Metatranscriptomic Characterization of COVID-19 Identified A Host Transcriptional Classifier Associated With Immune Signaling

Haocheng Zhang[1,*], Jing-Wen Ai[1,*], Wenjiao Yang[2,*], Xian Zhou[1,*], Fusheng He[2], Shumei Xie[2], Weiqi Zeng[2,3], Yang Li[1], Yiqi Yu[1], Xuejing Gou[2], Yongjun Li[2], Xiaorui Wang[2], Hang Su[2], Teng Xu[2,3,#], Wenhong Zhang[1,#]

[1]Department of Infection Diseases, Huashan Hospital affiliated to Fudan University, Shanghai 200040, China

[2]Vision Medicals Center for Infection Diseases, Guangzhou, Guangdong 510000, China

[3]Key Laboratory of Animal Gene Editing and Animal Cloning in Yunnan Province and College of Veterinary Medicine, Yunnan Agricultural University, Kunming 650201, China

[*]Author H.Z., J.A., and W.Y. contributed equally to this manuscript.

[#] Author T.X., and W.Z. contributed equally to this manuscript.

Corresponding author:

Wenhong Zhang

Department of Infection Diseases, Huashan Hospital, Fudan University

12 Wulumuqi Zhong Road, Shanghai, 200040, China

Email: wenhongzhang_hs@126.com

Summary

A disrupted airway microbiome with frequent potential concurrent infections and a special host immune response was found in patients with COVID-19 compared to other pneumonias. An immune-associated host gene classifier exhibited potential for improving COVID-19 diagnoses and indicating disease severity.

ABSTRACT :

BACKGROUND : The recent identification of a novel coronavirus, also known as SARS-CoV-2, has caused a global outbreak of respiratory illnesses. The rapidly developing pandemic has posed great challenges to diagnosis of this novel infection. However, little is known about the metatranscriptomic characteristics of patients with Coronavirus Disease 2019 (COVID-19).

METHODS: We analyzed metatranscriptomics in 187 patients (62 cases with COVID-19 and 125 with non-COVID-19 pneumonia). Transcriptional aspects of three core elements – pathogens, the microbiome, and host responses – were interrogated. Based on the host transcriptional signature, we built a host gene classifier and examined its potential for diagnosing COVID-19 and indicating disease severity.

RESULTS: The airway microbiome in COVID-19 patients had reduced alpha diversity, with 18 taxa of differential abundance. Potentially pathogenic microbes were also detected in 47% of the COVID-19 cases, 58% of which were respiratory viruses. Host gene analysis revealed a transcriptional signature of 36 differentially expressed genes significantly associated with immune pathways such as cytokine signaling. The host gene classifier built on such a signature exhibited potential for diagnosing COVID-19 (AUC of 0.75-0.89) and indicating disease severity.

CONCLUSIONS: Compared to those with non-COVID-19 pneumonias, COVID-19 patients appeared to have a more disrupted airway microbiome with frequent potential concurrent infections, and a special trigger host immune response in certain pathways such as interferon gamma signaling. The immune-associated host transcriptional signatures of COVID-19 hold promise as a tool for improving COVID-19 diagnosis and indicating disease severity.

Keywords: SARS-CoV-2, COVID-19, Host responses, Metatranscriptomics, Concurrent infection

Abbreviations: COVID-19: Coronavirus Disease 2019; SDI: Shannon Diversity Index; FDR: False Discovery Rate; ROC: Receiver Operating Characteristic Curve; AUC: Area Under Curve; NS: Nasopharyngeal Swabs; SP: Sputum; DEGs: Differential Expressed Genes; TPM: Transcript Per Million; ACE2: Angiotensin Converting Enzyme 2; HMGB3: High-Mobility-Group Protein B Family-3.

INTRODUCTION

An outbreak of serious pneumonias was first reported in Wuhan, China in December 2019 and spread rapidly globally [1-4]. The causative infectious agent was a novel coronavirus, first named 2019-nCoV and then SARS-CoV-2 [1, 5, 6]. As of now, over 80,000 patients in China alone were diagnosed with novel coronavirus pneumonia (COVID-19) [2]. The rapid spread of the COVID-19 pandemic internationally has created an urgent need for improved diagnosis and a better understanding of the infection [7, 8].

SARS-CoV-2 is a highly transmissible and pathogenic respiratory virus [9-11]. Infected patients often present with a wide range of symptoms. Although most patients seem to have mild disease, ~20% may progress to severe illness, including pneumonia, respiratory failure, and even death [8, 12-14]. Recent reports have shown a correlation between poor clinical outcome and disease severity as indicated by oxygen saturation, respiratory rate, blood leukocyte/lymphocyte counts [12, 14-16]. However, little is understood about the molecular mechanisms underlying COVID-19's pathogenesis and prognosis.

Rapid identification of this novel virus at the beginning of the outbreak was achieved by recent advances in clinical metagenomics and metatranscriptomics for infectious diseases [6, 15, 17-19]. By extracting all nucleic acids from a patient's respiratory specimens and subjecting them to high-throughput next-generation sequencing (NGS), an unbiased collection of genomic and transcriptional profiles from pathogens, the microbiome, and the host cells is analyzed [20, 21]. This information can be used for taxonomic classification of known or novel microbes and analyses of host responses [19].

In this study, we employed metatranscriptomic sequencing in 187 patients (62 with COVID-19 and 125 with non-COVID-19 pneumonias) to obtain unbiased, cross-organism transcriptional profiles. We aimed to characterize the three core elements of respiratory infection – pathogens, the airway microbiome, and host response – and apply this information to facilitate better understanding and diagnoses of COVID-19.

RESULTS

Identification of SARS-CoV-2 And Viral Gene Profiles by Metatranscriptomics

An initial cohort of 113 pneumonia cases was enrolled in our study and classified based on the sample type and the status of infection. Of these, 38 COVID-19 cases were diagnosed based on clinical and epidemiologic evidence, and laboratory-confirmed by either RT-PCR or metatranscriptomic assay (Figure 1). In this cohort, 60 samples were obtained by nasopharyngeal swabs (NS) (24 COVID-19 cases and 36 non-COVID-19 cases) and 53 samples of sputum (14 COVID-19 cases and 39 non-COVID-19 cases).

Among the 38 confirmed cases of COVID-19, metatranscriptomic sequencing and RT-PCR achieved the same sensitivity of 87% (33/38) and specificity of 100% (75/75). Despite the overall concordance between these two methods in diagnosis (91.2%, 103/113, Figure 2A), 10 positive cases were missed by one of the two assays, implying of sample-dependent performance.

NGS read numbers mapped to the SARS-CoV-2 genome varied across 6 orders of magnitudes, with a median read number of 1,484, and a range of 2 to 19,016,501. The median genome coverage and sequencing depth was 46.8% (2.8%-100%) and 12.0× (1.0×-7870.1×), respectively (Figure 2B).

The lengths of the viral genes were highly variable; *ORF1a* was the longest (21,290 bp) and *ORF10* was the shortest (117 bp) [5, 22, 23]. Therefore, we measured the viral gene expression by transcript per million (TPM), which included normalization of the length of each gene. As shown in Figure 2C, *N* and *ORF10* had the highest expression, at a level of two-fold or more than most of the other viral genes. *Orf1ab*, commonly used as a target gene in PCR-based SARS-CoV-2 assays, had an expression level in the middle range, but showed the least variation in expression level. As high and consistent expression of detection targets would facilitate better sensitivity and reliability for diagnostic assays, these viral profiles indicate the potential of three genes as candidate targets for developing COVID-19 assays.

Disrupted Microbiome and Potential Co-infection in COVID-19

To fully characterize the microbial profiles in COVID-19, we constructed a database consisting of 18,556 species of bacteria, viruses, fungi, and parasites. Unbiased metatranscriptomic sequencing data were then analyzed using this database for taxonomic classification. Previous studies have demonstrated disruption of the airway microbiome among patients with respiratory infection [24, 25]. To elucidate the potential effect of SARS-CoV-2 on the respiratory microbiome, we analyzed the metatranscriptomic data in the cohort of 113 samples. Although the Shannon diversity index (SDI) in the NS samples remained largely unchanged, that of SP samples was significantly lower in COVID-19 cases (Figure 3A), indicating that COVID-19 patients might have a more disrupted airway microbiome compared to patients with other pneumonias. We identified 31 species in NS samples and 178 species in SP samples with different abundance between COVID-19 and non-COVID-19 cases. Most were less abundant in COVID-19 cases, accounting for 83.9% (26/31) in NS samples and 95.5% (170/178) in SP samples. Furthermore, 18 species were less abundant in both groups, with reductions in $\log_2$fold-change varying from 1.1 to 7.8 in NS samples and from 2.1 to 27.0 in SP samples. No species common to both groups had increased abundance. Compared to the NS samples, most SP samples had a greater degree of reduction (Figure 3B, C and Table S1). We observed correlations between the abundance of certain species and Creactive protein levels, age, sex, and number of NGS reads mapped to the SARS-CoV-2 genome (Figure S1).

To examine possible concurrent infections, we identified 24 microbes with potential pathogenicity in 18 of the 38 (47.4%) COVID-19 patients. These included 16 different microbial species, with *Candida albicans* and *Human alphaherpesvirus 1* being the most frequently detected opportunistic pathogens, and 8 viral pathogens, such as *Human influenza virus* and *Respiratory syncytial viruses* (Figure 3D). We observed more potential co-infections with viruses (58%, 14/24) than with bacteria (25%, 6/24) or fungi (17%, 4/24). These data imply that caution should be taken when ruling out COVID-19 in patients already diagnosed with other infections such as *Human influenza virus*.

In the non-COVID-19 pneumonia cases, we identified a differential spectrum of 76 microbes with pathogenic potential in 52% (39/75) of patients. *Haemophilus parainfluenzae* was the most frequently detected opportunistic pathogen and *Rhinovirus C* the most frequently detected pathogen (Figure S2). Viral, bacterial, and fungal microbes accounted for 45% (34/76), 42% (32/76), and 13% (10/76) of cases, respectively.

Host Transcriptional Signature Associated With Immune Signaling in COVID-19

To better understand the host transcriptional response to SARS-CoV-2 infection, we compared gene expression between COVID-19 and non-COVID-19 subjects. We identified a total of 279 differentially expressed genes (DEGs) in the NS samples, 68.8% (192/279) of which had reduced expression and 31.2% (87/279) had increased expression in COVID-19 cases. The SP group appeared more differentiated in COVID-19, with a total of 4,454 DEGs, 73.1% (3254/4454) and 26.9% (1200/4454) were under- and over-expressed, respectively. A total of 36 common DEGs were shared by both of the groups, with their Log$_2$fold-change in expression ranged from 0.8 to 6.3 in NS and from 1.3 to 7.8 in SP. Among which, 30 had lower expression in COVID-19 and 6 were higher expressed (Figure 4A, Table S2 and S3). Notably, almost all common DEGS, except for *RNR1*, showed significantly greater expression changes in SP than in NS, with an average fold change of 2.7 vs 1.8 ($P<10^{-4}$, Figure 4B). We also examined the expression of angiotensin converting enzyme 2 (ACE2) [26-30], the receptor for SARS-CoV-2, but found no significant changes (Figure S3).

We further performed gene set enrichment analysis with the common DEGs using the KEGG and Reactome databases, and identified 16 differential pathways, half related to immune signaling (Figure 4C). The immune system itself considered as a pathway was significantly overrepresented, with 12 differentially expressed genes ($P<5\times10^{-6}$). Among the subcategories of immune signaling pathways, cytokine signaling was the most significantly deregulated, followed by innate immune system and neutrophil degranulation pathways (Figure 4C, D), indicating the critical roles of innate immune system and cytokine signaling in COVID-19 [31-34]. Dysregulated expression of the interleukin-7 receptor (IL7R) and the interferon-gamma (IFN-γ) pathways (CIITA and HLA-DPB1 genes) were in line with previous reports of coronavirus infection [31].

COVID-19 Diagnosis Improved By Host Transcriptional Classifier

We next sought to build a predictive classifier based on the host transcriptional signature that could aid in COVID-19 diagnosis. These 36 conmen genes along with age and sex were used as classifiers. A receiver-operated curve (ROC) analysis using this classifier of 38 variables yielded areas under the curve (AUCs) of 0.90 and 0.97 in initial cohort of NS group and SP group samples, respectively (Figure 5A). Similar predictive performance with AUCs of 0.91 and 0.96 was also found independent of age and sex as covariates in the classifier (Figure S4)

To further validate the diagnostic performance of this predictive classifier, we enrolled an independent cohort of 74 patients, including 24 laboratory-confirmed COVID-19 cases and 50 non-COVID-19 subjects (Figure 1 and S5). ROC analysis in this validation cohort yielded an AUC of 0.80 (Figure 5A). Consistently, when assessed by 6-fold cross-validation using all of the 187 cases from both cohorts, the host gene classifier again achieved an AUC of 0.82 with a standard deviation of 0.07 (Figure 5A).

Furthermore, the host signature identified 10.5% (4/38) of COVID-19 cases that would had been missed by a metagenomic assay if only analyzing for the presence of the SARS-CoV-2 sequences (Figure 5B). Thus, we designed a diagnostic metric that integrated scoring of all three assays. Although each of the three methods used individually had a false negative rate of 13%, 13%, and 24%, respectively, the integrated diagnosis achieved 100% clinical sensitivity and specificity in our cohort, demonstrating the value of integrating host transcriptional signatures as part of the COVID-19 diagnosis. Our model proposed that caution should be taken to employ further testing when the classifier tests positive, even in the absence of SARS-CoV-2 sequences by metatranscriptomics.

We also categorized the COVID-19 subjects to test the potential of this host gene classifier to discern disease severity. We found a clear segregated clustering of severe and mild COVID-19 cases (Figure 5C). Inter-group comparison also revealed significant differences between non-COVID-19 and COVID-19 subjects, and between the mild and severe COVID-19 cases (Figure 5D, *P*<0.001).

DISCUSSION

In December 2019, a group of patients with pneumonia were identified as infected with a novel coronavirus, known as SARS-CoV-2, in Wuhan, China [1, 3]. Within a few months, COVID-19 aggressively spread to 79 countries and was declared a pandemic by the WHO in March 2020 [7]. SARS-CoV-2 is highly pathogenic and transmissible, and clinical characteristics of COVID-19 differ from those of SARS, MERS, and seasonal influenza [12, 14, 33]. Understanding the pathogenesis is COVID-19 is essential for improving its diagnosis and treatment.

To our knowledge, this is the first study characterizing the microbiome and host response by applying unbiased metatranscriptomic data in COVID-19. We applied RT-PCR and metatranscriptomic sequencing in a cohort of COVID-19 patients, and investigated the three core elements of acute respiratory infections: the pathogen, airway microbiome, and host response. We found reduced diversity in the airway microbiome and identified a host gene signature associated with COVID-19. We further defined a classifier based on such a signature to predict SARS-CoV-2 infection and disease severity.

Previous studies found SARS-CoV-2 made use of a densely glycosylated spike (S) protein to gain entry into host cells [28]. SARS-CoV-2 S binds with high affinity to host cell receptor-angiotensin-converting enzyme 2 (ACE2), a type I membrane protein expressed in lungs, heart, kidneys, and intestine[26-29, 35]. However, the overall host–pathogen interaction network in SARS-CoV-2 infection remains largely unknown. Transcriptomic host heterogeneity has demonstrated its value for studying pathogenesis and disease prognosis in pneumonia [36, 37]. By harnessing an unbiased metatranscriptomic approach, we identified 36 host genes differentially expressed in COVID-19 subjects that were common in both NS and SP groups. For example, high-mobility-group protein B family-3 (HMGB3) was positively correlated with leukocyte and neutrophil counts, and negatively correlated with viral load. Previous studies showed that HMGB is actively released in response to inflammatory stimuli and functions as a danger signal to activate immune responses [38-40]. The HMGB family is expressed in almost all cell types, especially leukocytes. They transduce cellular signals by interacting with important inflammatory receptors such as TLR2 to activate the NF-κB, ERK and p38 pathways, and lead to cytokine production (TNF, IL-6 and IFN-γ) and greater cell survival [38-40]. Future studies are warranted to investigate the HMGB family's role as signaling molecules in COVID-19, especially in the process of cytokine secretion. To explore the potential cell-type specific effect, we also compared COVID-19 and non-COVID-19 cases within each sample type (Figure S3). Despite overlap of many pathways in both groups, some pathways were specifically enriched by sample type. For instance, the interferon pathway was significantly enriched in the COVID-19 cases in the NS group, which might reflect the viral responses of epithelial cells [41, 42]. Enrichment of the interleukin pathway was found in the SP samples, which might indicate viral responses of immune cells [43, 44]. Further studies to investigate the host responses in various cell types will be highly valuable to improve our understanding of COVID-19 pathogenesis.

Although various cytokines (e.g. IL7, tumor necrosis factor-α, TNFα, and IFNγ) have important roles in response to coronavirus infection [45, 46], understanding of their local immune response in the respiratory tract, especially to SARS-CoV-2, remains very limited. Our results showed even higher IFNγ signaling in COVID-19 cases comparison to non-COVID-19 pneumonias, suggesting that IFNγ may be one of the initial immune signals in response to SARS-CoV-2 infection.

By constructing a host gene classifier and assessing its differentiating performance, our results show promise for applying host transcriptional signatures to improve COVID-19 diagnosis and indicate disease severity. When the non-COVID-19 cases were categorized into four subgroups: viral, bacterial, clinically diagnosed infection with unknown pathogen etiology or non-infection pneumonia, we found that the classifier was able to distinguish COVID-19 pneumonia from all forms of non-COVID-19 pneumonia, but could not differentiate among the non-COVID-19 forms (Figure S7). In addition, we did pathway enrichment analyses between COVID-19 and each form of non-COVID-19 pneumonia. Interestingly, among the top 20 differential pathways identified by individual comparisons, 13 were common, many related to immune signaling such as cytokine signaling, innate immune system and neutrophil degranulation (Figure S8). Others have also reported that host transcriptome profiles held promise for diagnosing lower respiratory tract infections with an AUC of 0.88 (95% CI, 0.75-1.00) [37]. In addition, host transcriptional tools have been used to identify various acute respiratory viral infections. Moreover, certain transcriptomic signatures were associated with higher mortality in sepsis patients given corticosteroids, and may reflect immune response state and prognosis in those with severe sepsis [47-49]. Early identification of severely ill patients is essential in COVID-19 infections, where the median duration from ICU admission to death was 7 days (IQR 3–11) among patients who died [50]. Earlier mechanical ventilation or use of antiviral agents may improve disease prognosis.

By harnessing unbiased metatranscriptomic sequencing, we characterized the transcriptional profiles of viral genes, microbiome, and host responses in a cohort of COVID-19 patients. Compared to those with non-COVID-19 pneumonia, we found that COVID-19 patients had a disrupted microbiome with further reduced alpha diversity in COVID-19, and a host gene signature associated with immune signaling. Based on such a signature, we built a host transcriptional classifier that displayed promising performance in diagnosing COVID-19 and indicating disease severity.

MATERIALS AND METHODS

Study Participants

This study included an initial cohort of 113 patients and a validation cohort of 74 patients. For each patient, one respiratory specimen, either a nasopharyngeal swab or sputum, was collected and used for this study. The initial groups of samples consisted of 60 nasopharyngeal swabs (24 COVID-19 cases and 36 non-COVID-19 cases) and 53 sputum specimens (14 COVID-19 cases and 39 non-COVID-19 cases). The validation cohort included 74 sputum specimens (24 COVID-19s and 50 non-COVID-19s). Basic characteristics of the patients are summarized in Table S4 and Table S5. There were no significant differences in sex or age between COVID-19 and non-COVID-19 cases.

Pneumonia patients were confirmed by chest CT before being enrolled into our study. Among these patients, COVID-19 cases were diagnosed based on the WHO interim guidance [51]. According to Chinese guidelines [2], suspected COVID-19 cases were identified as: (a) having pneumonia after a chest CT (with one of the two following criteria met: fever or respiratory symptoms, normal or decreased white blood cell counts, decreased lymphocyte counts), and (b) a travel history or contact with patients with fever or respiratory symptoms from Hubei Province or confirmed cases within 2 weeks. A confirmed COVID-19 case was defined as a positive SARS-COV-2 nucleic acid testing result either by RT-PCR or sequencing-based assay with respiratory specimens. Those who did not meet the clinical and epidemiologic criteria for suspected COVID-19 and also tested negative by metatranscriptomics and RT-PCR assays were defined as non-COVID-19 cases.

Severe cases were defined as having any one of the following: 1) Respiratory distress, respiratory rate ≥ 30 per minute; 2) Pulse oxygen saturation ≤ 93% on room air ; and 3) oxygenation index (PaO_2/FiO_2) ≤300 mmHg [12]. Oral or written consents were obtained from all patients, and the study was approved by the ethics review committee of Huashan Hospital affiliated to Fudan University.

RT-PCR and Metatranscriptomic NGS Sequencing

RNA concentrations were measured by a Qubit Fluorometer (Thermo Fisher Scientific, Carlsbad, CA). Specific RT-PCR assays for COVID-19 were performed using a clinically validated kit (DaAn, Guangzhou, China) on a ABI-7500 Real-Time PCR System (Thermofisher Scientific, Carlsbad, CA) according to the instructions, with Ct values below 40 considered positive.

For metatranscriptomic sequencing, RNA was isolated with the QIAamp ViralRNA Mini kit (Qiagen, Valencia, CA) before being subjected to human rRNA depletion (Vazyme, Nanjing, China). Reverse transcription was performed with N_6 random primers after adaptor ligation with T4 ligase and library amplification. Sequencing was performed on a NextSeq sequencer (Illumina, San Diego, CA). At least 10 million single-end 75bp reads were generated for each sample. Quality control processes included removal of low-complexity, low-quality, and short reads, as well as adapter trimming. Reads derived from the host genome were then removed. Clean reads were aligned against the reference databases, including archaea, bacteria, fungi, protozoa, and viruses. A negative control sample was processed and sequenced in parallel for each sequencing run for contamination and background control.

Microbiome Analysis

The alpha diversity of the respiratory microbiome for each subject was assessed by Shannon Diversity Index (SDI) at the species level using the Vegan package in R (version 3.4.0). Bacteria were included in diversity analyses. Diversity values were then compared between patients with and without SARS-CoV-2 infection within each group using the Wilcoxon rank-sum test [52]. Species with differential abundance were identified within each group using DESeq2 in R at a false discovery rate (FDR) ≤ 0.1, fold change ≥ 2, and $P \leq 0.05$.

Host Gene Expression Analysis

Alpha diversity of the host transcriptional profile for each subject was assessed by SDI at the gene level using the Vegan package in R (version 3.4.0). SDIs between patients with and without SARS-CoV-2 infection were then compared within NS and SP samples separately using the Wilcoxon rank-sum test [52]. Differential expressed genes (DEGs) were identified in each group using DESeq2 in R with FDR ≤ 0.1, fold change ≥ 1.5, and $P \leq 0.05$. Gene set enrichment analysis was conducted using the Kobas program (http://kobas.cbi.pku.edu.cn/kobas3/genelist/) for differentially expressed genes that were over-represented in the KEGG and Reactome pathways. Pathways or biological processes with $P < 0.05$ by Fisher's exact test after Benjamini and Hochberg adjustment were considered significantly enriched [53].

Classifier Analysis

Classifiers were built based on the 36 common DEGs and sample type, with or without age and sex as covariates. Elastic–net penalty logistic regression implementation from the Python-based sklearn module was used to generate receiver operating characteristic (ROC) and area under the curve (AUC) calculations, with a regularization parameter α of 0.5 and a six-fold cross-validation strategy [54]. Wilcoxon rank-sum tests were used for statistical assessment of differences in probability value in each group.

Statistical Analysis

Statistical significance was defined as $P < 0.05$, using two-tailed tests of hypotheses unless indicated otherwise. Comparative analyses were conducted using Pearson $\chi 2$ tests, Fisher's exact tests, Student's t-tests, or log-rank tests where appropriate as specified.

Data Availability

Raw microbial sequences are available via GISAID accession ID CNP0001055.

Potential Conflicts of Interests

The authors declare no competing interests.

REFERENCES

1. Salata C, A Calistri, C Parolin and G Palù. Coronaviruses: a paradigm of new emerging zoonotic diseases. Pathog Dis. **2019**. 77(9).

2. National Health Commission of the People's Republic China. 2019 Novel Coronavirus. **2020**. Retrieved 1, 36, from http://www.nhc.gov.cn.

3. Tan W, X Zhao, X Ma, *et al.* A novel coronavirus genome identified in a cluster of pneumonia cases—Wuhan, China 2019− 2020. China CDC Weekly. **2020**. 2(4): 61-62.

4. Acter T, N Uddin, J Das, *et al.* Evolution of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) as coronavirus disease 2019 (COVID-19) pandemic: A global health emergency. Sci Total Environ. **2020**. 730: 138996.

5. Zhou P, X-L Yang, X-G Wang, *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. **2020**. 1-4.

6. Zhu N, D Zhang, W Wang, *et al.* A novel coronavirus from patients with pneumonia in China, 2019. New England Journal of Medicine. **2020**. 382(8):727-733.

7. WHO. Latest rolling update: WHO characterizes COVID-19 as a pandemic. **2020**. from https://www.who.int.

8. Baloch S, M A Baloch, T Zheng and X Pei. The Coronavirus Disease 2019 (COVID-19) Pandemic. Tohoku J Exp Med. **2020**. 250(4): 271-278.

9. Li Q, X Guan, P Wu, *et al.* Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. New England Journal of Medicine. **2020**. 382(13):1199-1207.

10. Bulut C and Y Kato. Epidemiology of COVID-19. Turk J Med Sci. **2020**. 50(Si-1): 563-570.

11. Riou J and C Althaus. Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020. Eurosurveillance. **2020**. 25(4).

12. Guan W-j, Z-y Ni, Y Hu, *et al.* Clinical characteristics of coronavirus disease 2019 in China. New England Journal of Medicine. **2020**. 382(18):1708-1720.

13. Chen N, M Zhou, X Dong, *et al.* Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. The Lancet.

2020. 395(10223): 507-513.

14. Wang D, B Hu, C Hu, *et al.* Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. Jama. **2020**. 323(11): 1061-1069.

15. Ai J-W, H-C Zhang, T Xu, *et al.* Optimizing diagnostic strategy for novel coronavirus pneumonia, a multi-center study in Eastern China. **2020**. 10.1101/2020.02.13.20022673.

16. Zhang X, Tan Y, Ling Y, *et al*. Viral and host factors related to the clinical outcome of COVID-19. Nature. **2020**. 10.1038/s41586-020-2355-0.

17. Wilson M R, S N Naccache, E Samayoa, *et al.* Actionable diagnosis of neuroleptospirosis by next-generation sequencing. N Engl J Med. **2014**. 370(25): 2408-2417.

18. Gu W, S Miller and C Chiu. Clinical metagenomic next-generation sequencing for pathogen detection. Annual Review of Pathology: Mechanisms of Disease. **2019**. 14: 319-338.

19. Ren L-L, Y-M Wang, Z-Q Wu, *et al.* Identification of a novel coronavirus causing severe pneumonia in human: a descriptive study. Chinese medical journal. **2020**. 10.3760/cma.j.issn.0366-6999.2020.00.E001.

20. Avraham R, N Haseley, A Fan, *et al.* A highly multiplexed and sensitive RNA-seq protocol for simultaneous analysis of host and pathogen transcriptomes. Nature protocols. **2016**. 11(8): 1477.

21. Westermann A J, S A Gorski and Vogel. Dual RNA-seq of pathogen and host. Nature Reviews Microbiology. **2012**. 10(9): 618-630.

22. Hussain S, J Pan, Y Chen, *et al.* Identification of novel subgenomic RNAs and noncanonical transcription initiation signals of severe acute respiratory syndrome coronavirus. J Virol. **2005**. 79(9): 5288-5295.

23. Masters P S. The molecular biology of coronaviruses. Adv Virus Res. **2006**. 66: 193-292.

24. Hanada S, M Pirzadeh, K Y Carver and J C Deng. Respiratory Viral Infection-Induced Microbiome Alterations and Secondary Bacterial Pneumonia. Front Immunol. **2018**. 9: 2640.

25. Groves H T, L Cuthbertson, P James, *et al.* Respiratory Disease following Viral Lung Infection Alters the Murine Gut Microbiota. Front Immunol. **2018**. 9: 182.

26. Xia S, Y Zhu, M Liu, *et al.* Fusion mechanism of 2019-nCoV and fusion inhibitors targeting HR1 domain in spike protein. Cell Mol Immunol. **2020**. 1-3.

27. Ashour H M, W F Elkhatib, M M Rahman and H A Elshabrawy. Insights into the Recent 2019 Novel Coronavirus (SARS-CoV-2) in Light of Past Human Coronavirus Outbreaks. Pathogens. **2020**. 9(3).

28. Hoffmann M, H Kleine-Weber, S Schroeder, *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. Cell. **2020**. 181(2): 271-280.e278.

29. Ou X, Y Liu, X Lei, *et al.* Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. Nat Commun. **2020**. 11(1): 1620.

30. Wrapp D, N Wang, K S Corbett, *et al.* Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science. **2020**. 367(6483): 1260-1263.

31. Ye Q, B Wang and J Mao. The pathogenesis and treatment of the `Cytokine Storm' in COVID-19. J Infect. **2020**.

32. Tufan A, A Avanoğlu Güler and M Matucci-Cerinic. COVID-19, immune system response, hyperinflammation and repurposing antirheumatic drugs. Turk J Med Sci. **2020**. 50(Si-1): 620-632.

33. Huang C, Y Wang, X Li, *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. **2020**. 395(10223): 497-506.

34. Qin C, L Zhou, Z Hu, *et al.* Dysregulation of immune response in patients with COVID-19 in Wuhan, China. Clin Infect Dis. **2020**. 10.1093/cid/ciaa248.

35. Li H, L Liu, D Zhang, *et al.* SARS-CoV-2 and viral sepsis: observations and hypotheses. Lancet. **2020**. 395(10235): 1517-1520.

36. Herberg J A, M Kaforou, S Gormley, *et al.* Transcriptomic profiling in childhood H1N1/09 influenza reveals reduced expression of protein synthesis genes. The Journal of infectious diseases. **2013**. 208(10): 1664-1668.

37. Langelier C, K L Kalantar, F Moazed, *et al.* Integrating host response and unbiased microbe detection for lower respiratory tract infection diagnosis in critically ill adults. Proceedings of the National Academy of Sciences. **2018**. 115(52): E12353-E12362.

38. Lotze M T and K J Tracey. High-mobility group box 1 protein (HMGB1): nuclear weapon in the immune arsenal. Nat Rev Immunol. **2005**. 5(4): 331-342.

39. Kang R, R Chen, Q Zhang, *et al.* HMGB1 in health and disease. Mol Aspects Med. **2014**.

40: 1-116.

40. Avgousti D C, C Herrmann, K Kulej, *et al.* A core viral protein binds host nucleosomes to sequester immune danger signals. Nature. **2016**. 535(7610): 173-177.

41. Hillyer P, R Shepard, M Uehling, *et al.* Differential Responses by Human Respiratory Epithelial Cell Lines to Respiratory Syncytial Virus Reflect Distinct Patterns of Infection Control. J Virol. **2018**. 92(15).

42. Lee S, M Hirohama, M Noguchi, K Nagata and A Kawaguchi. Influenza A Virus Infection Triggers Pyroptosis and Apoptosis of Respiratory Epithelial Cells through the Type I Interferon Signaling Pathway in a Mutually Exclusive Manner. J Virol. **2018**. 92(14).

43. Rouse B T, P P Sarangi and S Suvas. Regulatory T cells in virus infections. Immunol Rev. **2006**. 212: 272-286.

44. van Leeuwen E M, G J de Bree, I J ten Berge and R A van Lier. Human virus-specific CD8+ T cells: diversity specialists. Immunol Rev. **2006**. 211: 225-235.

45. Li G, Y Fan, Y Lai, *et al.* Coronavirus infections and immune responses. J Med Virol. **2020**. 92(4): 424-432.

46. Merad M and J C Martin. Pathological inflammation in patients with COVID-19: a key role for monocytes and macrophages. Nat Rev Immunol. **2020**. 1-8.

47. Walter N D, M A Miller, J Vasquez, *et al.* Blood Transcriptional Biomarkers for Active Tuberculosis among Patients in the United States: a Case-Control Study with Systematic Cross-Classifier Evaluation. J Clin Microbiol. **2016**. 54(2): 274-282.

48. Walter N D, R Reves and J L Davis. Blood transcriptional signatures for tuberculosis diagnosis: a glass half-empty perspective. Lancet Respir Med. **2016**. 4(6): e28.

49. Antcliffe D B, K L Burnham, F Al-Beidh, *et al.* Transcriptomic Signatures in Sepsis and a Differential Response to Steroids. From the VANISH Randomized Trial. Am J Respir Crit Care Med. **2019**. 199(8): 980-986.

50. Yang X, Y Yu, J Xu, *et al.* Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. Lancet Respir Med. **2020**. 8(5): 475-481.

51. WHO. Clinical management of severe acute respiratory infection when Novel coronavirus (nCoV) infection is suspected: interim guidance. publications detail/clinical management of severe acute respiratory infection when novel coronavirus (ncov) infection is suspected (accessed Jan 20, 2020). **2020**. from https://www.who.int/internal.

52. Liao C, S Li and Z Luo (**2007**). Gene Selection for Cancer Classification using Wilcoxon Rank Sum Test and Support Vector Machine. 2006 International Conference on Computational Intelligence and Security.

53. Xie C, X Mao, J Huang, *et al.* KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic acids research. **2011**. 39(suppl_2): W316-W322.

54. Géron A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, O'Reilly Media.**2019**.

Figure Legends

Figure 1. Summary of study cohorts.

Figure 2. (A) Pie diagram showing the results by two different nucleic-acid assays for SARS-CoV-2 diagnosis; (B) Detection of SARS-CoV-2 by metatranscriptomic sequencing as shown by the number of mapped reads, viral genome coverage and sequencing depth; (C) Transcriptional profile of SARS-CoV-2 genes measured by TMP.

Figure 3. (A) Comparison of alpha diversity in the microbiome between SARS-CoV-2 positive and negative cases in each group; (B-C) Species with differential abundance common in both NS and SP groups and their respective fold change (B) and normalized abundance (C). **, *P*<0.01; (D) Potential concurrent infection identified in COVID-19 by metatranscriptomics. Percentages of COVID-19 cases in the presence or absence of potential pathogenic microbes (upper panel), and the frequency of each microbe (lower panel).

Figure 4. (A-B) Gene with differential expression common in both NS and SP groups and their respective fold change (A) and normalized abundance (B); (C) Gene set enrichment analysis using KEGG and Reactome databases; (D) Deregulated host genes and immue-pathways in COVID-19.

Figure 5. (A) ROC curves by elastic-net penalty logistic regression showing host classifier performance in the NS group, NP groups, the independent validation cohort and the consolidated cohort of all specimens by six-fold validation; (B) Diagnostic metric of three assays for COVID-19, in which a sample with negative viral result by sequencing but positive by host signature would receive a score of 0.5 and call for PCR validation. Black and white cells indicate positive and negative test results, respectively. PCR or sequencing; (C) Samples were ranked from the lowest to the highest logit probability defined by the classifier. Those with inconsistent PCR/metatranscriptomic results (i.e., false negative in one of the assays) were arrowed. (D) Logit probability among different groups were compared. ***, *P*<0.001.

Figure 1

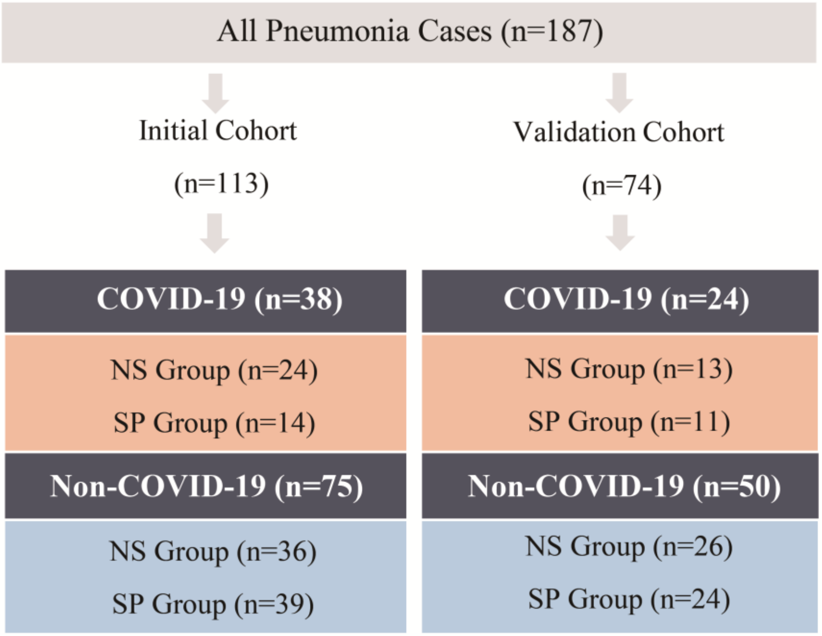
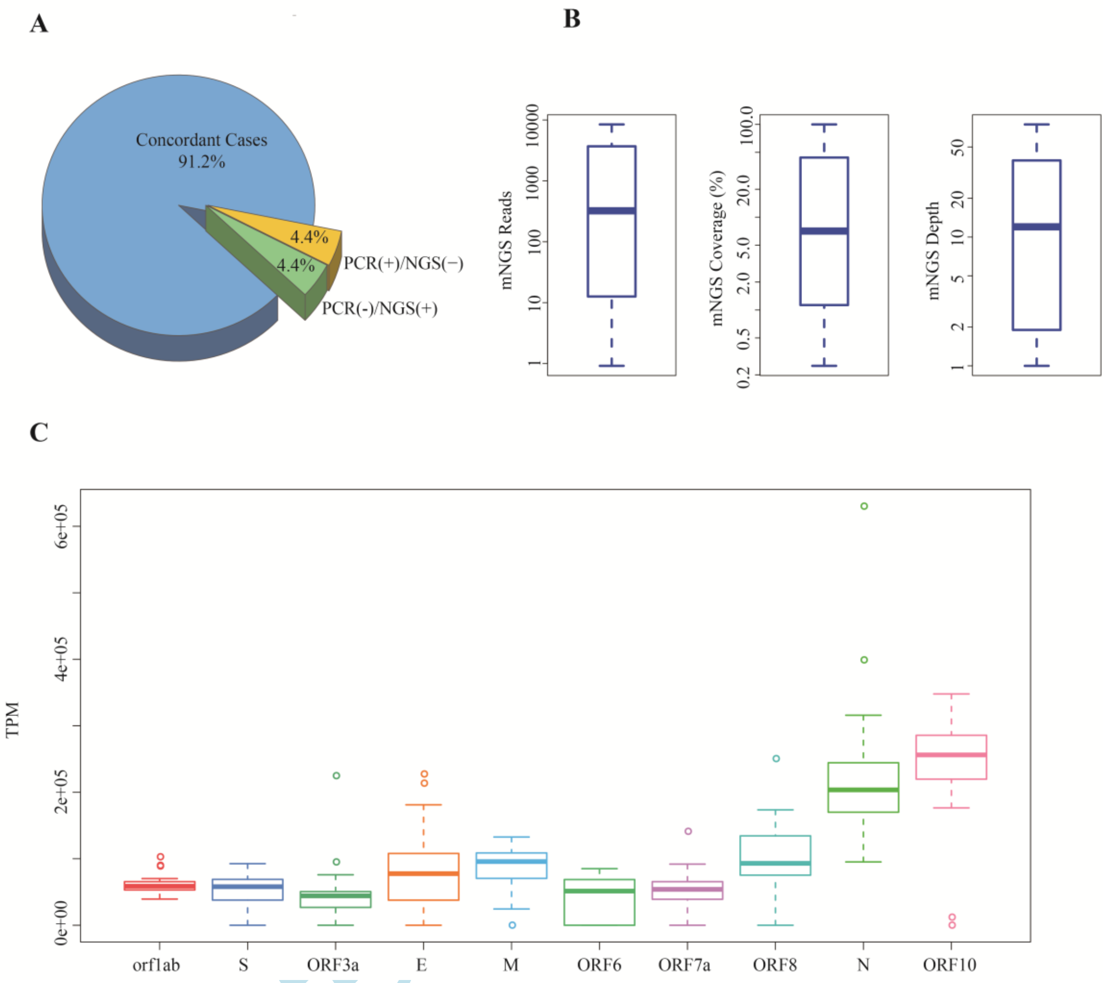

All Pneumonia Cases (n=187)
Initial Cohort
(n=113)
Validation Cohort
(n=74)
COVID-19 (n=38)
NS Group (n=24)
SP Group (n=14)
Non-COVID-19 (n=75)
NS Group (n=36)
SP Group (n=39)
COVID-19 (n=24)
NS Group (n=13)
SP Group (n=11)
Non-COVID-19 (n=50)
NS Group (n=26)
SP Group (n=24)

Figure 2



A
Concordant Cases
91.2%
4.4%
4.4%
PCR(+)/NGS(−)
PCR(-)/NGS(+)
B
mNGS Reads
1
10
100
1000
10000
mNGS Coverage (%)
0.2
0.5
2.0
5.0
20.0
100.0
mNGS Depth
1
2
5
10
20
50
C
TPM
0e+00
2e+05
4e+05
6e+05
orf1ab
S
ORF3a
E
M
ORF6
ORF7a
ORF8
N
ORF10

Figure 3

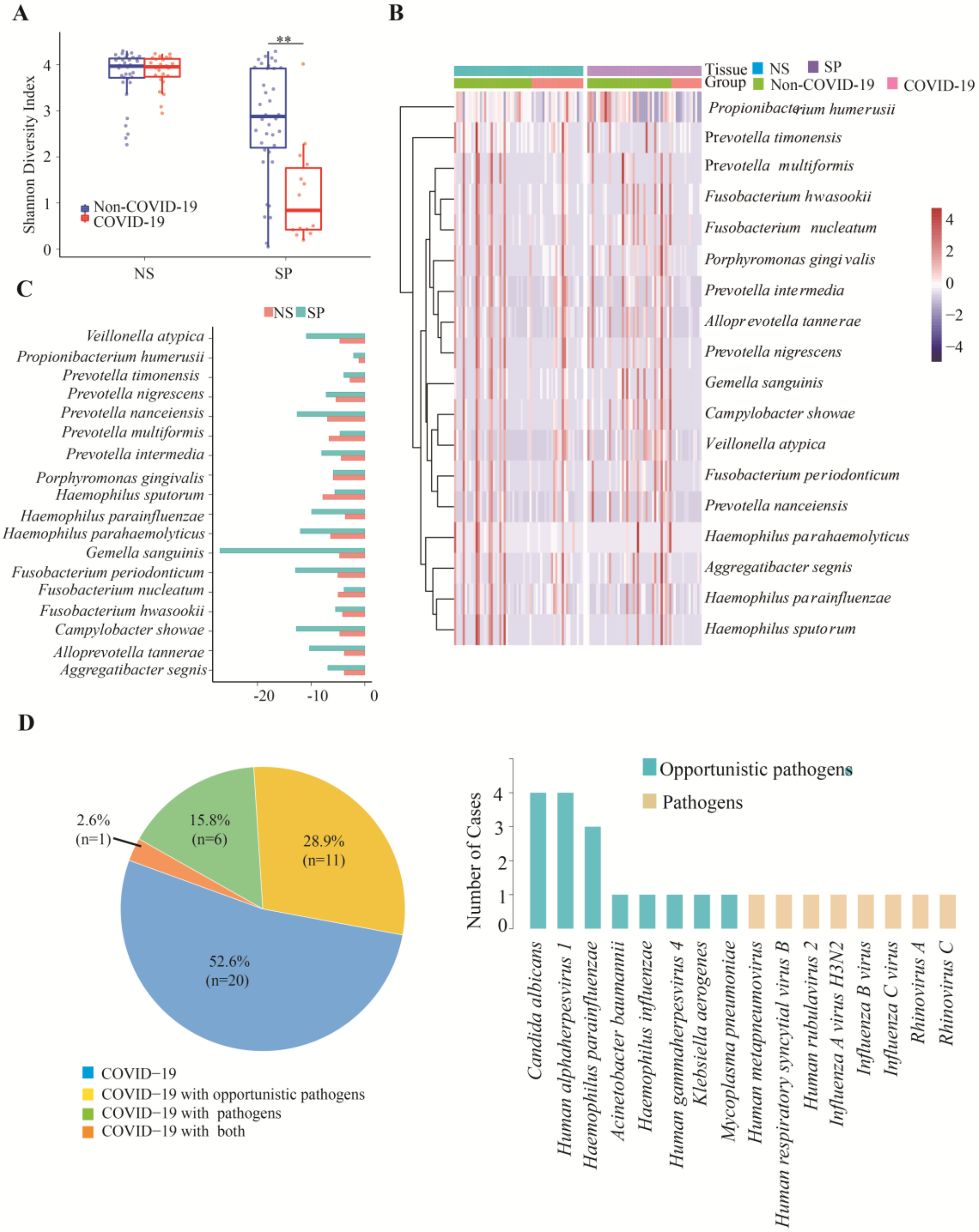

A
Shannon Diversity Index
0
1
2
3
4
**
Non-COVID-19
COVID-19
NS
SP
B
Tissue NS SP
Group Non-COVID-19 COVID-19
Propionibacterium humerusii
Prevotella timonensis
Prevotella multiformis
Fusobacterium hwasookii
Fusobacterium nucleatum
Porphyromonas gingivalis
Prevotella intermedia
Alloprevotella tannerae
Prevotella nigrescens
Gemella sanguinis
Campylobacter showae
Veillonella atypica
Fusobacterium periodonticum
Prevotella nanceiensis
Haemophilus parahaemolyticus
Aggregatibacter segnis
Haemophilus parainfluenzae
Haemophilus sputorum
4
2
0
−2
−4
C
NS SP
Veillonella atypica
Propionibacterium humerusii
Prevotella timonensis
Prevotella nigrescens
Prevotella nanceiensis
Prevotella multiformis
Prevotella intermedia
Porphyromonas gingivalis
Haemophilus sputorum
Haemophilus parainfluenzae
Haemophilus parahaemolyticus
Gemella sanguinis
Fusobacterium periodonticum
Fusobacterium nucleatum
Fusobacterium hwasookii
Campylobacter showae
Allopreovotella tannerae
Aggregatibacter segnis
-20
-10
0
D
2.6% (n=1)
15.8% (n=6)
28.9% (n=11)
52.6% (n=20)
COVID−19
COVID−19 with opportunistic pathogens
COVID−19 with pathogens
COVID−19 with both
Number of Cases
0
1
2
3
4
Opportunistic pathogens
Pathogens
Candida albicans
Human alphaherpesvirus 1
Haemophilus parainfluenzae
Acinetobacter baumannii
Haemophilus influenzae
Human gammaherpesvirus 4
Klebsiella aerogenes
Mycoplasma pneumoniae
Human metapneumovirus
Human respiratory syncytial virus B
Human rubulavirus 2
Influenza A virus H3N2
Influenza B virus
Influenza C virus
Rhinovirus A
Rhinovirus C

Figure 4

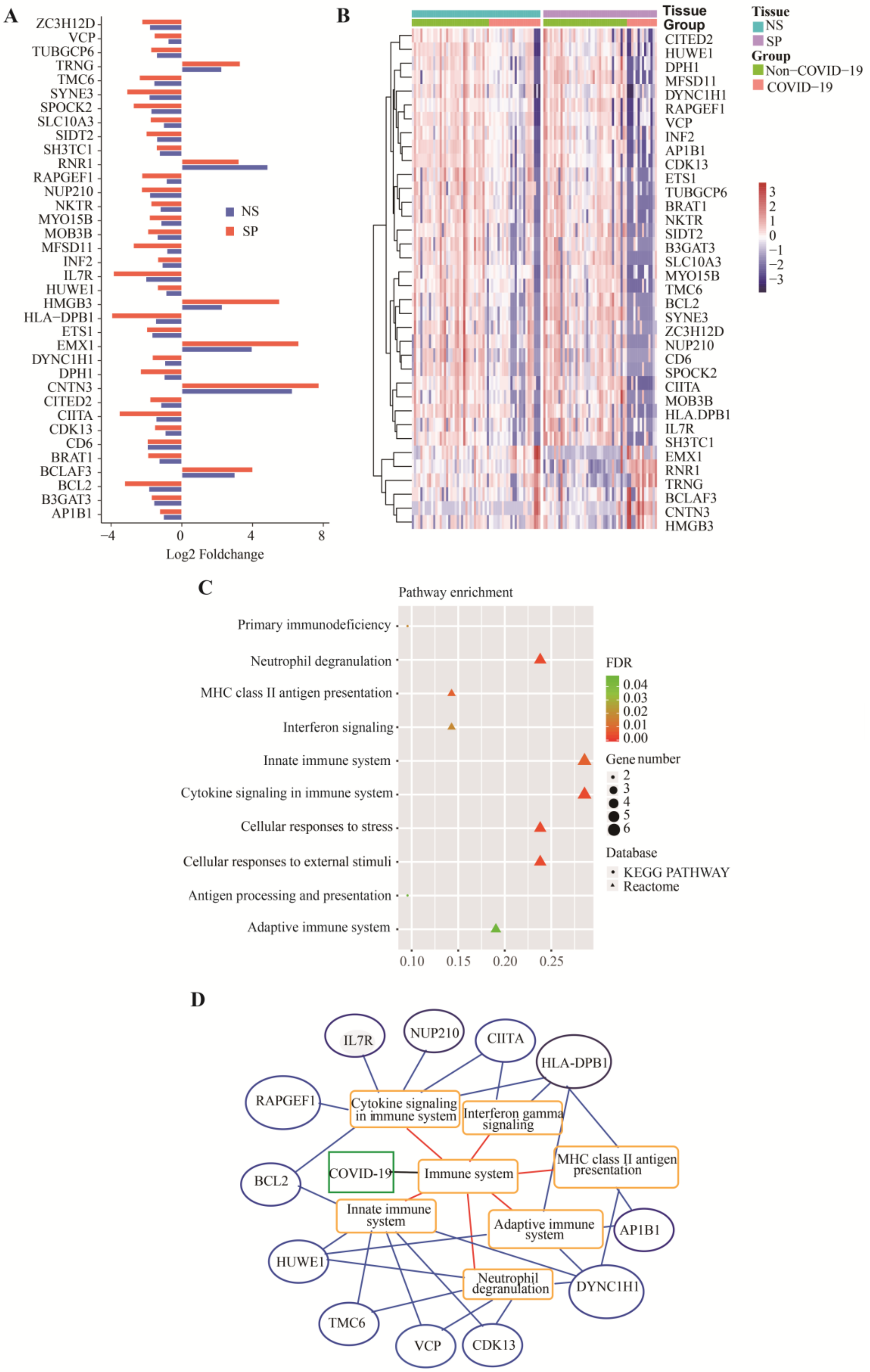
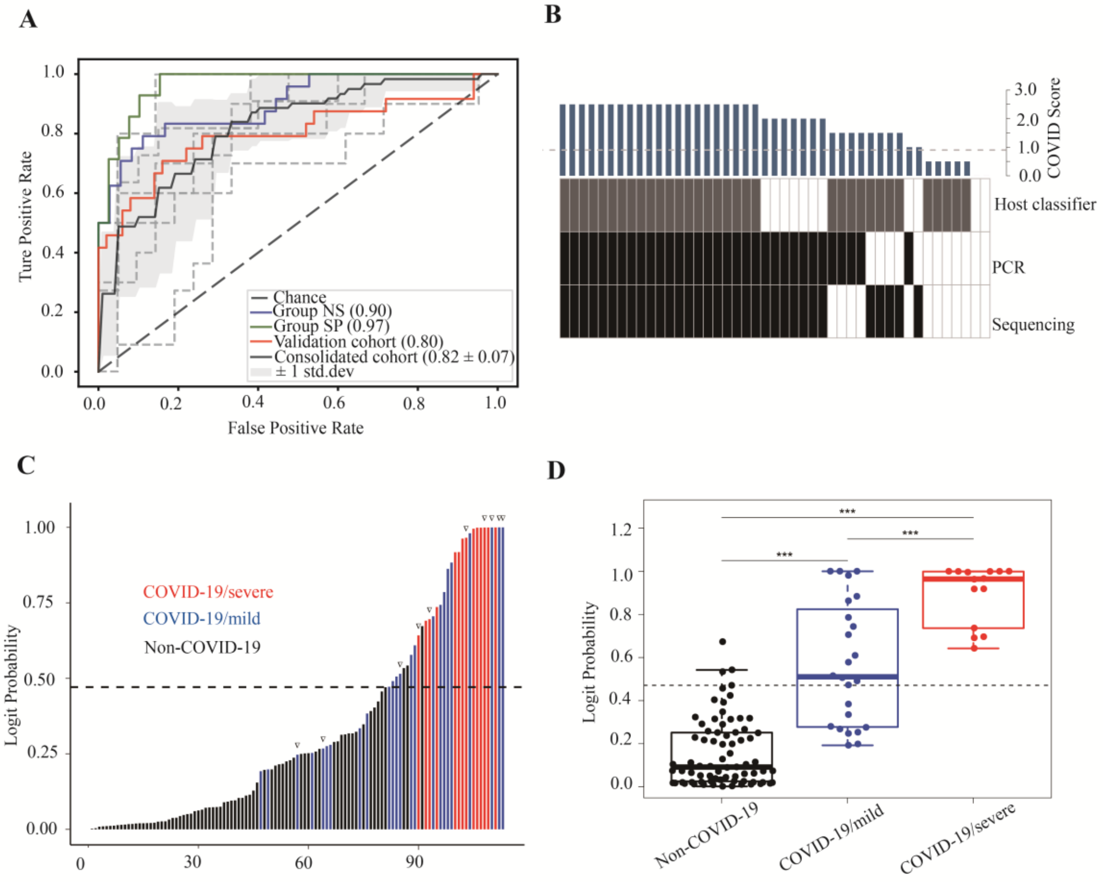

A
ZC3H12D
VCP
TUBGCP6
TRNG
TMC6
SYNE3
SPOCK2
SLC10A3
SIDT2
SH3TC1
RNR1
RAPGEF1
NUP210
NKTR
MYO15B
MOB3B
MFSD11
INF2
IL7R
HUWE1
HMGB3
HLA−DPB1
ETS1
EMX1
DYNC1H1
DPH1
CNTN3
CITED2
CIITA
CDK13
CD6
BRAT1
BCLAF3
BCL2
B3GAT3
AP1B1
NS
SP
−4
0
4
8
Log2 Foldchange
B
Tissue
Group
CITED2
HUWE1
DPH1
MFSD11
DYNC1H1
RAPGEF1
VCP
INF2
AP1B1
CDK13
ETS1
TUBGCP6
BRAT1
NKTR
SIDT2
B3GAT3
SLC10A3
MYO15B
TMC6
BCL2
SYNE3
ZC3H12D
NUP210
CD6
SPOCK2
CIITA
MOB3B
HLA.DPB1
IL7R
SH3TC1
EMX1
RNR1
TRNG
BCLAF3
CNTN3
HMGB3
Tissue
NS
SP
Group
Non−COVID−19
COVID−19
3
2
1
0
−1
−2
−3
C
Pathway enrichment
Primary immunodeficiency
Neutrophil degranulation
MHC class II antigen presentation
Interferon signaling
Innate immune system
Cytokine signaling in immune system
Cellular responses to stress
Cellular responses to external stimuli
Antigen processing and presentation
Adaptive immune system
0.10
0.15
0.20
0.25
FDR
0.04
0.03
0.02
0.01
0.00
Gene number
2
3
4
5
6
Database
KEGG PATHWAY
Reactome
D
IL7R
NUP210
CIITA
HLA-DPB1
RAPGEF1
Cytokine signaling in immune system
Interferon gamma signaling
BCL2
COVID-19
Immune system
MHC class II antigen presentation
Innate immune system
Adaptive immune system
AP1B1
HUWE1
Neutrophil degranulation
DYNC1H1
TMC6
VCP
CDK13

Figure 5


A
1.0
0.8
0.6
0.4
0.2
0.0
Ture Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
Chance
Group NS (0.90)
Group SP (0.97)
Validation cohort (0.80)
Consolidated cohort (0.82 ± 0.07)
± 1 std.dev
B
3.0
2.0
1.0
0.0
COVID Score
Host classifier
PCR
Sequencing
C
1.00
0.75
0.50
0.25
0.00
Logit Probability
0
30
60
90
COVID-19/severe
COVID-19/mild
Non-COVID-19
D
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Logit Probability
***
***
***
Non-COVID-19
COVID-19/mild
COVID-19/severe