients with COVID-19 were analyzed by GSVA.

was set to 3, the consistency score of each subtype was close to or greater than 0.8 (Figure 4D), and the relative change of the area under the CDF curve was smaller (Figure 4E), so this classification is more robust. The age and gender ratio of patients in three subtypes were statistically analyzed, and it was found that there was no significant difference in age composition and gender ratio among COVID-19 patients in different subtypes (Figure 4F and Figure 4G).

GSVA in Different Molecular Subtypes of COVID-19 Based on KEGG (Kyoto Encyclopedia of Genes and Genomes) Pathway Enrichment Analysis

GSVA analysis was performed among molecular subtypes of COVID-19 patients, and the first 20 differentially activated KEGG pathways were drawn. 82 differential gene pathways **(Supplementary Table I)** were screened between subtypes I and II, 131 differential gene path-

ways (Supplementary Table II) were screened between subtypes I and III, and 107 differential gene pathways (Supplementary Table III) were screened between subtypes II and III. A heatmap (Figure 5A) showed the differential activation pathways between subtypes I and II, which showed significant differences in activation of "STEROID BIOSYNTHESIS", "DRUG METABOLISM CYTOCHROME P450" and "PROPANOATE METABOLISM" pathways between

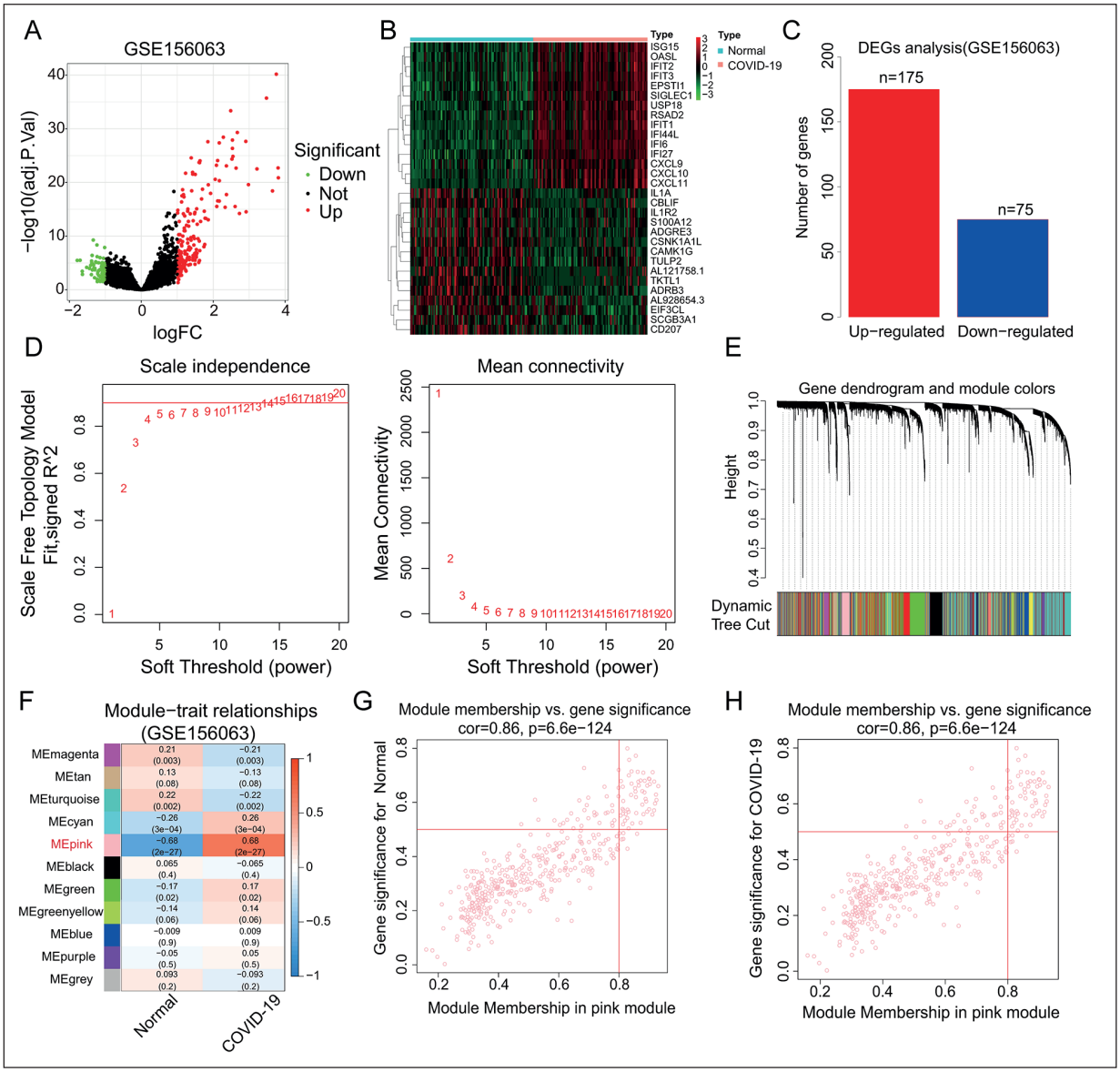

Figure 2. Differential gene screening and WGCNA analysis of the GSE156063 dataset. (**A**) Volcano plot, red dots represent up-regulated differentially expressed genes, while green dots represent down-regulated differentially expressed genes. (**B**) Heatmap, differentially expressed genes between normal subjects and COVID-19 patients. Different colors represent the relative expression of genes. Red represents high expression, and green represents low expression. The first 15 up-regulated and down-regulated genes are shown, respectively. (**C**) Differential gene expression histogram. 175 up-regulated genes and 75 down-regulated genes were screened by the DEGs method. (**D**) Analysis of scale-free index and average connectivity for various soft-threshold powers. (**E**) In the GSE156063 dataset, gene clustering dendrograms are based on different topological overlaps and module colors. (**F**) 11 gene modules were identified by consistent clustering. The correlation coefficients between the 11 gene modules and normal persons and COVID-19 patients were shown by heatmap. The pink gene modules with the strongest correlation with normal persons and COVID-19 patients were selected for follow-up analysis. (**G-H**) In normal subjects and COVID-19 patients, the correlation analysis between pink module group and module-related genes, $p < 0.05$. It shows that the genes in this module are of great significance for the study of COVID-19 diseases.

subtypes I and II. The heatmap in Figure 5B showed differential activation pathways between subtypes I and III, which showed significant differences in activation of "GLYCOSAMINOGLYCAN BIOSYNTHESIS CHONDROITIN SULFATE", "PRIMARY IMMUNODEFICIENCY" and "ADIPOCYTOKINE SIGNALING PATHWAY" pathways between subtypes I and III. The heatmap in Figure 5C showed differential activation pathways between subtypes II and III, indicating that subtypes II and III were significantly different in activation of "MAPK SIGNALING PATHWAY", "CHEMOKINE SIGNALING PATHWAY" and "NATURAL KILLER CELL

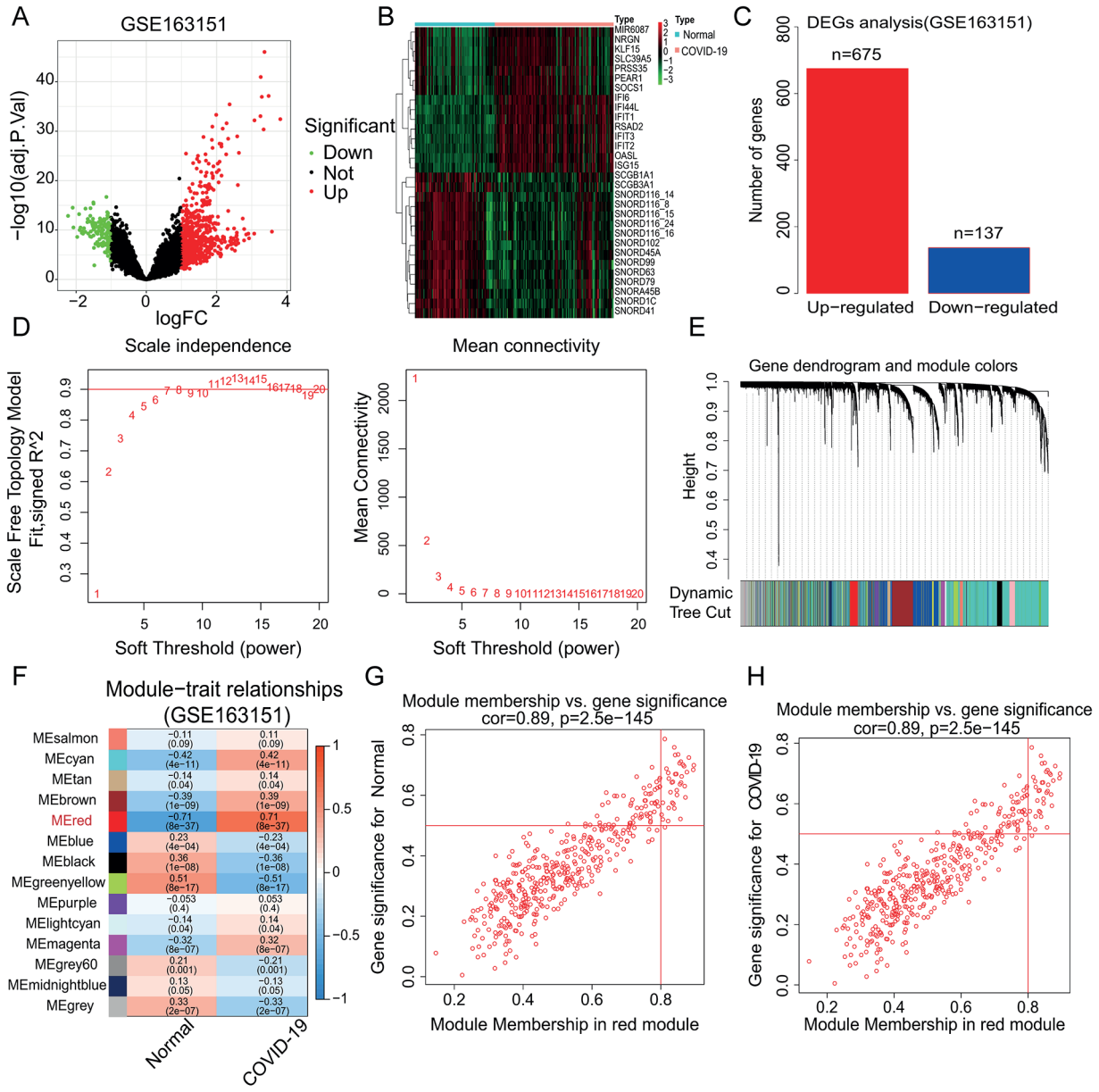

Figure 3. Differential gene screening and WGCNA analysis of the GSE163151 dataset. (**A**) Volcano plot. (**B**) Differential gene expression heatmap. (**C**) Differential gene expression histogram. 675 up-regulated genes and 137 downregulated genes. (**D**) Analysis of scale-free index and average connectivity for various soft-threshold powers. (**E**) In the GSE163151 dataset, gene clustering dendrograms are based on different topological overlaps and module colors. (**F**) 14 gene modules were identified by consistent clustering. The red gene modules with the strongest correlation with normal persons and COVID-19 patients were selected for follow-up analysis. (**G-H**) In normal subjects and COVID-19 patients, the correlation analysis between red module group and module-related genes, $p < 0.05$.

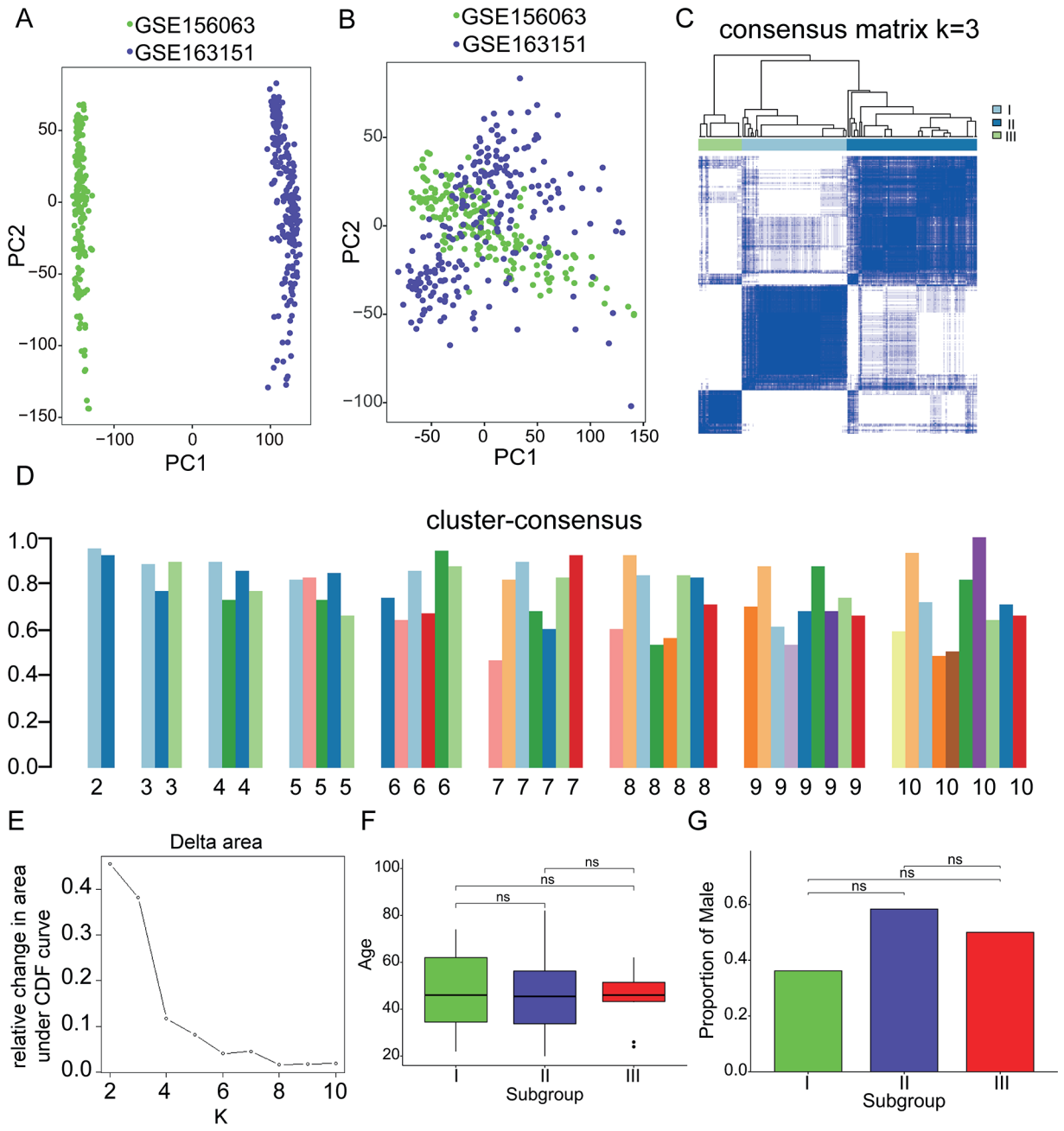

Figure 4. The consistency clustering for COVID-19 patients: (**A**) PCA clustering diagram, the batch clustering of the two datasets before batch correction. (**B**) Batch clustering of the two datasets after batch correction. (**C**) Consensus clustering plot, 231 COVID-19 patients were divided into three molecular subtypes. (**D**) The consistency clustering score of each subtype in different clustering methods. When K=3, the consistency score of each subtype cluster is close to or greater than 0.8. (**E**) CDF curve, the smaller the relative change of the area under the CDF curve, the more stable the classification. (**F**) Analysis of age composition differences among three molecular subtypes. (**G**) Analysis of gender composition ratio differences among 3 molecular subtypes.

MEDIATED CYTOTOXICITY" pathways. The results show that the activation pathways are significantly different among molecular subtypes of COVID-19 patients, and each subtype should be treated with a different approach.

Identification of Key Genes and Pathway Enrichment Analysis

Intersection of the differentially expressed genes screened by DEGs, the most relevant gene modules screened by WGCNA (GSE156063: pink

module, GSE163151: red module) and the specific genes related to subtypes of COVID-19 patients, and finally, 44 key genes were screened out (Figure 6A). All the 44 genes had a significant positive correlation (Figure 6B), and genes with a correlation coefficient greater than 0.85 (Figure 6C) were as follows: HERC6, OAS2, CMPK2, IFI6, MX1, TRIM22, IFIT1, DDX58, RSAD2, IFIH1, HERC5, CXCL10, CXCL11. Furthermore, the 44 key genes were analyzed by gene enrichment analysis. GO enrichment analysis showed that the genes were enriched in the functions of "response to virus," "defense response to virus," and "negative regulation of viral life cycle" (BP: 235 enrichment pathways, CC: 10 enrichment pathways, MF: 36 enrichment pathways). KEGG enrichment

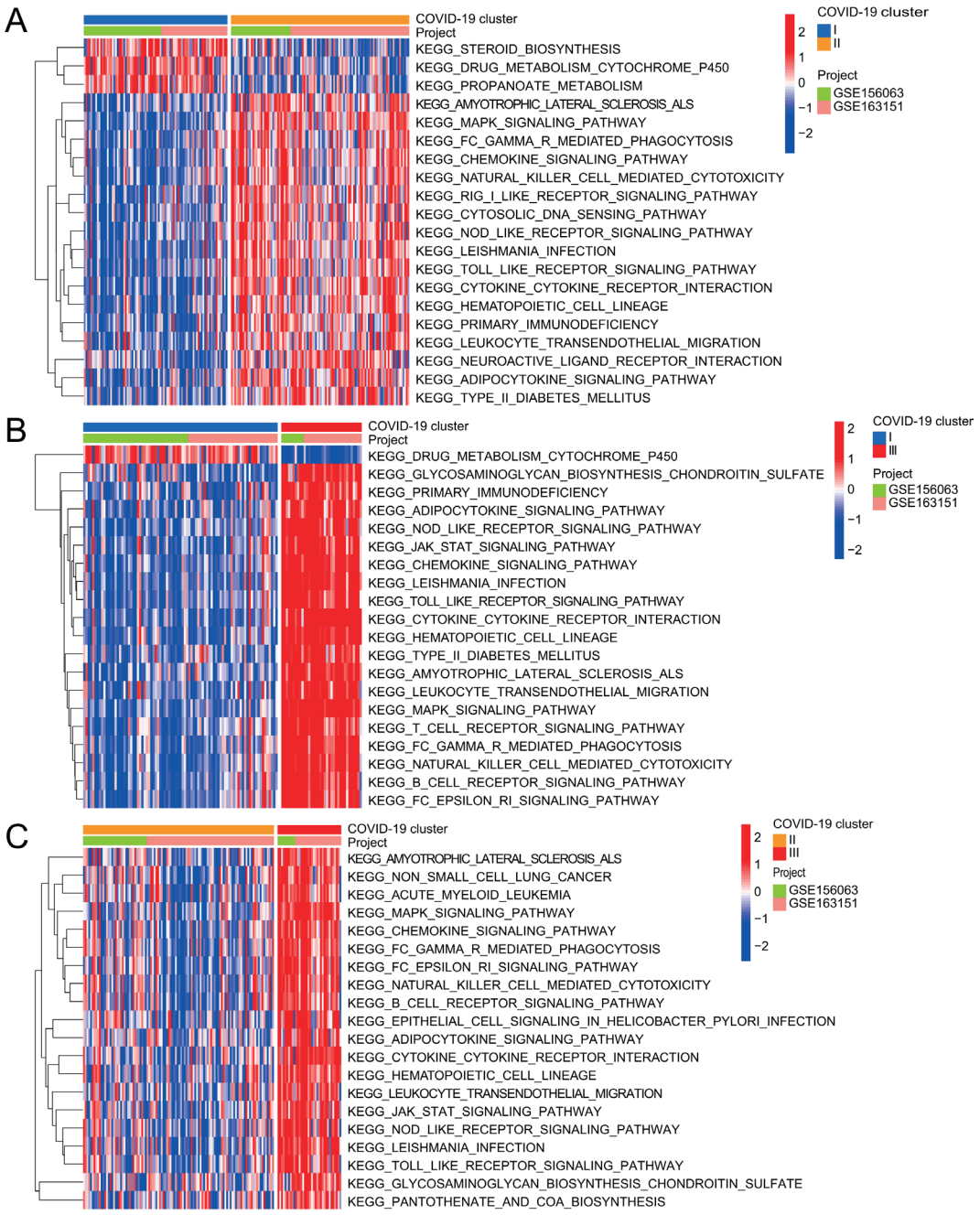

Figure 5. GSVA in different molecular subtypes of COVID-19. **(A)** The first 20 differentially activated gene pathways between molecular subtypes I and II. (**B**) The first 20 differentially activated gene pathways between molecular subtypes I and III. (**C**) The first 20 differentially activated gene pathways between molecular subtypes II and III.

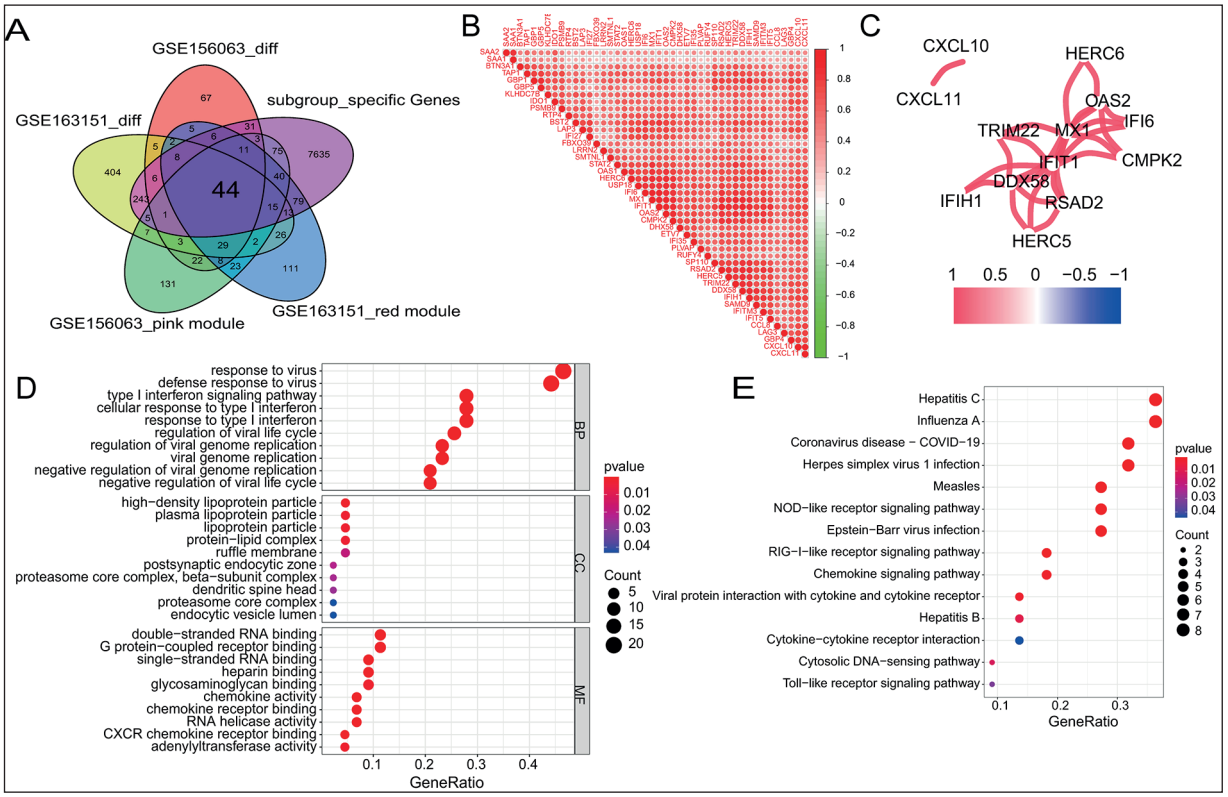

Figure 6. Key genes screening and pathways enrichment analysis. (**A**) Venn diagram, screening intersection genes. (**B**) The correlation coefficient heatmap of these 44 genes. (**C**) The gene relationship network diagram with a correlation coefficient greater than 0.85. (**D**) GO enrichment analysis; BP, CC, MF methods show the first ten enrichment pathways, respectively. (**E**) KEGG enrichment analysis; There are 14 activation gene pathways.

analysis showed that the genes were significantly enriched in "Coronavirus disease-COVID-19", "Chemokine signaling pathway" and "Cytokine-cytokine receptor interaction" pathways (14 enrichment pathways).

PPI Network Construction and Identification of Hub Genes

The PPI network of 44 key genes was constructed by the STRING database (Version:11.0), and the minimum required interaction score was set to 0.9. Finally, 28 protein-coding genes were included to construct the PPI network (Figure 7A). The hub gene (the number of gene-related nodes greater than or equal to 10) is as follows (Figure 7B): RSAD2, IFIT1, MX1, OAS1, OAS2, BST2, IFI27, IFI35, IFI6, IFITM3, STAT2.

Correlation Analysis of Hub Gene Expression with COVID-19 Patients and ROC Curve for Diagnosis COVID-19 Patients

These 11 genes were all highly expressed in COVID-19 patients (Figure 8A), and their gene expression levels were significantly higher in young patients (Figure 8C), but not related to gender (Figure 8D). In patients with different COVID-19 molecular subtypes, the gene expression levels were significantly different; the difference was statistically significant (Figure 8E). ROC curve analysis showed that all genes had high diagnostic ability in patients with COVID-19, and the area under the curve was greater than 0.8. therefore, the diagnostic ability was more significant (Figure 8B).

Pan-Cancer Correlation Analysis of Hub Gene Expression

Malignant tumor is a high-risk factor of COVID-19 associated mortality. Accordingly, it is highly significant to analyze expression in patients with malignant tumors. **Supplementary Figure 1A** shows the overall expression level of genes in all malignant tumors. **Supplementary Figure 1B** is a heatmap of gene expression differences among different cancerous tumors. **Supplementary Figure 1C** shows the correlation coefficient among 11 genes. **Supplementary Figure 2**

and Supplementary Figure 3 showed the differential expression of 11 genes in different types of cancers and their corresponding normal tissues. It can be seen that many genes are highly expressed in cancer tissues. **Supplementary Figure 4** shows the correlation heatmap between gene expression, immune cell score and tumor stem cell index in pan-cancer, which showed a significant positive correlation between gene expression and immune score in different tumor types[34]. **Supplementary Table IV** shows the results of drug sensitivity analysis, which showed a significant correlation between gene activation and a variety of drugs, which can help screen drugs for treatment in patients with malignant tumors

Discussion

COVID-19 is a highly infectious disease, which has spread to many countries and regions worldwide, turning into a humanitarian disaster with devastating social and economic repercussions[35,36]. With the progress made in COVID-related research and the continuous improvement of treatment methods, the cure rate of patients continues to improve, along with a decrease in the case fatality rate. Reducing the mortality rate is the ultimate goal of COVID-19 treatments. Although many vaccines have been developed, the constant mutation of the virus has attenuated the efficiency and usefulness of these vaccines. The prognosis of COVID-19 patients is highly heterogeneous. An increasing body of evidence suggests that advanced age and comorbidities are high-risk factors for mortality in patients infected with the SARS-COV-2 virus[37,38]. Since its emergence in December 2019, the SARS-COV-2 virus gene has mutated and increased its infectivity and viral load, and the virus-negative time is longer. Accordingly, it is essential to study the key genes and pathways of the SARS-COV-2 virus in the human body. We studied the differential gene expression of nasopharyngeal swab high-throughput sequencing microarray (GSE156063 and GSE163151) in 231 COVID-19 patients and screened the specific gene modules related to COVID-19 using the WGCNA method. The "ConsensusClusterPlus" package was used to cluster the patients for subtyping of COVID-19 patients. Differentially expressed genes, specific module genes, and subtype-related genes were selected and intersected to screen for the most critical genes. Finally, 44 key genes closely related to the pathogenesis of SARS-COV-2 virus were obtained. Gene pathway enrichment analysis of the selected key genes was performed. GO enrichment analysis showed that the genes were obviously enriched in the functions of "response to virus", "defense response to virus" and "negative regulation of viral life cycle". KEGG enrichment analysis showed that the genes were significantly enriched in the pathways "Coronavirus disease-COVID-19", "Chemo-

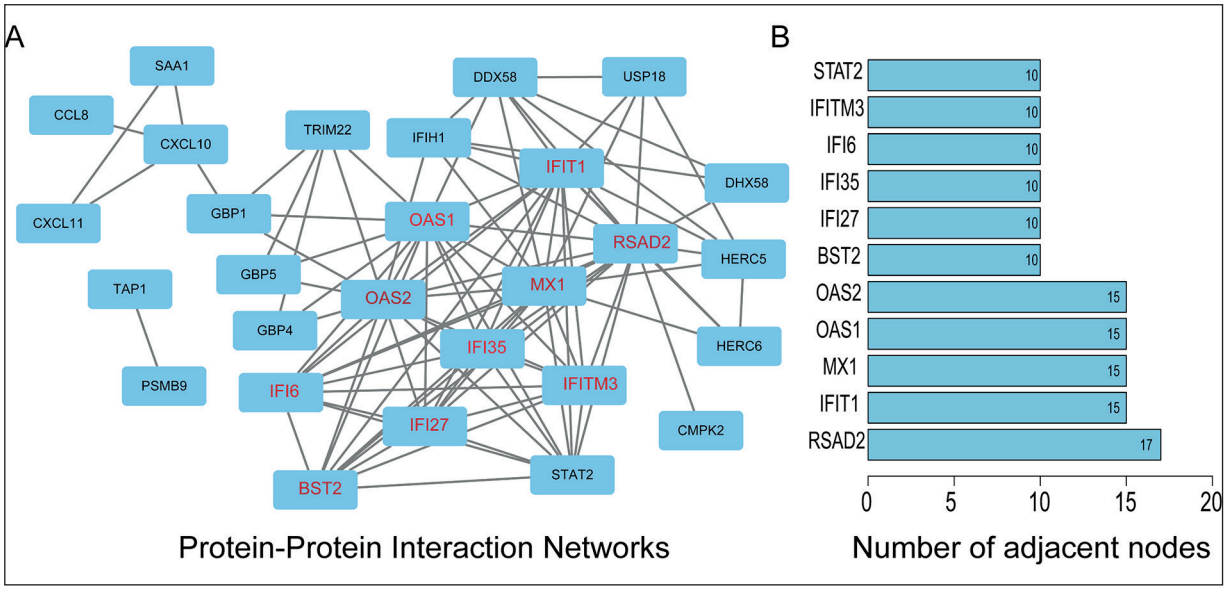

Figure 7. PPI network. (**A**) 28 crucial protein-coding genes in PPI network. (**B**) Genes with more than or equal to 10 gene-related nodes in the PPI network.

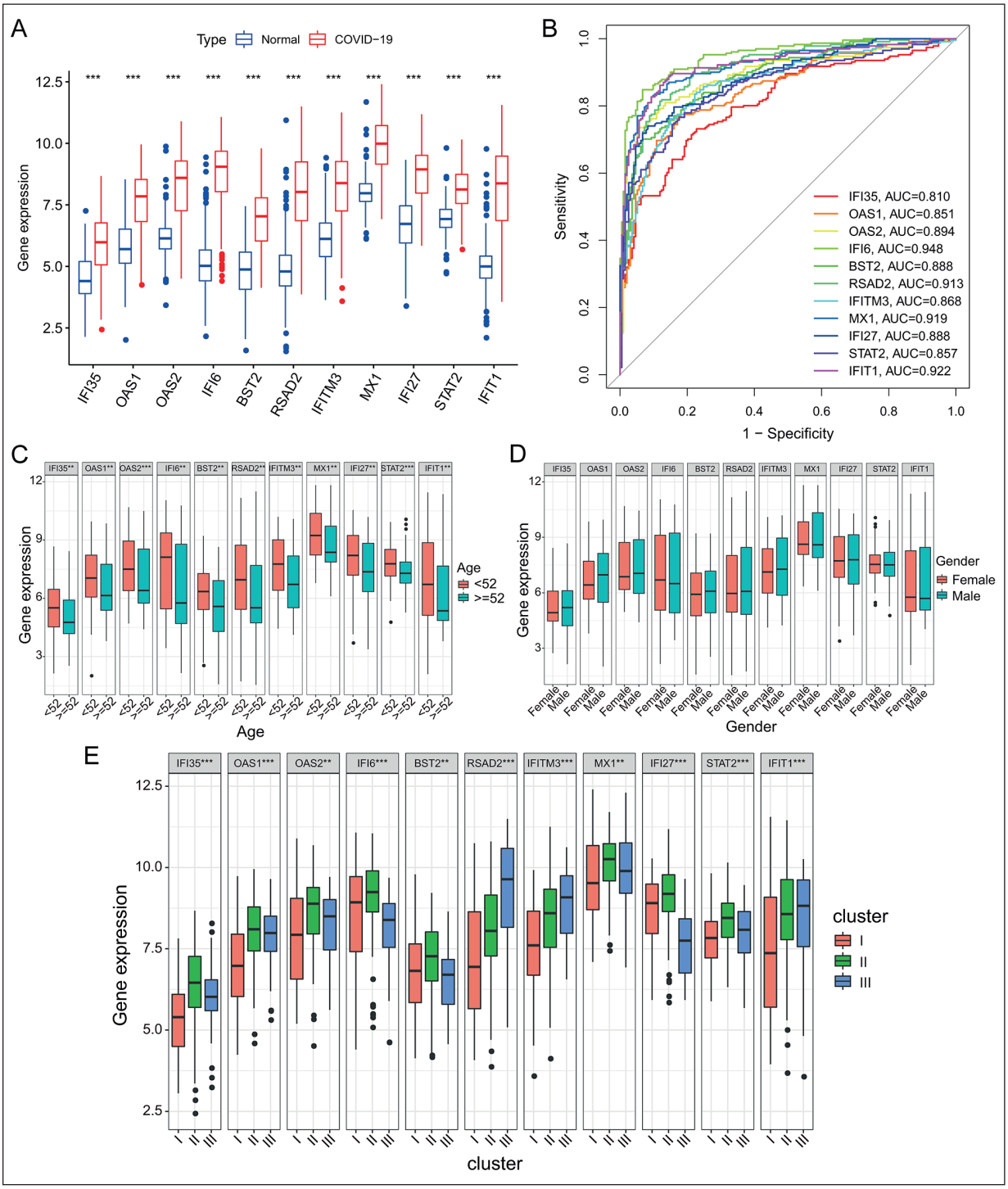

Figure 8. Correlation analysis of hub Genes and Diagnostic ROC curves. (**A**) Differential expression histogram of 11 hub genes in normal subjects and COVID-19 patients. (**B**) ROC curves of 11 hub genes for diagnosing COVID-19 patients. (**C**) Differential expression histogram of 11 hub genes in different age groups. (**D**) Differential expression histogram of 11 hub genes in different genders. (**E**) Differential expression histogram of 11 hub genes in different molecular subtypes.

kine signaling pathway" and "Cytokine-cytokine receptor interaction". Through the enrichment analysis of key genes, the scope of our follow-up research has been significantly reduced. This can help us develop more targeted drugs for the treatment of COVID-19.

Different COVID-19 patients have different prognoses; this may be explained by the different pathways activated by the SARS-COV-2 virus. Using the GSVA method[21] to find the specific activation pathway between different molecular subtyping of COVID-19 patients can explain the differences in prognosis of different patients. 82 differentially activated gene pathways between subtypes I and II were screened, 131 differentially activated gene pathways between subtypes I and III were screened, and 107 differential gene pathways between subtypes II and III were screened. There were significant differences in activation of "STEROID BIOSYNTHESIS", "DRUG METABOLISM CYTOCHROME P450" and "PROPANOATE METABOLISM" pathways between subtypes I and II. Furthermore, significant differences in activation of "GLYCOSAMINOGLYCAN BIOSYNTHESIS CHONDROITIN SULFATE", "PRIMARY IMMUNODEFICIENCY" and "ADIPOCYTOKINE SIGNALING PATHWAY" pathways were observed between subtypes I and III. Furthermore, significant differences were observed in activation of "MAPK SIGNALING PATHWAY", "CHEMOKINE SIGNALING PATHWAY" and "NATURAL KILLER CELL MEDIATED CYTOTOXICITY" pathways between subtypes II and III. The analysis of GSVA can help us choose appropriate treatment approaches and develop different targeted drugs for patients with different molecular subtypes of COVID-19.

In order to further narrow the scope of research, a PPI network was constructed using the STRING database, and the hub genes (RSAD2, IFIT1, MX1, OAS1, OAS2, BST2, IFI27, IFI35, IFI6, IFITM3, STAT2) with ten or more gene-related nodes were screened. These genes were highly expressed in COVID-19 patients, and they all have a high diagnostic yield. By using the Pearson correlation coefficient method, it was found that the expression of these genes was not related to gender but significantly correlated with age and subtypes of COVID-19 patients. These genes were highly expressed in young patients (age < 52 years old). Thus, more emphasis should be laid on specific molecular subtypes of COVID-19 patients, and different treatment approaches should be adopted accordingly.

During the epidemic of COVID-19, patients with malignant tumors are a population that requires special attention[39,40]. The presence of comorbidities often worsens the condition of these patients; patients often have relatively lower immunity which is further undermined during the treatment for tumors (such as radiotherapy and chemotherapy). These patients are thus more susceptible to be infected with the SARS-COV-2 virus and have higher mortality rates. Through the analysis of these 11 hub genes in 33 types of malignant tumors, we can discover the high-risk tumor types of SARS-COV-2 virus infection and key prognostic genes. Gene-related sensitive drugs were screened using the CellMiner drug database[41]. This information will help in providing better treatment and improving survival in patients with cancer.

Here, we improved our understanding of COVID-19, and identified specific gene pathways and key pathogenic genes involved in the SARS-COV-2 virus infection. Different approaches should be used to treat COVID-19 patients with different molecular subtypes, and high-risk populations should be identified earlier. For high-risk patients, multidisciplinary consultation should be carried out as soon as possible, and tailored treatment plans should be made accordingly to improve the cure rate and reduce mortality.

Conclusions

There are significant differences in gene activation and pathway enrichment among different molecular subtypes of COVID-19, which may account for the heterogeneity in clinical presentation and the prognosis of patients.

Author Contribution

Renwang Hu drafted the manuscript. Dan Li conceived the idea and recommended this journal. All authors read and approved the final manuscript.

Funding

Not applicable.

Conflict of Interest

The Authors declare that they have no conflict of interests.

Reference

1) Basnarkov L. SEAIR Epidemic spreading model of COVID-19. Chaos Solitons Fractals 2021; 142: 110394.

2) Niu B, Liang R, Zhang S, Zhang H, Qu X, Su Q, Zheng L, Chen Q. Epidemic analysis of COVID-19 in Italy based on spatiotemporal geographic information and Google Trends. Transbound Emerg Dis 2021; 68: 2384-2400.
3) Palaiodimos L, Kokkinidis DG, Li W, Karamanis D, Ognibene J, Arora S, Southern WN, Mantzoros CS. Severe obesity, increasing age and male sex are independently associated with worse in-hospital outcomes, and higher in-hospital mortality, in a cohort of patients with COVID-19 in the Bronx, New York. Metabolism 2020; 108: 154262.
4) Peckham H, de Gruijter NM, Raine C, Radziszewska A, Ciurtin C, Wedderburn LR, Rosser EC, Webb K, Deakin CT. Male sex identified by global COVID-19 meta-analysis as a risk factor for death and ITU admission. Nat Commun 2020; 11: 6317.
5) Sansone A, Mollaioli D, Ciocca G, Limoncin E, Colonnello E, Vena W, Jannini EA. Addressing male sexual and reproductive health in the wake of COVID-19 outbreak. J Endocrinol Invest 2021; 44: 223-231.
6) Davies NG, Klepac P, Liu Y, Prem K, Jit M, CMMID COVID-19 working group, Eggo RM. Age-dependent effects in the transmission and control of COVID-19 epidemics. Nat Med 2020; 26: 1205-1211.
7) Liu Y, Mao B, Liang S, Yang J, Lu H, Chai Y, Wang L, Zhang L, Li Q, Zhao L, He Y, Gu X, Ji X, Li L, Jie Z, Li Q, Li X, Lu H, Zhang W, Song Y, Qu J, Xu J, Shanghai Clinical Treatment Experts Group for COVID-19. Association between age and clinical characteristics and outcomes of COVID-19. Eur Respir J 2020; 55: 2001112.
8) Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Holko M, Ayanbule O, Yefanov A, Soboleva A. NCBI GEO: archive for functional genomics data sets--10 years on. Nucleic Acids Res 2011; 39: D1005-D10010.
9) Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res 2013; 41: D991-D995.
10) Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008; 9: 559.
11) Niemira M, Collin F, Szalkowska A, Bielska A, Chwialkowska K, Reszec J, Niklinski J, Kwasniewski M, Kretowski A. Molecular Signature of Subtypes of Non-Small-Cell Lung Cancer by Large-Scale Transcriptional Profiling: Identification of Key Modules and Genes by Weighted Gene Co-Expression Network Analysis (WGCNA). Cancers (Basel) 2019; 12: 37.
12) Pu L, Wang M, Li K, Feng T, Zheng P, Li S, Yao Y, Jin L. Identification micro-RNAs functional modules and genes of ischemic stroke based on weighted gene co-expression network analysis (WGCNA). Genomics 2020; 112: 2748-2754.
13) Qian J, Yang J, Liu X, Chen Z, Yan X, Gu H, Xue Q, Zhou X, Gai L, Lu P, Shi Y, Yao N. Analysis of lncRNA-mRNA networks after MEK1/2 inhibition based on WGCNA in pancreatic ductal adenocarcinoma. J Cell Physiol 2020; 235: 3657-3668.
14) Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics 2010; 26: 1572-1573.
15) Li L, Cai S, Liu S, Feng H, Zhang J. Bioinformatics analysis to screen the key prognostic genes in ovarian cancer. J Ovarian Res 2017; 10: 27.
16) Tao C, Huang K, Shi J, Hu Q, Li K, Zhu X. Genomics and Prognosis Analysis of Epithelial-Mesenchymal Transition in Glioma. Front Oncol 2020; 10: 183.
17) Gene Ontology Consortium. Gene Ontology Consortium: going forward. Nucleic Acids Res 2015; 43: D1049-D1056.
18) Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000; 28: 27-30.
19) Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ, von Mering C. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res 2011; 39: D561-D568.
20) Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, Doncheva NT, Legeay M, Fang T, Bork P, Jensen LJ, von Mering C. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res 2021; 49: D605-D612.
21) Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics 2013; 14: 7.
22) Ferreira MR, Santos GA, Biagi CA, Silva Junior WA, Zambuzzi WF. GSVA score reveals molecular signatures from transcriptomes for biomaterials comparison. J Biomed Mater Res A 2021; 109: 1004-1014.
23) Zhang J, Gu J, Guo S, Huang W, Zheng Y, Wang X, Zhang T, Zhao W, Ni B, Fan Y, Wang H. Establishing and validating a pathway prognostic signature in pancreatic cancer based on miRNA and mRNA sets using GSVA. Aging (Albany NY) 2020; 12: 22840-22858.
24) Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015; 43: e47.
25) Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics 2012; 28: 882-883.
26) Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 2016; 44: D457-D462.
27) Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes,

pathways, diseases and drugs. Nucleic Acids Res 2017; 45: D353-D361.

28) Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M. New approach for understanding genome variations in KEGG. Nucleic Acids Res 2019; 47: D590-D595.

29) Lu Y, Rosenfeld R, Simon I, Nau GJ, Bar-Joseph Z. A probabilistic generative model for GO enrichment analysis. Nucleic Acids Res 2008; 36: e109.

30) Taboada B, Verde C, Merino E. High accuracy operon prediction method based on STRING database scores. Nucleic Acids Res 2010; 38: e130.

31) Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P, Jensen LJ, von Mering C. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res 2017; 45: D362-D368.

32) Wang S, Xiong Y, Zhao L, Gu K, Li Y, Zhao F, Li J, Wang M, Wang H, Tao Z, Wu T, Zheng Y, Li X, Liu XS. UCSCXenaShiny: An R/CRAN Package for Interactive Analysis of UCSC Xena Data. Bioinformatics 2021; 29: btab561.

33) Reinhold WC, Sunshine M, Liu H, Varma S, Kohn KW, Morris J, Doroshow J, Pommier Y. CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. Cancer Res 2012; 72: 3499-3511.

34) Hu RW, Liu C, Gong JP, Cao ZX. Differential expression and immune correlation analysis of COVID-19 receptor ACE2 and TMPRSS2 genes in all normal and tumor tissues. Eur Rev Med Pharmacol Sci 2021; 25: 1724-1731.

35) Pullano G, Di Domenico L, Sabbatini CE, Valdano E, Turbelin C, Debin M, Guerrisi C, Kengne-Kuetche C, Souty C, Hanslik T, Blanchon T, Boëlle PY, Figoni J, Vaux S, Campèse C, Bernard-Stoecklin S, Colizza V. Underdetection of cases of COVID-19 in France threatens epidemic control. Nature 2021; 590: 134-139.

36) Shin HY. A multi-stage SEIR(D) model of the COVID-19 epidemic in Korea. Ann Med 2021; 53: 1159-1169.

37) Khalili MA, Leisegang K, Majzoub A, Finelli R, Panner Selvam MK, Henkel R, Mojgan M, Agarwal A. Male Fertility and the COVID-19 Pandemic: Systematic Review of the Literature. World J Mens Health 2020; 38: 506-520.

38) Lee J, Yousaf A, Fang W, Kolodney MS. Male balding is a major risk factor for severe COVID-19. J Am Acad Dermatol 2020; 83: e353-e354.

39) Cai M, Wang G, Wu Y, Wang Z, Wang G, Tao K. Study of the gastrointestinal tumor progression during the COVID-19 epidemic in Wuhan. Br J Surg 2020; 107: e502-e503.

40) Gatson NTN, Barnholtz-Sloan J, Drappatz J, Henriksson R, Hottinger AF, Hinoul P, Kruchko C, Puduvalli VK, Tran DD, Wong ET, Glas M. Tumor Treating Fields for Glioblastoma Therapy During the COVID-19 Pandemic. Front Oncol 2021; 11: 679702.

41) Liu H, D'Andrade P, Fulmer-Smentek S, Lorenzi P, Kohn KW, Weinstein JN, Pommier Y, Reinhold WC. mRNA and microRNA expression profiles of the NCI-60 integrated with drug activities. Mol Cancer Ther 2010; 9: 1080-1091.