Genomic diversity of SARS-CoV-2 in Coronavirus Disease 2019 patients

Zijie Shen [1,2#], Yan Xiao [3#], Lu Kang [1,2#], Wentai Ma [1,2#], Leisheng Shi [1,2], Li Zhang[1], Zhuo Zhou[4], Jing Yang [1,2], Jiaxin Zhong [1,2], Donghong Yang[5], Li Guo[3], Guoliang Zhang[6], Hongru Li[7], Yu Xu[5], Mingwei Chen[8], Zhancheng Gao[5], Jianwei Wang [3], Lili Ren [3*], Mingkun Li[1,9*].

1. Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, and China National Center for Bioinformation, Beijing, 101300, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China.
3. National Health Commission of the People's Republic of China Key Laboratory of Systems Biology of Pathogens and Christophe Mérieux Laboratory, Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100730, China;
4. Biomedical Pioneering Innovation Center, Beijing Advanced Innovation Center for Genomics, Peking-Tsinghua Center for Life Sciences, Peking University Genome Editing Research Center, State Key Laboratory of Protein and Plant Gene Research, School of Life Sciences, Peking University, Beijing 100871, China
5. Department of Respiratory and Critical Care Medicine, Peking University People's Hospital, Beijing 100044, China
6. National Clinical Research Center for Infectious Diseases, Guangdong Key Laboratory for Emerging Infectious Diseases, Shenzhen Third People's Hospital, Southern University of Science and Technology, Shenzhen 518112, China
7. Fujian Provincial Hospital, Fujian 350000, PR China
8. Department of Respiratory and Critical Care Medicine, the First Affiliated Hospital of Xi'an Jiaotong University, Shaanxi Province 710061, P.R. China
9. Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, 650223, China;

# Author Z.S., Y.X., L.K., and W.M. contributed equally to this manuscript.

* Author L.R. and M.L. contributed equally to this manuscript.

Corresponding author:

Mingkun Li

E-mail: limk@big.ac.cn

Beijing Institute of Genomics, Chinese Academy of Sciences

No. 1-104, Beichen West Road, Chaoyang District, Beijing, 100101, China

Tel/Fax: 86-10-84097716

Summary

An elevated level of viral diversity was found in some SARS-CoV-2 infected patients, indicating the risk of rapid evolution of the virus. Although no evidence for the transmission of intra-host variants was found, the risk should not be overlooked.

Abstract:

Background A novel coronavirus (SARS-CoV-2) has infected more than 75,000 individuals and spread to over 20 countries. It is still unclear how fast the virus evolved and how the virus interacts with other microorganisms in the lung.

Methods We have conducted metatranscriptome sequencing for the bronchoalveolar lavage fluid of eight SARS-CoV-2 patients, 25 community-acquired pneumonia (CAP) patients, and 20 healthy controls.

Results The median number of intra-host variants was 1-4 in SARS-CoV-2 infected patients, which ranged between 0 and 51 in different samples. The distribution of variants on genes was similar to those observed in the population data (110 sequences). However, very few intra-host variants were observed in the population as polymorphism, implying either a bottleneck or purifying selection involved in the transmission of the virus, or a consequence of the limited diversity represented in the current polymorphism data. Although current evidence did not support the transmission of intra-host variants in a person-to-person spread, the risk should not be overlooked. The microbiota in SARS-CoV-2 infected patients was similar to those in CAP, either dominated by the pathogens or with elevated levels of oral and upper respiratory commensal bacteria.

Conclusion SARS-CoV-2 evolves *in vivo* after infection, which may affect its virulence, infectivity, and transmissibility. Although how the intra-host variant spreads in the population is still elusive, it is necessary to strengthen the surveillance of the viral evolution in the population and associated clinical changes.

Keywords: SARS-CoV-2, COVID-19, intra-host variant, microbiota, transmission

Abbreviations: COVID-19: Coronavirus disease 2019; CAP: Community-acquired pneumonia; SARS: Severe acute respiratory syndrome; CoV: Coronavirus; Healthy: Healthy controls; BALF: Bronchoalveolar lavage fluid; nCoV: COVID-19 patients; NC: Negative controls.

Introduction

Since the outbreak of a novel coronavirus (SARS-CoV-2) in Wuhan, China, the virus had spread to more than 20 countries, resulting in over 75,000 cases and more than 2,300 deaths (Until Feb 22, 2020) [1, 2]. The basic reproduction number was estimated to range from 2.2 to 3.5 at the early stage[3], making it a severe threat to public health. Recent studies have identified bat as the possible origin of SARS-CoV-2, and the virus likely uses the same cell surface receptor as SARS-CoV [4], namely ACE2. These studies have advanced our understanding of SARS-CoV-2. However, our knowledge of the novel virus is still limited.

The virus undergoes a strong immunologic pressure in humans, and may thus accumulate mutations to outmaneuver the immune system [5]. These mutations could result in changes in viral virulence, infectivity, and transmissibility [6]. Therefore, it is imperative to investigate the pattern and frequency of mutations occurred. Aside from the pathogen, microbiota in the lung is associated with disease susceptibility and severity [7]. Alterations of lung microbiota could potentially modify immune response against the viral and secondary bacterial infection [8, 9]. Thus, understanding the microbiota, which comprises bacteria that could cause secondary infection or exert effects on the mucosal immune system, might help to predict the outcome and reduce complications.

In our study, we conducted metatranscriptome sequencing on bronchoalveolar lavage fluid (BALF) samples from 8 subjects with Coronavirus disease 2019 (COVID-19, the disease caused by SARS-CoV-2) patients. We found that the number

of intra-host variants ranged from 0 to 51 with a median number of 4, suggesting a high evolution rate of the virus. By investigating a person-to-person spread event, we found no evidence for the transmission of intra-host variants. Meanwhile, we found no specific microbiota alteration in the BALF of COVID-19 patients comparing to CAP patients with other suspected viral causes.

Results

Data summary

By metatranscriptome sequencing, more than 20 million reads were generated for each BALF of COVID-19 patients (nCoV) as well as a negative control (nuclease-free water, NC). For comparison, the metatranscriptome sequencing data with similar number of reads from 25 virus-like community-acquired pneumonia patients (CAP, determined by at least 100 reads and 10-fold higher than those in the NC), 20 healthy controls without any known pulmonary diseases (Healthy), and two extra NCs (two saline solutions passing through the bronchoscope) were used in this study. Demographic and clinical information was collected and summarized in Supplementary Table 1.

After quality control, a median number of 55,571 microbial reads were generated for each sample. nCoV had the highest proportion of microbial reads compared to CAP and Healthy (nCoV: median proportion of 7%, CAP: 0.8%, Healthy:0.1%, $p < 0.001$, Figure 1A), and 49% of the microbial reads could be mapped to SARS-CoV-2, which was not different from the viral proportion in CAP (Figure 1B). Only SARS-CoV-2 was identified in nCoV, and no read was mapped to other species

belonging to *Betacoronavirus*. Moreover, besides the detection of HCoV-OC43 in one Healthy and HCoV-NL63 in a CAP, no other samples showed any signal of *Betacoronavirus*, which proved the authenticity of the data and methods used in our analysis.

High level of intra-host variants in SARS-CoV-2 patients

The sequencing depth of SARS-CoV-2 ranged from 18-fold in nCoV-5 to 32,291-fold in nCoV1, with more than 80% of the genome covered by at least 50-fold in five samples (Figure 2A, Supplementary Table 2). In total, 84 intra-host variants were identified with minor allele frequency (MAF) greater than 5%, and 25 variants were with MAF greater than 20% (Supplementary Table 3, Figure 2B, nCoV5 was excluded from the analysis due to large gaps on its genome coverage). Notably, the number of variants was not associated with the sequencing depth (Supplementary Figure 1). The overall Ka/Ks ratio was significantly smaller than 1, which was similar for intra-host variants and the polymorphisms observed in the population data, suggesting a purifying selection acting on both types of mutations (Table 1). The numbers of variants observed in the gene were proportional to gene lengths (cor = 0.950, p = 8E-06 for the intra-host variant; cor = 0.957, p = 4E-06 for the polymorphisms) . Although only a small fraction of the variants was observed in multiple patients (2 out of 84, Figure 2C), some positions were more prone to mutate, such as position 10779, where the mutant allele A was observed in all seven patients, with the frequency ranging from 15% to 100% (Figure 2D).

The number of intra-host variants per individual showed a large variation (0 to 51,

median 4 for variants with MAF ≥ 5%; 0 to 19, median 1 for variants with MAF ≥ 20%), which could not be explained by the batch effect, coverage variance, or contamination (Supplementary Figure 1; nCoV1-4 were in one batch, nCoV5-8 were in another batch; most mutations were not observed in the population data). We also noted that the number of variations was not relevant to the days after symptom onset or the age of patients (Supplementary Figure 2). Collectively, we did not find any reason for the extremely high level of variants in nCoV6 (51 variants). A larger population size is needed to investigate how frequent such outliers are, and whether they are associated with the level of host immune response or the viral replication rate. We also noted similar outliers for other viruses [11]. Of note, the origin of variants could be either mutation occurred *in vivo* after infection or multiple transmitted SARS-CoV-2 strains.

No evidence for transmission of intra-host variants between samples

Among the eight COVID-19 patients, nCoV4 and nCoV7 were from the same household, with dates of symptom onset differing by five days; thus a transmission from nCoV4 to nCoV7 is highly suspected, especially considering that only nCoV4 had been to the Huanan seafood market in Wuhan, which is the starting point of the outbreak and suspected to be the source. First, the consensus sequence of the virus was the same for two samples, and all four intra-host variants passing the selection criteria in nCoV4 were not detected in nCoV7 (Table 2). We further expanded the investigation to all variants with MAF ≥ 2% and supported by at least 3 reads. By doing so, we detected seven variants (out of 25) shared between the two samples.

However, the MAF in both nCoV4 and nCoV7 were similar to those in other samples, suggesting that these positions were either error-prone or mutation-prone; hence they cannot support the transmission of these variants.

Meanwhile, among all 84 intra-host variants, only three of them were found to be polymorphic in the population data (position 7866 G/T; 27493 C/T; 28253 C/T). This small number of overlap also suggests that intra-host variants were rarely transmitted to other samples. However, we cannot rule out the possibility that the sequence diversity in the population is underestimated by the current database.

Missing microbiota signature associated with SARS-CoV-2 infection

Metatranscriptome data also enabled us to profile the transcriptionally active microbiota in different types of pneumonia, which is associated with the immunity response in the lung [12, 13]. In general, a significant difference in microbiota composition was observed among the nCoV, CAP, and Healthy groups ($R^2 = 0.07$, $p = 0.001$; Figure 3A). However, the clustering of some samples with NC indicated a barren microbiota in some samples. After removing the problematic samples and ambiguous components, we still found that nCoV and CAP were both different from the healthy controls (nCoV *vs.* Healthy: $R^2 = 0.45$, $p = 0.001$; CAP *vs.* Healthy: $R^2 = 0.10$, $p = 0.002$), implying a dysbiosis occurred in their lung microbiota. Microbiota could be classified into three different types (Figure 3B). In particular, the microbiota in cluster I was dominated by the possible pathogens, whereas the microorganisms in other clusters were more diverse. By further inspecting the species belonging to each cluster (Supplementary Table 4-5), we found that bacteria in Type III were mainly

commensal species frequently observed in the oral and respiratory tract, whereas bacteria in Type II were mostly environmental organisms, thereby contamination was highly suspected. Therefore, the microbiota was either pathogen-enriched (Type I) or commensal-enriched (Type III) or undetermined due to low microbial load (Type II).

The microbiota in six nCoV samples were pathogen-enriched, and the other two were commensal-enriched (Figure 3B). Moreover, two nCoV samples (2, 6) with an excess number of intra-host SARS-CoV-2 variants both possessed the pathogen-enriched microbiota. The overwhelming proportion of the virus may associate with a higher replication rate, and could also potentially stimulate the intense immune response against the virus, under which circumstance, an excess number of intra-host mutations would be expected. However, as only eight nCoV patients were included in this analysis, and the absolute microbial load was unknown, more data is needed for further investigation.

Discussion

RNA viruses have a high mutation rate due to the lack of proofreading activity of polymerases. Consequently, RNA viruses are prone to evolve resistance to drugs and escape from immune surveillance. The mutation rate of SARS-CoV-2 is still unclear. However, considering that the median number of pairwise sequence differences was 4 (Interquartile Range: 3-6) for 110 sequences collected between Dec 24, 2019 and Feb 9, 2020, the mutation rate should be at the same order of magnitude in SARS-CoV ($0.80\text{-}2.38\times10^{-3}$ nucleotide substitution per site per year)[14]. The high mutation rate also results in a high level of intra-host variants in RNA viruses [11, 15]. The median

number of intra-host variant in COVID-19 patients was 4 for variant with frequency ≥ 5%, and this incidence was not significantly different from that reported in a study on Ebola (655 variants with frequency ≥ 5% in 134 samples) ($p>0.05$)[11], suggesting that the mutation rate of SARS-CoV-2 was also comparable to Ebola virus. An exoribonuclease (ExoN) has been proposed to provide proofreading activity in SARS-CoV[16, 17], and we noted that all three key motifs in the gene were identical between SARS-CoV and SARS-CoV-2 (Supplementary Figure 3). In addition, neither polymorphism nor intra-host variant was detected in these motifs, suggesting that the gene is highly conserved, and thereby it could be a potential target for antiviral therapy. Although we did not find any mutation hotspot genes in either polymorphism or intra-host variants, the observation of shared intra-host variants among different individuals implied the possibility of adaptive evolution of the virus in patients, which could potentially affect the antigenicity, virulence, and infectivity of the virus [6].

It is worth noting that the SARS-CoV-2 genome in patients could be highly diverse, which was also observed in other viruses [11]. The high diversity could potentially increase the fitness of the viral population, making it hard to be eliminated[15]. Further studies are needed to explore how this may influence the immune response towards the virus and whether there is a selection acting on different strains in the human body or during the transmission. In a single transmission event investigated in this study, we found no evidence for the transmission of multiple strains. However, it is unclear whether these intra-host variants occurred before the transmission or after the transmission, which would result

in different conclusions. Additionally, a bottleneck may be involved in the transmission, which could also result in the loss of diversity [18]. Nevertheless, the observation of high mutation burden in some patients emphasized the possibility of rapid-evolving of this virus.

Recent studies have shown that the microbiota in the lung contributed to the immunological homeostasis and potentially altered the susceptibility to viral infection. Meanwhile, the lung microbiota could also be regulated by invading viruses [9, 19]. However, besides the feature that the microbial diversity was significantly lower in pneumonia than that in healthy controls (Figure 3B), we did not identify any specific microbiota pattern shared among COVID-19 patients, neither for CAP patients. A possible reason for this could be the use of antibiotics in pneumonia patients. However, this was not true for all pneumonia samples, as a substantial proportion of bacteria were observed in some samples, including two COVID-19 patients. It is well known that a common complication of viral infection, especially for respiratory viruses, secondary bacterial infection often results in a significant increase in morbidity [20]. Thus, the elevated level of bacteria in the BALF of some COVID-19 patients might increase the risk of secondary infection. In the clinical data, the secondary infection rate for COVID-19 was between 1%-10% [2, 21]. However, the quantitative relationship between bacterial relative abundance/titer and infection is unclear.

Overall, our study has revealed the evolution of SARS-CoV-2 in the patient, a common feature shared by most RNA viruses. How these variants influence the

fitness of viruses and genetic diversity in the population awaits further investigation. Currently, only limited sequences are shared in public databases (Supplementary Table 6); hence there is an urgent need to accumulate more sequences to trace the evolution of the viral genome and associate the changes with clinical symptoms and outcomes.

Methods.

Subjects and samples collection

Eight COVID-19 pneumonia samples were collected from hospitals in Wuhan from December 18 to 29, 2019; 25 virus-like community-acquired pneumonia (CAP) samples were collected from Beijing Peking University People's Hospital, The Shenzhen Third People's Hospital, Fujian Provincial Hospital, and The First-affiliated hospital of Xi'an Jiaotong University between 2014 and 2018. CAP was diagnosed following the guidelines of the Infectious Diseases Society of America and the American Thoracic Society [22]. Pneumonia patients with chronic pulmonary diseases were excluded. Meanwhile, BALF from 20 healthy volunteers were collected and used as healthy controls. Demographic information and clinical information were included in Supplementary Table 1.

For each patient, BALF samples were collected using a bronchoscope as part of normal clinical management. The volume of BALF samples ranged between 5ml and 30ml, most of which were used for bacterial culture and the remnant were aliquoted and stored at -80 ℃before processing.

Metatranscriptome sequencing

A 200 ul aliquot of each SARS-CoV-2 infected whole-BALF sample was used to extract RNA using Direct-zol RNA Miniprep kit (Zymo Research, Irvine, CA, USA) and Trizol LS (Thermo Fisher Scientific, Carlsbad, CA, USA) in biosafety III laboratory, and the rest samples were operated following the same protocol in biosafety II laboratory. The RNA was then reverse transcribed, and amplified using an Ovation Trio RNA-Seq library preparation kit (NuGEN, CA, USA) and was sequenced on an Illumina HiSeq 2500/4000 platform (Illumina, United Kingdom).

Data availability

The raw sequencing data reported in this paper have been deposited in the Genome Warehouse in National Genomics Data Center [23], under project number PRJCA002202 that is publicly accessible at https://bigd.big.ac.cn/gsa. Meanwhile, the data have also been submitted to NCBI Sequence Read Archive (SRA) database under project number PRJNA605907.

Data processing and taxonomic assignments

Quality control processes included adapter trimming, low quality reads removal, short reads removal by fastp (-l 70, -x, --cut-tail, --cut_tail_mean_quality 20, version: 0.20.0)[24], low complexity reads removal by Komplexity (-F, -k 8, -t 0.2, version: Nov 2019)[25], host removal by bmtagger (ftp://ftp.ncbi.nlm.nih.gov/pub/agarwala/bmtagger)[26, 27], and ribosomal reads removal by SortMeRNA (version:2.1b)[28].

The resultant reads were mapped against NCBI nt database (version: Jul 1 2019)

using BLAST+ (version:2.9.0)(-task megablast, -evalue 1e-10, -max_target_seqs 10, -max_hsps 1, -qcov_hsp_perc 60, -perc_identity 60)[29]. Taxonomic assignment was done by MEGAN using lowest common ancestor algorithm (-ms 100, -supp 0, -me 0.01, -top 10, -mrc 60, version: 6.11.0)[30]. After performing an overall PCoA and Permanova test, samples and microorganisms were filtered for further analyses with the following criteria. Samples with less than 5000 microbial reads were discarded. Microorganisms satisfying the following criteria were considered in the microbiota analysis, 1) archaea, bacteria, fungi, or virus; 2) with relative abundance ≥ 1% in the raw data and filtered data; 3) supported by at least 100 reads; 4) abundance higher than 10-fold of that in the negative control; 5) no batch effect; 6) abundance was not negatively correlated with bacteria titer; 7) not known contamination.

Intra-individual variants detection

Clean reads were mapped to the reference genome of SARS-CoV-2 (GenBank: MN908947.3) using BWA mem (version:0.7.12)[31]. Duplicate reads were removed by Picard (http://broadinstitute.github.io/picard; version: 2.18.22) [32] . Mpileup file was generated by samtools (version 1.8)[33], and intra-host variants were called using VarScan (version: 2.3.9)[34] and an in-house scripts. All variants had to satisfy the following requirements: 1) Sequencing depth ≥ 50; 2) Minor allele frequency ≥ 5%; 3) Minor allele frequency ≥ 2% on each strand; 4) Minor allele count ≥ 5 on each strand; 5) The minor allele was supported by the inner part of the read (excluding 10 bp on each end); 6) Both alleles could be identified in at least 3 reads that specifically assigned to genus *Betacoronavirus*.

For comparison with the polymorphism in the population, we obtained 110 sequences from GISAID (www.gisaid.org)[35, 36]. The accession number and acknowledgment were included in Supplementary Table S6.

Statistical analysis.

Pearson's chi-square test or Fisher's exact test was used for categorical variables, and the Mann-Whitney U test or Kruskal-Wallis rank sum test was used for continuous variables that do not follow a normal distribution. A comparison of microbiota was done by Permanova test.

Ethics statement.

The study was approved by the Institutional Review Board of Beijing Peking University People's Hospital, The Shenzhen Third People's Hospital, Fujian Provincial Hospital, and The First-affiliated hospital of Xi'an Jiaotong University. The data collection for the COVID-19 patients were deemed by the National Health Commission of the People's Republic of China as the contents of the public health outbreak investigation. Written informed consent was obtained from other pneumonia patients and healthy controls.

Acknowledgments

We thank Dr. Xue Yongbiao and colleagues from National Genomics Data Center for helpful discussion and computational resource support. We thank Dr. Huang Yanyi (Peking University, Beijing, China) and Wang Jianbin (Tsinghua University, Beijing, China) for providing the sequencing platform. We also thank Dr. Huang Chaolin for assist in sample collection. We gratefully acknowledge the Authors, the Originating and Submitting Laboratories for their sequence and metadata shared through GISAID, on which some of our analysis is based, a full name list of all submitters was given in Table S6.

Funding

This work was supported by grants from Innovation Fund for Medical Sciences [2016-I2M-1-014], the National Major Science & Technology Project for Control and Prevention of Major Infectious Diseases in China [2017ZX10103004, 2018ZX10305409, 2018ZX10301401, 2018ZX10732401]; National Natural Science Foundation of China [31670169, 31871263]; and the Open Project of Key Laboratory of Genomic and Precision Medicine, Chinese Academy of Sciences.

Potential conflicts of Interests

The authors declare no competing interests.

Reference

1. Zhu N, Zhang D, Wang W, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. N Engl J Med **2020**.
2. Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet **2020**.
3. Zhao S. LQ, Ran J., Musa S., Yang G., Wang W., Lou Y., Gao D., Yang L., He D., Wang M. Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. Int J Infect Dis **2020**.
4. Zhou P, Yang, X., Wang, X. et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature **2020**.
5. Lucas M, Karrer U, Lucas A, Klenerman P. Viral escape mechanisms--escapology taught by viruses. Int J Exp Pathol **2001**; 82(5): 269-86.
6. Berngruber TW, Froissart R, Choisy M, Gandon S. Evolution of virulence in emerging epidemics. PLoS Pathog **2013**; 9(3): e1003209.
7. O'Dwyer DN, Dickson RP, Moore BB. The Lung Microbiome, Immunity, and the Pathogenesis of Chronic Lung Disease. J Immunol **2016**; 196(12): 4839-47.
8. Hanada S, Pirzadeh M, Carver KY, Deng JC. Respiratory Viral Infection-Induced Microbiome Alterations and Secondary Bacterial Pneumonia. Front Immunol **2018**; 9: 2640.
9. Huffnagle GB, Dickson RP, Lukacs NW. The respiratory tract microbiome and lung inflammation: a two-way street. Mucosal Immunol **2017**; 10(2): 299-306.

10. Subissi L, Imbert I, Ferron F, et al. SARS-CoV ORF1b-encoded nonstructural proteins 12-16: replicative enzymes as antiviral targets. Antiviral Res **2014**; 101: 122-30.

11. Ni M, Chen C, Qian J, et al. Intra-host dynamics of Ebola virus during 2014. Nat Microbiol **2016**; 1(11): 16151.

12. Ren L, Zhang R, Rao J, et al. Transcriptionally Active Lung Microbiome and Its Association with Bacterial Biomass and Host Inflammatory Status. mSystems **2018**; 3(5).

13. Segal LN, Clemente JC, Tsay JC, et al. Enrichment of the lung microbiome with oral taxa is associated with lung inflammation of a Th17 phenotype. Nat Microbiol **2016**; 1: 16031.

14. Zhao Z, Li H, Wu X, et al. Moderate mutation rate in the SARS coronavirus genome and its implications. BMC Evol Biol **2004**; 4: 21.

15. Domingo E, Sheldon J, Perales C. Viral quasispecies evolution. Microbiol Mol Biol Rev **2012**; 76(2): 159-216.

16. Minskaia E, Hertzig T, Gorbalenya AE, et al. Discovery of an RNA virus 3'->5' exoribonuclease that is critically involved in coronavirus RNA synthesis. Proc Natl Acad Sci U S A **2006**; 103(13): 5108-13.

17. Smith EC, Blanc H, Surdel MC, Vignuzzi M, Denison MR. Coronaviruses lacking exoribonuclease activity are susceptible to lethal mutagenesis: evidence for proofreading and potential therapeutics. PLoS Pathog **2013**; 9(8): e1003565.

18. Sigal D, Reid JNS, Wahl LM. Effects of Transmission Bottlenecks on the Diversity of Influenza A Virus. Genetics **2018**; 210(3): 1075-88.

19. Tsang TK, Lee KH, Foxman B, et al. Association between the respiratory microbiome and susceptibility to influenza virus infection. Clin Infect Dis **2019**.

20. Hendaus MA, Jomha FA, Alhammadi AH. Virus-induced secondary bacterial infection: a concise review. Ther Clin Risk Manag **2015**; 11: 1265-71.

21. Chen N. ZM, Dong X., Qu J., Gong F., Han Y., Qiu Y., Wang J., Liu Y., Wei Y., Xia J., Yu T., Zhang X., Zhang L. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. The Lancet **2020**.

22. Mandell LA, Wunderink RG, Anzueto A, et al. Infectious Diseases Society of America/American Thoracic Society consensus guidelines on the management of community-acquired pneumonia in adults. Clin Infect Dis **2007**; 44 Suppl 2: S27-72.

23. National Genomics Data Center M, Partners. Database Resources of the National Genomics Data Center in 2020. Nucleic Acids Res **2020**; 48(D1): D24-D33.

24. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics **2018**; 34(17): i884-i90.

25. Clarke EL, Taylor LJ, Zhao C, et al. Sunbeam: an extensible pipeline for analyzing metagenomic sequencing experiments. Microbiome **2019**; 7(1): 46.

26. ftp://ftp.ncbi.nlm.nih.gov/pub/agarwala/bmtagger.

27. Wang J, Wang W, Li R, et al. The diploid genome sequence of an Asian individual. Nature **2008**; 456(7218): 60-5.

28. Kopylova E, Noe L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics **2012**; 28(24): 3211-7.

29. Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. BMC Bioinformatics **2009**; 10: 421.

30. Huson DH, Beier S, Flade I, et al. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLoS Comput Biol **2016**; 12(6): e1004957.

31. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:13033997 [q-bioGN] **2013**.

32. http://broadinstitute.github.io/picard.

33. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics **2009**; 25(16): 2078-9.

34. Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res **2012**; 22(3): 568-76.

35. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Glob Chall **2017**; 1(1): 33-46.

36. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro Surveill **2017**; 22(13).

37. Zhang Z, Li J, Zhao XQ, Wang J, Wong GK, Yu J. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genomics Proteomics Bioinformatics **2006**; 4(4): 259-63.

38. Salter SJ, Cox MJ, Turek EM, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol **2014**; 12: 87.

Table 1. The number of intra-host variants and polymorphisms in the genome of SARS-CoV-2

Gene	length	Intra-host variants			Polymorphisms			P-value[2]
		NS	S	Ka/Ks[1]	NS	S	Ka/Ks	
orf1a	13203	30	9	0.676	34	14	0.493*	0.627
orf1b	8088	8	5	0.355	10	8	0.277*	1
S	3822	8	3	0.599	10	6	0.375	0.692
ORF3a	828	2	1	0.561	4	2	0.561	1
E	228	0	0	NA	0	0	NA	1
M	669	2	0	NA	1	1	0.318	1
ORF6	186	1	0	NA	0	0	NA	1
ORF7a	366	2	0	NA	3	0	NA	1
ORF8	366	2	2	0.228	2	1	0.456	1
N	1260	7	2	1.048	5	7	0.214*	0.184
ORF10	117	0	0	NA	1	0	NA	1
Sum	29133	62	22	0.578*	70	39	0.368*	0.164

1. Ka/Ks was calculated using KaKs_Calculator2.0 (MS model)[37], an asterisk is added if Ka/Ks is significantly different from 1 ($p < 0.05$).
2. P-value indicating whether a significant difference of Ka/Ks ratio was observed between two types of mutations in the gene, Fisher Exact test.

Table 2. The allele frequency changes in transmission from nCoV4 to nCoV7

POS[1]	Ref_nCoV4[2]	Alt_nCoV4[3]	FRE	Ref_nCoV7	Alt_nCoV7	FRE	P-value[4]
376	119	177	0.598	9	0	0.000	0.0003
769	777	17	0.021	16	0	0.000	1
2037	1496	33	0.022	8	0	0.000	1
3290	2249	112	0.047	17	0	0.000	1
3306	1523	137	0.083	17	0	0.000	0.389
3321	1232	29	0.023	16	0	0.000	1
4511	685	26	0.037	11	0	0.000	1
4518	710	16	0.022	8	0	0.000	1
10771	1416	185	0.116	24	3	0.111	1
10773	1467	48	0.032	24	1	0.040	0.557
10779	987	401	0.289	18	8	0.308	0.829
10814	581	53	0.084	15	1	0.063	1
11387	653	15	0.022	4	0	0.000	1
13693	1237	38	0.030	9	0	0.000	1
15682	1321	46	0.034	8	1	0.111	0.269
15685	1342	47	0.034	8	1	0.111	0.270
18499	1783	108	0.057	19	0	0.000	0.621
18699	1013	46	0.043	12	0	0.000	1
21641	520	24	0.044	5	0	0.000	1
22270	2282	55	0.024	27	2	0.069	0.153
23127	1151	176	0.133	19	0	0.000	0.159
26177	492	11	0.022	6	0	0.000	1
27493	1535	1554	0.503	40	0	0.000	1.46E-12
28253	3600	487	0.119	34	0	0.000	0.028
29398	4671	127	0.026	47	0	0.000	0.636
Sum	34842	3968	0.102	421	17	0.028	1.45E-06

1. The four intra-host variant positions with MAF≥5% and passing selection criteria were highlighted in bold.
2. Number of reads supporting the reference allele.
3. Number of reads supporting the alternative (mutant) allele.
4. P-value indicates whether the difference of allele frequency between nCoV4 and nCoV7 is significant or not (Fisher Exact test).

Figure legends

Figure 1. Overview of the sequencing data. (A) The proportion of microbial reads in different groups; (B) Proportion of the viral read in patients infected with different viruses.

Figure 2. Intra-host variants in SARS-CoV-2 genome. (A) Genome coverage for SARS-CoV-2. A dash line indicates coverage of 50; (B) Frequency distribution of all intra-host variants, and the frequency of different mutations in polymorphism data was shown on the right side; (C) Distribution of the intra-host variations and polymorphisms on the genome of SARS-CoV-2. The outer ring displays the structure of the genome, following by the polymorphisms distribution on the genome. The length of each bar represents the number of sequences with this mutation. Due to a large variation of the number (1-27), 5% of the bar length was added for each additional sequence. The inner rings represent the distribution of intra-host variants in different patients (ID of each patient was labeled on each ring). Red bar indicates a synonymous mutation, and blue bar indicates a nonsynonymous mutation; (D) Frequency of the mutant allele at each high level (with frequency ≥ 20%) intra-host variant position. Nucleotides at the position (reference allele/alternative allele), mutation type (nonsynonymous, synonymous, noncoding), gene name, amino acid change were labeled on the right side of the heatmap. The total number of variants (with frequency ≥ 20%) in each sample was labeled on top of the heatmap. The name of five samples with more than 80% of the genome covered by at least 50-fold was labeled in blue. An open circle was added if the sample had a sequencing depth less than 50-fold at this position.

Figure 3. Microbiota in the BALF of COVID-19 patients, CAP patients, and healthy controls. A. Principal Coordinates Analysis (PCoA) of all samples. B. Heatmap of microbiota composition after QC filter (filters were described in Methods). The CAP samples were labeled as virus names followed by numbers. COVID-19 patients were highlighted by black rectangles, and two co-occurring bacterial clusters were highlighted by red rectangles. The names of all viruses are labeled in blue, and contaminant genera reported by Salter and colleagues are labeled in red[38].

Figure 1

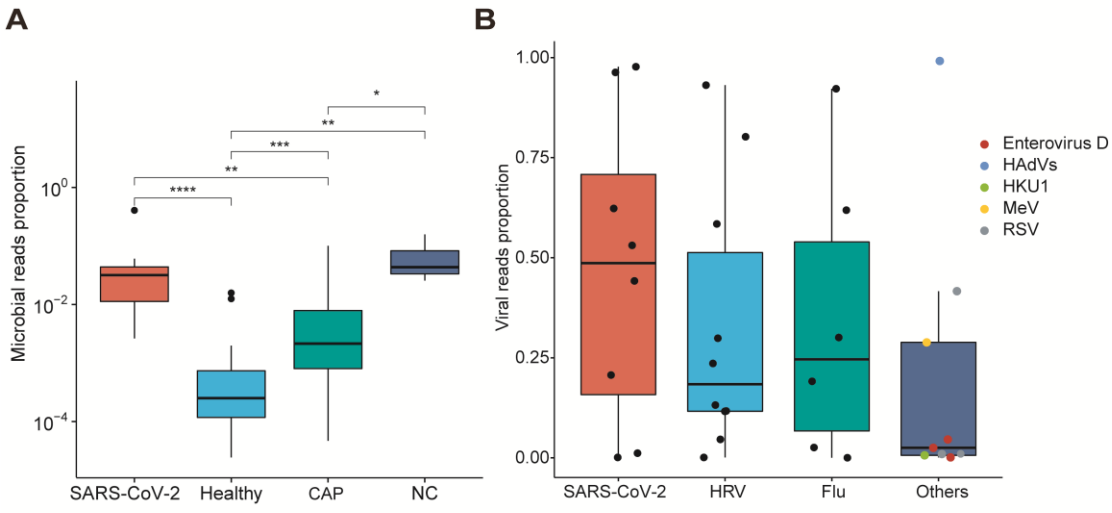

Figure 2

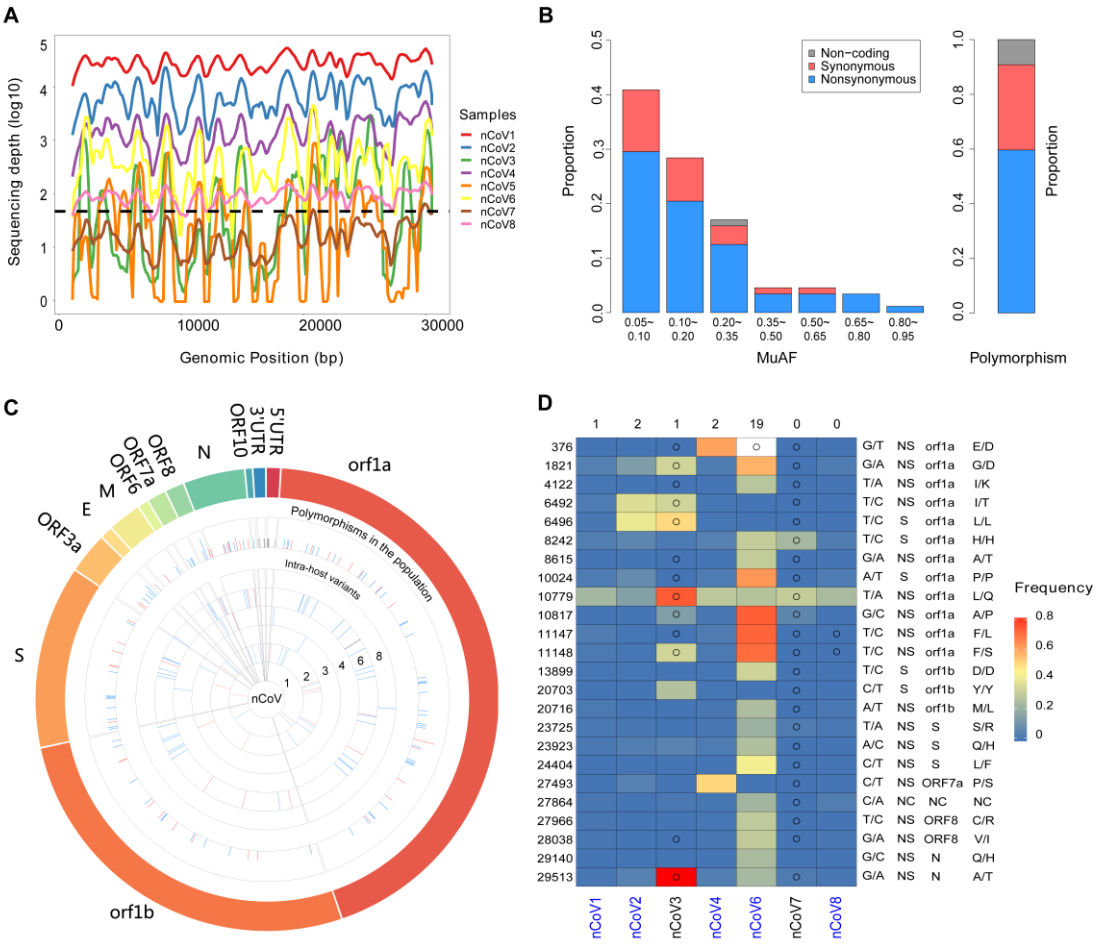

Figure 3