frontiers in Medicine

BRIEF RESEARCH REPORT
published: 15 February 2021
doi: 10.3389/fmed.2021.585358

Population Bottlenecks and Intra-host Evolution During Human-to-Human Transmission of SARS-CoV-2

Daxi Wang[1,2†], Yanqun Wang[3†], Wanying Sun[1,2,4†], Lu Zhang[3,5†], Jingkai Ji[1,2†], Zhaoyong Zhang[3†], Xinyi Cheng[1,2,6†], Yimin Li[3†], Fei Xiao[7], Airu Zhu[3], Bei Zhong[8], Shicong Ruan[9], Jiandong Li[1,2,4], Peidi Ren[1,2], Zhihua Ou[1,2], Minfeng Xiao[1,2], Min Li[1,2,4], Ziqing Deng[1,2], Huanzi Zhong[1,2,10], Fuqiang Li[1,2,11], Wen-jing Wang[1,11], Yongwei Zhang[1], Weijun Chen[4,12], Shida Zhu[1,13], Xun Xu[1,14], Xin Jin[1], Jingxian Zhao[3], Nanshan Zhong[3], Wenwei Zhang[1], Jincun Zhao[3,5]*, Junhua Li[1,2,6]* and Yonghao Xu[3]**

[1] BGI-Shenzhen, Shenzhen, China, [2] Shenzhen Key Laboratory of Unknown Pathogen Identification, BGI-Shenzhen, Shenzhen, China, [3] State Key Laboratory of Respiratory Disease, National Clinical Research Center for Respiratory Disease, Guangzhou Institute of Respiratory Health, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China, [4] BGI Education Center, University of Chinese Academy of Sciences, Shenzhen, China, [5] Institute of Infectious Disease, Guangzhou Eighth People's Hospital of Guangzhou Medical University, Guangzhou, China, [6] School of Biology and Biological Engineering, South China University of Technology, Guangzhou, China, [7] Guangdong Provincial Key Laboratory of Biomedical Imaging, Department of Infectious Diseases, Guangdong Provincial Engineering Research Center of Molecular Imaging, The Fifth Affiliated Hospital, Sun Yat-sen University, Zhuhai, China, [8] The Sixth Affiliated Hospital of Guangzhou Medical University, Qingyuan People's Hospital, Qingyuan, China, [9] Yangjiang People's Hospital, Yangjiang, China, [10] Laboratory of Genomics and Molecular Biomedicine, Department of Biology, University of Copenhagen, Copenhagen, Denmark, [11] Guangdong Provincial Key Laboratory of Human Disease Genomics, Shenzhen Key Laboratory of Genomics, BGI-Shenzhen, Shenzhen, China, [12] BGI PathoGenesis Pharmaceutical Technology, BGI-Shenzhen, Shenzhen, China, [13] Shenzhen Engineering Laboratory for Innovative Molecular Diagnostics, BGI-Shenzhen, Shenzhen, China, [14] Guangdong Provincial Key Laboratory of Genome Read and Write, BGI-Shenzhen, Shenzhen, China

The emergence of the novel human coronavirus, SARS-CoV-2, causes a global COVID-19 (coronavirus disease 2019) pandemic. Here, we have characterized and compared viral populations of SARS-CoV-2 among COVID-19 patients within and across households. Our work showed an active viral replication activity in the human respiratory tract and the co-existence of genetically distinct viruses within the same host. The inter-host comparison among viral populations further revealed a narrow transmission bottleneck between patients from the same households, suggesting a dominated role of stochastic dynamics in both inter-host and intra-host evolutions.

Keywords: SARS-CoV-2, population bottleneck, intra-host variation, human to human transmission, evolution

AUTHOR SUMMARY

In this study, we compared SARS-CoV-2 populations of 13 Chinese COVID-19 patients from three hospitals in different cities of Guangdong province. Those viral populations contained a considerable proportion of viral subgenomic messenger RNAs (sgmRNAs), reflecting an active viral replication activity in the respiratory tract tissues. The comparison of identified intra-host variants further showed a low viral genetic distance between intra-household patients and a narrow transmission bottleneck size. Despite the co-existence of genetically distinct viruses within the same host, most intra-host minor variants were not shared between transmission pairs, suggesting a

OPEN ACCESS

Edited by:
Michael Kemp, University of Southern Denmark, Denmark

Reviewed by:
Aine Niamh O'Toole, University of Edinburgh, United Kingdom
Roshan Kumar, Magadh University, India

***Correspondence:**
Wenwei Zhang
zhangww@genomics.cn
Jincun Zhao
zhaojincun@gird.cn
Junhua Li
lijunhua@genomics.cn
orcid.org/0000-0001-6784-1873
Yonghao Xu
dryonghao@163.com

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Infectious Diseases - Surveillance, Prevention and Treatment, a section of the journal Frontiers in Medicine

Received: *20 July 2020*
Accepted: *11 January 2021*
Published: *15 February 2021*

Citation:
Wang D, Wang Y, Sun W, Zhang L, Ji J, Zhang Z, Cheng X, Li Y, Xiao F, Zhu A, Zhong B, Ruan S, Li J, Ren P, Ou Z, Xiao M, Li M, Deng Z, Zhong H, Li F, Wang W-j, Zhang Y, Chen W, Zhu S, Xu X, Jin X, Zhao J, Zhong N, Zhang W, Zhao J, Li J and Xu Y (2021) Population Bottlenecks and Intra-host Evolution During Human-to-Human Transmission of SARS-CoV-2. Front. Med. 8:585358. doi: 10.3389/fmed.2021.585358

dominated role of stochastic dynamics in both inter-host and intra-host evolutions. Furthermore, the narrow bottleneck and active viral activity in the respiratory tract show that the passage of a small number of virions can cause infection. Our data have therefore delivered a key genomic resource for the SARS-CoV-2 transmission research and enhanced our understanding of the evolutionary dynamics of SARS-CoV-2.

INTRODUCTION

The rapid spread of the novel human coronavirus, SARS-CoV-2, has been causing millions of COVID-19 (coronavirus disease 2019) cases with high mortality rate worldwide (1, 2). As an RNA virus, SARS-CoV-2 mutates frequently (8.5×10^{-4} nucleotide substitutions per site per year) during genome replication (3–5), leading to the development of genetically different viruses within the same host. Several studies have reported intra-host single nucleotide variants (iSNVs) in SARS-CoV-2 (7, 8, 25). Recently, we investigated the intra-host evolution of SARS-CoV-2 and revealed genetic differentiation among tissue-specific populations (9). However, it is still not clear how the intra-host variants circulate among individuals. Here, we described and compared viral populations of SARS-CoV-2 among COVID-19 patients within and across households. Our work here demonstrated the utilization of viral genomic information to identify transmission linkage of this virus.

RESULTS AND DISCUSSION

Using both metatranscriptomic and hybrid-capture based techniques, we newly deep sequenced respiratory tract (RT) samples of seven COVID-19 patients in Guangdong, China, including two pairs of patients from the same households, respectively (P03 and P11; P23 and P24). Among the two intra-household pairs, patient P03, P23, and P24 had a travel history to Wuhan city during the early pandemic. The data were then combined with those of 23 RT samples used in our previous study (9), yielding a combined data set of 30 RT samples from 13 COVID-19 patients (**Supplementary Table 1**).

A sustained viral population should be supported by an active viral replication (10). We firstly estimated the viral transcription activity within RT samples using viral subgenomic messenger RNAs (**sgmRNAs**), which is only synthesized in infected host cells (11). The sgmRNA abundance was measured as the ratio of short reads spanning the transcription regulatory sequence (TRS) sites to the viral genomic reads (as demonstrated in **Supplementary Figure 1**). It should be noted that the sgmRNA abundance might be underestimated, given that only the short reads with sufficient length to simultaneously cover both leader and coding flanking regions of the TRS site, which might be improved with long read sequencing in future. Nonetheless, the sgmRNA abundance within nasal and throat swab samples was similar to that within sputum samples (**Figure 1A**), reflecting an active viral replication in the upper respiratory tract. Notably, the patient P01, who eventually passed away due to COVID-19, showed the highest level of sgmRNA abundance (**Supplementary Figure 2**). However, due to the limited samples of mild cases, we did not observe a significant difference of sgmRNA abundance between severe and mild cases. For the patients with chronological samples and improved clinical outcomes (P10 and P13), their viral load measured by real-time reverse transcription PCR (qRT-PCR) negatively correlate with the days post symptoms onset with marginal significance (**Figures 1B,C**). Interestingly, the sgmRNA abundance showed a similar trend across time (**Figures 1D,E**), reflecting a direct biological association between viral replication and viral shedding in the respiratory tract tissues.

Using the metatranscriptomic data, we identified 66 iSNVs in protein encoding regions with the alternative allele frequency (AAF) ranged from 5 to 95% (**Supplementary Tables 2, 3**; **Supplementary Figure 3**). Here, an alternative allele was defined as the allele that is different from the allele at the same position of the reference genome. The identified iSNVs showed a high concordance between the AAFs derived from metatranscriptomic and that from hybrid-capture sequences (Spearman's $\rho = 0.81$, $P < 2.2e\text{-}16$; **Supplementary Figure 4**). We firstly looked for signals of natural selection against intra-host variants. Using the Fisher's exact test, we compared the number of iSNV sites on each codon position against that of the other two positions and detected a marginal but significant difference among them (codon position 1 [$n = 10$, $P = 0.02$], 2 [$n = 21$; $P = 1$], and 3 [$n = 35$; $P = 0.03$]). In contrast to the numbers of iSNV sites, the alternative allele frequency of those iSNVs did not discriminate among the non-synonymous and synonymous categories (**Figure 2A**), suggesting that most non-synonymous intra-host variants were not under an effective purifying selection within the host. Among the 66 identified iSNVs, 30 were coincided with the consensus variants in the public database as of April 5, 2020 (**Supplementary Table 2**). Those iSNVs were categorized into common iSNVs, while the iSNVs presented in a single patient were categorized into rare iSNVs. Interestingly, the common iSNVs had a significant higher minor allele frequency compared to the rare iSNVs (**Supplementary Figure 5**; Wilcoxon rank sum test, $P = 2.7e\text{-}05$), suggesting that they may have been developed in earlier strains before the most recent infection.

We then estimated the viral genetic distance among samples in a pairwise manner based on their iSNVs and allele frequencies. The samples were firstly categorized into intra-host pairs (serial samples from the same host), intra-household pairs and inter-household pairs (**Figure 2B** and **Supplementary Table 4**). As expected, the intra-host pairs had the lowest genetic distance compared to either intra-household pairs (Wilcoxon rank sum test, $P = 0.018$) and inter-household pairs (Wilcoxon rank sum test, $P < 2.22e\text{-}16$). Interestingly, the genetic distance between intra-household pairs was significantly lower than that of inter-household pairs (**Figure 2B**; Wilcoxon rank sum test, $P = 0.03$), supporting a direct passage of virions among intra-household individuals. Nonetheless, we only observed a small proportion of (3/14 for P03 and P11; 1/20 for P23 and P24) minor intra-host variants shared among intra-household pairs, suggesting that the estimated genetic similarity was mostly determined by consensus nucleotide differences (**Figures 2C,D**). Based on the AAF of iSNVs in transmission pairs, it seems only the minor virion groups carrying three (from P03) and one variants (from

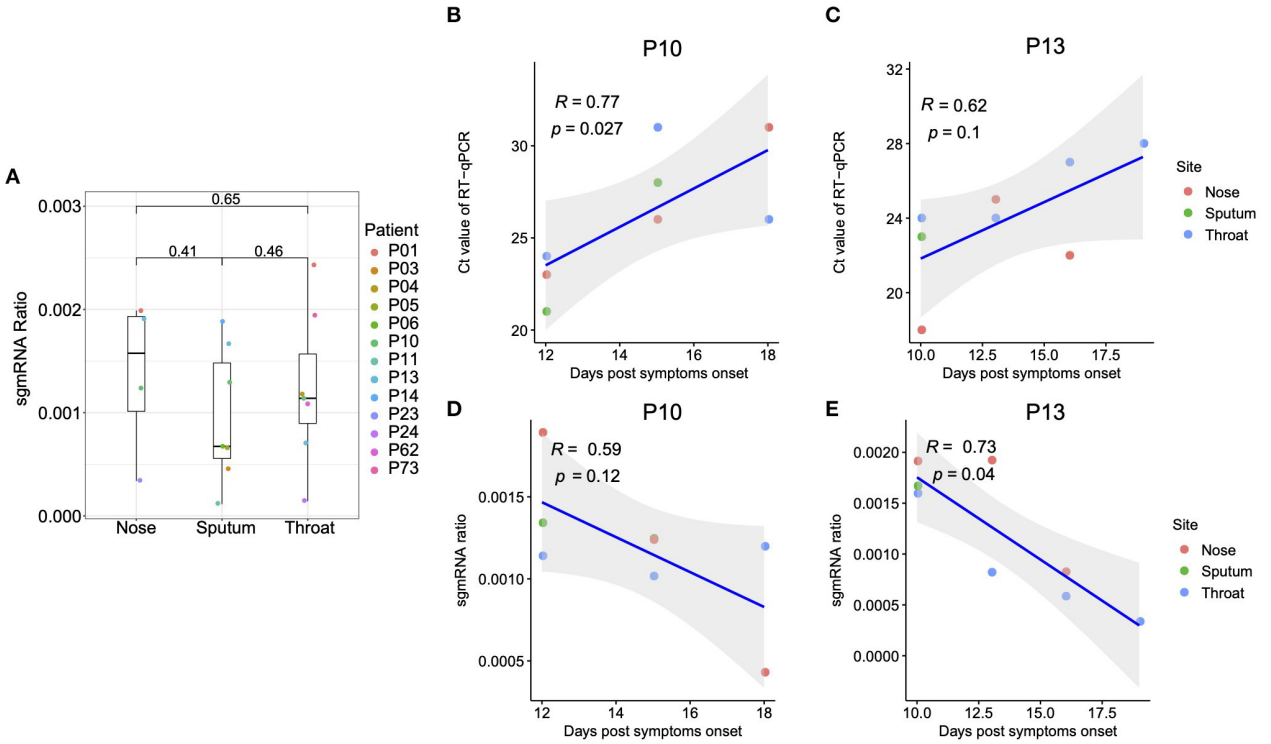

FIGURE 1 | Subgenomic messenger RNAs (sgmRNAs). **(A)** The ratio of sgmRNA of each respiratory sample type (nasal, throat swabs, and sputum). For the patients with multiple samples from the same sample type, only the samples with the median sgmRNA ratio were displayed. **(B,C)** Correlation between the cycle threshold and the days post symptoms onset for patient P10 and P13, respectively. **(D,E)** Correlation between the sgmRNA ratio and the days post symptoms onset for patient P10 and P13, respectively.

P23) were passed to the recipient, respectively. Specifically, in one intra-household pair (P23 and P24), one patient (P23) contained iSNVs that were coincided with the linked variants, C8782T and T28144C, suggesting that this patient may have been co-infected by genetically distinct viruses. However, the strain carrying 8782C and 28144T was not observed in the intra-household counterpart (P24). Given the small number of COVID-19 cases in Guangdong (about 2,000 total cases), secondary infection from other sources is not likely. Within this intra-household pair, it is likely that there is a narrow transmission bottleneck allowing only the major strain to be circulated, if P23 was infected by all the observed viral strains before the transmission.

The transmission bottlenecks among intra-household pairs were estimated using a beta binomial model, which was designed to allow some temporal stochastic dynamics of viral population in the recipient (12). Here, we defined the donor and recipient within the intra-household pairs according to their dates of the first symptom onset. The estimated bottleneck sizes were 6 (P03 and P11) and 8 (P23 and P24) for the two intra-household pairs (**Supplementary Table 5**). The observed narrow bottleneck is consistent with two recent studies of SARS-CoV-2 (13, 14). Nonetheless, a loose transmission bottleneck was also observed (8). Similarly, many animal viruses and human respiratory viruses showed a narrow transmission bottleneck (15, 16), while the only study reporting a loose bottleneck among human respiratory viral infections (17) was argued as the generic consequence of shared iSNVs caused by read mapping artifacts (18). The relatively narrow transmission bottleneck sizes are expected to increase the variance of viral variants being circulated between transmission pairs (19). However, given that we can only measure the viral population that were descendants of the founding population, the actual population could have been much larger. Even after successful transmission, virions carrying the minor variants are likely to be purged out due to the frequent stochastic dynamics within the respiratory tract (9), which is also consistent with the low diversity and instable iSNV observed among the RT samples.

The observed narrow transmission bottleneck suggests that, in general, only a few virions successfully enter host cells and eventually cause infection. Although the number of transmitted virions is sparse, they can easily replicate in the respiratory tract, given the observed viral replication activities in all the RT sample types and the high host-cell receptor binding affinity of SARS-CoV-2 (6). The narrow transmission bottleneck also indicate that instant hand hygiene and mask-wearing might be particular effective in blocking the transmission chain of SARS-CoV-2.

In summary, we have characterized and compared SARS-CoV-2 populations of patients within and across households using both metatranscriptomic and hybrid-capture based techniques. Our work showed an active viral replication activity

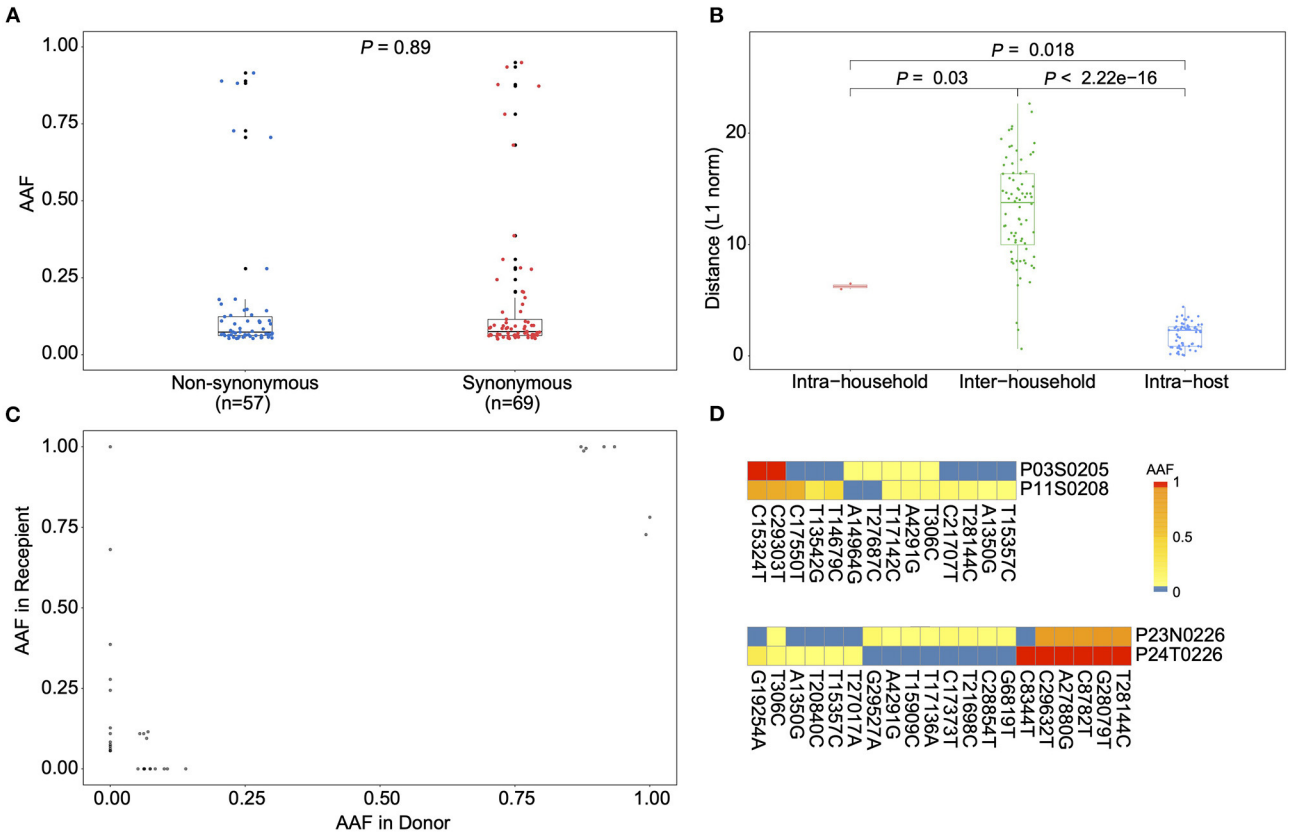

FIGURE 2 | Allele frequency changes of transmission pairs. **(A)** Box plots showing the alternative allele frequency (AAF) distribution of synonymous and non-synonymous intra-host variants. Each dot indicates the median AAF of each iSNV sites of samples from same patient. **(B)** Box plots representing the L1-norm distance distribution among sample pairs. Each dot represents the genetic distance between each sample pair. **(C)** The AAF of donor iSNVs in transmission pairs. Allele frequencies under 5% and over 95% were adjusted to 0% and 1, respectively. **(D)** Heatmap representing the alternative allele frequencies (AAFs) of consensus and intra-host single nucleotide variants (iSNVs) of the two transmission pairs.

in the human respiratory tract and the co-existence of genetically distinct viruses within the same host. The inter-host comparison among viral populations further revealed a narrow transmission bottleneck between patients from the same households, suggesting a dominated role of stochastic dynamics in both inter-host and intra-host evolution. The present work enhanced our understanding of SARS-CoV-2 virus transmission and shed light on the integration of genomic and epidemiological in the control of this virus.

MATERIALS AND METHODS

Patients

Respiratory tract (RT) samples, including nasal swabs, throat swabs, sputum, were collected from 13 COVID-19 patients during the early outbreak of the pandemic (from January 25 to February 10 of 2020). Those patients were hospitalized at the first affiliated hospital of Guangzhou Medical University (nine patients) in Guangzhou, the fifth affiliated hospital of Sun Yat-sen University (two patients), Qingyuan People's Hospital (1 patient) in Zhuhai and Yangjiang People's Hospital (one patient) Yangjiang. The research plan was assessed and approved by the Ethics Committee of each hospital. All the privacy information was anonymized.

Dataset Description

Public consensus sequences were downloaded from GISAID on April 5, 2020.

Sample Preparation and Sequencing

RNA was extracted from the clinical RT samples using QIAamp Viral RNA Mini Kit (Qiagen, Hilden, Germany), which was then tested for SARS-CoV-2 using qRT-PCR. Human DNA was removed using DNase I and RNA concentration was measured using Qubit RNA HS Assay Kit (Thermo Fisher Scientific, Waltham, MA, USA). After human DNA-depletion, the samples were RNA purified and then subjected to double-stranded DNA library construction using the MGIEasy RNA Library preparation reagent set (MGI, Shenzhen, China) following the method used in the previous study (20). Possible environmental or cross contamination during library preparation was tracked using the control RNA samples from human breast cell lines

(Michigan Cancer Foundation-7). The constructed libraries were converted to DNA nanoballs (DNBs) and then sequenced on the DNBSEQ-T7 platform (MGI, Shenzhen, China), generating paired-end short reads with 100 bp in length. Most samples were also sequenced using hybrid capture-based enrichment approach that was described in previous study (20). Briefly, the SARS-CoV-2 genomic content was enriched from the double-stranded DNA libraries using the 2019-nCoVirus DNA/RNA Capture Panel (BOKE, Jiangsu, China). The enriched SARS-CoV-2 genomic contents were converted to DNBs and then sequenced on the MGISEQ-2000 platform, generating paired-end short reads with 100 bp in length.

Data Filtering

Read data from both metatranscriptomic and hybrid capture based sequencing were filtered following the steps described in the previous research (20). Short read data were mapped to a database that contains all the available reference genomes of coronaviridae (including SARS, SARS-CoV-2 and MERS genomes from GISAID, NCBI and CNGB) using Kraken v0.10.5 with default parameters. Low-quality, adaptor contaminations, duplications within the mapped reads were removed using fastp v0.19.5 and SOAPnuke v1.5.6. Low-complexity reads were then filtered using PRINSEQ v0.20.4.

Profiling of Subgenomic Messenger RNAs (SgmRNAs)

Coronaviridae-like short reads were mapped to the reference genome (EPI_ISL_402119) using the aligner HISAT2 (21). Reads spanning the transcription regulatory sequence (TRS) sites of both leader region and the coding genes (S gene, ORF3a, 6, 7a, 8, E, M, and N gene) were selected to represent the sgmRNAs. The junction sites were predicted using RegTools junctions extract (22). The ratio of sgmRNA reads to the viral genomic RNA reads (sgmRNA ratio) was used to estimate the relative transcription activity of SARS-CoV-2. The sgmRNA ratio and its correlation with days post the first symptom were plotted using the R package *ggplot* (v.3.3.0). To avoid oversampling, for the patients with more than one sample, only the median sgmRNA ratio from samples of that patient was used for comparison among patients.

Detection of Intra-Host Variants

We defined an intra-host single nucleotide variant (iSNV) as the co-existence of an alternative allele and the reference allele at the same genomic position within the same sample. To identify iSNV sites, paired-end metatranscriptomic coronaviridae-like short read data were mapped to the reference genome (EPI_ISL_402119) using BWA aln (v.0.7.16) with default parameters (23). The duplicated reads were detected and marked using Picard MarkDuplicates (v. 2.10.10) (http://broadinstitute.github.io/picard). Nucleotide composition of each genomic position was characterized from the read mapping results using pysamstats (v. 1.1.2) (https://github.com/alimanfoo/pysamstats). The variable sites of each sample were identified using the variant caller LoFreq with default filters and the cut-off of 5% minor allele frequency ($n = 89$). After removing variable sites at UTR regions ($n = 12$), the sites with more than one alternative allele ($n = 0$), and the sites with only fixed variants (AAF > 95%) were filtered out ($n = 9$). All the iSNVs with less than five metatranscriptomic reads were verified using the hybrid capture data (at least two reads), and thus removed two iSNV sites. The rest 66 sites were regarded as iSNV sites. The identified iSNVs were then annotated using the SnpEff (v.2.0.5) with default settings (24). Alternative allele frequencies between synonymous and non-synonymous iSNV sites were tested with Wilcoxon rank sum test. Each dot indicates the median AAF of the same iSNV site of samples from same patient. All the plots were visualized using the R package ggplot (v.3.3.0).

Genetic Distance

The genetic distance between sample pairs was calculated using L1-norm distance, as defined by the following formula. To avoid oversampling, for the patients with more than one sample, only the median AAF among all samples of that patient was used for distance comparison. The L1-norm distance (D) between sample pairs is calculated by summing the distance of all the variable loci (N). The distance on each variable locus is calculated between vectors (p and q for each sample) of possible base frequencies ($n = 4$).

$$D = \sum_{k=1}^{N} \sum_{i=1}^{n} |p_i - q_i|$$

To verify the result, L2-norm distance (Euclidean distance) between sample pairs was calculated. The L2-norm distance $d(p, q)$ between two samples (p, q) is the square root of sum of distance across all the variable loci (N), as defined by the following formula.

$$d(p, q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

The comparison of genetic distances among sample pair categories was performed using the Wilcoxon rank-sum test.

Beta Binomial Model of Bottleneck Size Estimation

A beta-binomial model was used to estimate bottleneck sizes between donor and recipient (12) (https://github.com/weissmanlab/BB_bottleneck). The beta-binomial model can estimate the probability of variant being detected in the recipient viral population under the prior condition of founding population, allowing variant frequency changes between founding time and sampling time. Here, the bottleneck size represents the number of virions that pass into the recipient and finally shape the sequenced viral population. The patient with the earlier symptom onset date was defined as the donor, while the other was defined as the recipient. For each variable site, variant frequencies within both donor and recipient, read depth and number of reads supporting the mutation within the recipient were used as input of the beta-binomial model. In this model, the virus transmission from donor to the recipient was regarded as a Bernoulli trial, and the probability of a given number of

virions carrying this mutation follows a binomial distribution. The maximum-likelihood estimates (MLE) of bottleneck sizes were estimated within 95% confidence intervals. In our data, we got 6 and 8 virions as the estimated transmission bottleneck size of the two donor-recipient pairs, as the probabilities of their beta-binomial distributions reached maximums, respectively.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study have been deposited into CNSA (CNGB Sequence Archive) of CNGBdb with the accession number CNP0001111 (https://db.cngb.org/cnsa/).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of the first affiliated hospital of Guangzhou Medical University, the fifth affiliated hospital of Sun Yat-sen University, Qingyuan People's Hospital and Yangjiang People's Hospital, respectively. Informed consent was obtained from all participants enrolled in the study. All the privacy information was anonymized.

AUTHOR CONTRIBUTIONS

DW, YX, JL, WZ, and JZ conceived the study. YW, LZ, and YL collected clinical specimen and executed the experiments. DW, WS, XC, and JJ analyzed the data. DW, YW, and ZZ wrote the manuscript. All the authors participated in discussion and result interpretation and revised and approved the final version.

FUNDING

This study was funded by grants from The National Key Research and Development Program of China (2018YFC1200100, 2018ZX10301403, 2018YFC1311900), the emergency grants for prevention and control of SARS-CoV-2 of Ministry of Science and Technology (2020YFC0841400), Guangdong province (2020B111107001, 2020B111108001, 2020B111109001, 2018B020207013, 2020B111112003), the Guangdong Province Basic and Applied Basic Research Fund (2020A1515010911), Guangdong Science and Technology Foundation (2019B030316028), Shenzhen Peacock Team Plan grant (No. KQTD2015033117210153), Guangdong Provincial Key Laboratory of Genome Read and Write (2017B030301011), and Guangzhou Medical University High-level University Innovation Team Training Program (Guangzhou Medical University released [2017] No.159). This work was supported by the Shenzhen Municipal Government of China Peacock Plan (KQTD2015033017150531). This work was supported by China National GeneBank (CNGB).

ACKNOWLEDGMENTS

This study was approved by the Health Commission of Guangdong Province to use patients' specimen for this study. We thank the patients who took part in this study. This manuscript has been released as a pre-print at the bioRxiv platform (Wang et al.).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2021.585358/full#supplementary-material

Supplementary Figure 1 | Structure of subgenomic messenger RNAs (sgmRNAs).

Supplementary Figure 2 | Transcription profile of subgenomic messenger RNAs (sgmRNAs) of each patient.

Supplementary Figure 3 | Heatmap representing the alternative allele frequencies (AAFs) of consensus and intra-host single nucleotide variants (iSNVs) of 30 patients.

Supplementary Figure 4 | Concordance between minor alternative allele frequencies (AAFs) derived from metagenomic and hybrid capture data.

Supplementary Figure 5 | Alternative allele frequency (AAF) distribution of rare and common iSNVs. Each dot indicates the median AAF of each iSNV sites of samples from same patient.

Supplementary Table 1 | Demography and clinical outcomes of COVID-19 patients.

Supplementary Table 2 | Summary of iSNVs.

Supplementary Table 3 | Frequency of iSNVs.

Supplementary Table 4 | Inter-host genetic distance (L1 and L2-norm).

Supplementary Table 5 | Bottleneck size of intra-household pairs.

REFERENCES

1. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. *Nature*. (2020) 579:265–9. doi: 10.1038/s41586-020-2008-3
2. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. (2020) 579:270–3. doi: 10.1038/s41586-020-2012-7
3. Smith EC, Sexton NR, Denison MR. Thinking outside the triangle: replication fidelity of the largest RNA viruses. *Annu Rev Virol.* (2014) 1:111–32. doi: 10.1146/annurev-virology-031413-085507
4. Day T, Gandon S, Lion S, Otto SP. On the evolutionary epidemiology of SARS-CoV-2. *Curr Biol.* (2020) 30:R849–57. doi: 10.1016/j.cub.2020.06.031
5. Nakagawa S, Miyazawa T. Genome evolution of SARS-CoV-2 and its virological characteristics. *Inflamm Regen.* (2020) 40:17. doi: 10.1186/s41232-020-00126-7
6. Shang J, Ye G, Shi K, Wan Y, Luo C, Aihara H, et al. Structural basis of receptor recognition by SARS-CoV-2. *Nature*. (2020) 581:221–4. doi: 10.1038/s41586-020-2179-y
7. Butler DJ, Mozsary C, Meydan C, Danko D, Foox J, Rosiene J, et al. Shotgun transcriptome and isothermal profiling of SARS-CoV-2 infection reveals unique host responses, viral diversification, and drug interactions. *bioRxiv [Preprint]*. (2020). doi: 10.1101/2020.04.20.048066
8. Lythgoe KA, Hall M, Ferretti L, de Cesare M, MacIntyre-Cockett G, Trebes A, et al. Shared SARS-CoV-2 diversity suggests localised transmission of minority variants. *bioRxiv [Preprint]*. (2020). doi: 10.1101/2020.05.28.118992

9. Wang Y, Wang D, Zhang L, Sun W, Zhang Z, Chen W, et al. Intra-host variation and evolutionary dynamics of SARS-CoV-2 population in COVID-19 patients. *bioRxiv [Preprint]*. (2020). doi: 10.1101/2020.05.20.103549
10. Wölfel R, Corman VM, Guggemos W, Seilmaier M, Zange S, Müller MA, et al. Virological assessment of hospitalized patients with COVID-2019. *Nature*. (2020) 581:465–9. doi: 10.1038/s41586-020-2196-x
11. Sola I, Almazán F, Zúñiga S, Enjuanes L. Continuous and discontinuous RNA synthesis in coronaviruses. *Annu Rev Virol.* (2015) 2:265–88. doi: 10.1146/annurev-virology-100114-055218
12. Sobel Leonard A, Weissman DB, Greenbaum B, Ghedin E, Koelle K. Transmission bottleneck size estimation from pathogen deep-sequencing data, with an application to human influenza A virus. *J Virol.* (2017) 91:0–19. doi: 10.1128/JVI.00171-17
13. Braun KM, Moreno GK, Halfmann PJ, Baker DA, Weiler AM, Haj AK, et al. Transmission of SARS-CoV-2 in domestic cats imposes a narrow bottleneck. *bioRxiv [Preprint]*. (2020). doi: 10.1101/2020.11.16.384917
14. James SE, Ngcapu S, Kanzi AM, Tegally H, Fonseca V, Giandhari J, et al. High resolution analysis of transmission dynamics of SARS-CoV-2 in two major hospital outbreaks in South Africa leveraging intrahost diversity. *medRxiv [Preprint]*. (2020). doi: 10.1101/2020.11.15.20231993
15. Zwart MP, Elena SF. Matters of size: genetic bottlenecks in virus infection and their potential impact on evolution. *Annu Rev Virol.* (2015) 2:161–79. doi: 10.1146/annurev-virology-100114-055135
16. McCrone JT, Woods RJ, Martin ET, Malosh RE, Monto AS, Lauring AS. Stochastic processes constrain the within and between host evolution of influenza virus. *Elife.* (2018) 7:e35962. doi: 10.7554/eLife.35962.036
17. Poon LLM, Song T, Rosenfeld R, Lin X, Rogers MB, Zhou B, et al. Quantifying influenza virus diversity and transmission in humans. *Nat Genet.* (2016) 48:195–200. doi: 10.1038/ng.3479
18. Xue KS, Bloom JD. Reconciling disparate estimates of viral genetic diversity during human influenza infections. *Nat Genet.* (2019) 51:1298–301. doi: 10.1038/s41588-019-0349-3
19. Xue KS, Moncla LH, Bedford T, Bloom JD. Within-host evolution of human influenza virus. *Trends Microbiol.* (2018) 26:781–93. doi: 10.1016/j.tim.2018.02.007
20. Xiao M, Liu X, Ji J, Li M, Li J, Sun W, et al. Multiple approaches for massively parallel sequencing of HCoV-19 genomes directly from clinical samples. *Genome Med.* (2020) 12:57. doi: 10.1186/s13073-020-00751-4
21. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods*. (2015) 12:357–60. doi: 10.1038/nmeth.3317
22. Cotto KC, Feng Y-Y, Ramu A, Skidmore ZL, Kunisaki J, Conrad DF, et al. RegTools: Integrated analysis of genomic and transcriptomic data for the discovery of splicing variants in cancer. *bioRxiv [Preprint]*. (2020). doi: 10.1101/43663423
23. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*. (2010) 26:589–95. doi: 10.1093/bioinformatics/btp698
24. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*. (2012) 6:80–92. doi: 10.4161/fly.19695
25. Shen Z, Xiao Y, Kang L, Ma W, Shi L, Zhang L, et al. Genomic diversity of severe acute respiratory syndrome-coronavirus 2 in patients with coronavirus disease 2019. *Clin Infect Dis*. (2020) 71:713–20. doi: 10.1093/cid/ciaa203

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.