1Population Bottlenecks and Intra-host Evolution during Human-to- 2Human Transmission of SARS-CoV-2

3

4Daxi Wang^{1,3#}, Yanqun Wang^{2#}, Wanying Sun^{1,3,4#}, Lu Zhang^{2,5#}, Jingkai Ji^{1,3#}, Zhaoyong
5Zhang^{2#}, Xinyi Cheng^{1,3,13#}, Yimin Li^{2#}, Fei Xiao⁶, Airu Zhu², Bei Zhong⁷, Shicong Ruan⁸,
6Jiandong Li^{1,3,4}, Peidi Ren^{1,3}, Zhihua Ou^{1,3}, Minfeng Xiao^{1,3}, Min Li^{1,3,4}, Ziqing Deng^{1,3}, Huanzi
7Zhong^{1,3,9}, Fuqiang Li^{1,3,10}, Wen-jing Wang^{1,10}, Yongwei Zhang¹, Weijun Chen^{4,11}, Shida
8Zhu^{1,12}, Xun Xu^{1,14}, Xin Jin¹, Jingxian Zhao², Nanshan Zhong², Wenwei Zhang^{1*}, Jincun
9Zhao^{2,5*}, Junhua Li^{1,3,13*}, Yonghao Xu^{2*}

10

11¹BGI-Shenzhen, Shenzhen, 518083, China.

12²State Key Laboratory of Respiratory Disease, National Clinical Research Center for Respiratory
13Disease, Guangzhou Institute of Respiratory Health, the First Affiliated Hospital of Guangzhou
14Medical University, Guangzhou, Guangdong, 510120, China.

15³Shenzhen Key Laboratory of Unknown Pathogen Identification, BGI-Shenzhen, Shenzhen, 518083,
16China.

17⁴BGI Education Center, University of Chinese Academy of Sciences, Shenzhen, 518083, China.

18⁵Institute of Infectious disease, Guangzhou Eighth People's Hospital of Guangzhou Medical
19University, Guangzhou, Guangdong, 510060, China.

20⁶Department of Infectious Diseases, Guangdong Provincial Key Laboratory of Biomedical Imaging,
21Guangdong Provincial Engineering Research Center of Molecular Imaging, The Fifth Affiliated
22Hospital, Sun Yat-sen University, Zhuhai, Guangdong, 519000, China.

23⁷The Sixth Affiliated Hospital of Guangzhou Medical University, Qingyuan People's Hospital,
24Qingyuan, Guangdong, China.

25⁸Yangjiang People's Hospital, Yangjiang, Guangdong, China

26⁹Laboratory of Genomics and Molecular Biomedicine, Department of Biology, University of
27Copenhagen, 2100 Copenhagen, Denmark.

28¹⁰Guangdong Provincial Key Laboratory of Human Disease Genomics, Shenzhen Key Laboratory of
29Genomics, BGI-Shenzhen, Shenzhen, 518083, China.

30¹¹BGI PathoGenesis Pharmaceutical Technology, BGI-Shenzhen, Shenzhen, 518083, China.

31¹²Shenzhen Engineering Laboratory for Innovative Molecular Diagnostics, BGI-Shenzhen, Shenzhen, 32518120, China.

33¹³School of Biology and Biological Engineering, South China University of Technology, Guangzhou, 34China.

35¹⁴Guangdong Provincial Key Laboratory of Genome Read and Write, BGI-Shenzhen, Shenzhen, 36518120, China.

37[#]These authors contributed equally to this work.

38^{*}These authors are corresponding authors.

39Email address: zhangww@genomics.cn

40Email address: zhaojincun@gird.cn

41Email address: lijunhua@genomics.cn

42Email address: dryonghao@163.com

43

44Abstract

45The emergence of the novel human coronavirus, SARS-CoV-2, causes a global COVID-19
46(coronavirus disease 2019) pandemic. Here, we have characterized and compared viral populations
47of SARS-CoV-2 among COVID-19 patients within and across households. Our work showed an active
48viral replication activity in the human respiratory tract and the co-existence of genetically distinct
49viruses within the same host. The inter-host comparison among viral populations further revealed a
50narrow transmission bottleneck between patients from the same households, suggesting a dominated
51role of stochastic dynamics in both inter-host and intra-host evolutions.

52

53Author summary

54In this study, we compared SARS-CoV-2 populations of 13 Chinese COVID-19 patients. Those viral
55populations contained a considerable proportion of viral sub-genomic messenger RNAs (sgmRNA),
56reflecting an active viral replication activity in the respiratory tract tissues. The comparison of 66
57identified intra-host variants further showed a low viral genetic distance between intra-household
58patients and a narrow transmission bottleneck size. Despite the co-existence of genetically distinct
59viruses within the same host, most intra-host minor variants were not shared between transmission
60pairs, suggesting a dominated role of stochastic dynamics in both inter-host and intra-host evolutions.
61Furthermore, the narrow bottleneck and active viral activity in the respiratory tract show that the
62passage of a small number of virions can cause infection. Our data have therefore delivered a key
63genomic resource for the SARS-CoV-2 transmission research and enhanced our understanding of the
64evolutionary dynamics of SARS-CoV-2.

65

66Introduction

67The rapid spread of the novel human coronavirus, SARS-CoV-2, has been causing millions of COVID-
6819 (coronavirus disease 2019) cases with high mortality rate worldwide [1,2]. As an RNA virus, SARS-
69CoV-2 mutates frequently due to the lack of sufficient mismatch repairing mechanisms during genome
70replication [3], leading to the development of genetically different viruses within the same host.
71Several studies have reported intra-host single nucleotide variants (iSNVs) in SARS-CoV-2 [4–6].
72Recently, we investigated the intra-host evolution of SARS-CoV-2 and revealed genetic differentiation
73among tissue-specific populations [7]. However, it is still not clear how the intra-host variants circulate

74among individuals. Here, we described and compared viral populations of SARS-CoV-2 among
75COVID-19 patients within and across households. Our work here demonstrated the utilization of viral
76genomic information to identify transmission linkage of this virus.

77

78Results and discussion

79Using both metatranscriptomic and hybrid-capture based techniques, we newly deep sequenced
80respiratory tract (RT) samples of seven COVID-19 patients in Guangdong, China, including two pairs
81of patients from the same households, respectively (P03 and P11; P23 and P24). The data were then
82combined with those of 23 RT samples used in our previous study [7], yielding a combined data set of
8330 RT samples from 13 COVID-19 patients (**Table S1**).

84 A sustained viral population should be supported by an active viral replication [8]. We firstly
85estimated the viral transcription activity within RT samples using viral sub-genomic messenger RNAs
86(**sgmRNAs**), which is only synthesised in infected host cells [9]. The sgmRNA abundance was
87measured as the ratio of short reads spanning the transcription regulatory sequence (TRS) sites to
88the viral genomic reads. The sgmRNA abundance within nasal and throat swab samples was similar
89to that within sputum samples (**Figure 1a**), reflecting an active viral replication in the upper respiratory
90tract. Notably, the patient P01, who eventually passed away due to COVID-19, showed the highest
91level of sgmRNA abundance (**Figure S1**). Among the samples from patients with improved clinical
92outcomes, their viral Ct (cycle threshold) value of reverse transcriptase quantitative PCR (RT-qPCR)
93negatively correlated with the days post symptoms onset (**Figure 1b**). Interestingly, the sgmRNA
94abundance showed a similar trend across time (**Figure 1c**). This result is further strengthened by the
95positive correlation between sgmRNA abundance and the Ct value (**Figure 1d**), reflecting a direct
96biological association between viral replication and viral shedding in the respiratory tract tissues.

97 Using the metatranscriptomic data, we identified 66 iSNVs in protein encoding regions with the
98alternative allele frequency (AAF) ranged from 5% to 95% (**Table S2 and Table S3**). The identified
99iSNVs showed a high concordance between the AAFs derived from metatranscriptomic and that from
100hybrid-capture sequences (Spearman's $\rho = 0.81$, $P < 2.2e-16$; **Figure S2**). We firstly looked for
101signals of natural selection against intra-host variants. Using the Fisher's exact test, we compared the
102number of iSNV sites on each codon position against that of the other two positions and detected a
103significant difference among them (codon position 1 [$n = 10$, $P = 0.02$], 2 [$n = 21$; $P = 1$] and 3 [$n = 35$;

104 $P = 0.03$). However, those iSNVs did not show a discriminated AAF among the non-synonymous and
 105 synonymous categories (**Figure 2a**), suggesting that most non-synonymous variants were not under
 106 an effective purifying selection within the host. Among the 66 identified iSNVs, 30 were coincided with
 107 the consensus variants in the public database (**Table S2**). Those iSNVs were categorised into
 108 common iSNVs, while the iSNVs presented in a single patient were categorised into rare iSNVs.
 109 Interestingly, the common iSNVs had a significant higher minor allele frequency compared to the rare
 110 iSNVs (**Figure S3**; Wilcoxon rank sum test, $P = 2.7e-05$), suggesting that they may have been
 111 developed in earlier strains before the most recent infection.

112 We then estimated the viral genetic distance among samples in a pairwise manner based on their
 113 iSNVs and allele frequencies. The samples were firstly categorised into intra-host pairs (serial
 114 samples from the same host), intra-household pairs and inter-household pairs (**Figure 2b and Table**
 115 **S4**). As expected, the intra-host pairs had the lowest genetic distance compared to either intra-
 116 household pairs (Wilcoxon rank sum test, $P = 0.018$) and inter-household pairs (Wilcoxon rank sum
 117 test, $P < 2.22e-16$). Interestingly, the genetic distance between intra-household pairs was significantly
 118 lower than that of inter-household pairs (**Figure 2b**; Wilcoxon rank sum test, $P = 0.03$), supporting a
 119 direct passage of virions among intra-household individuals. Nonetheless, we only observed a few
 120 minor variants shared among intra-household pairs, suggesting that the estimated genetic similarity
 121 was mostly determined by consensus nucleotide differences (**Figure 2c,d**). Specifically, in one intra-
 122 household pair (P23 and P24), one patient (P23) contained iSNVs that were coincided with the linked
 123 variants, C8782T and T28144C, suggesting that this patient may have been co-infected by genetically
 124 distinct viruses. However, the strain carrying C8782T and T28144C was not observed in the intra-
 125 household counterpart (P24). It is likely that there is a narrow transmission bottleneck allowing only
 126 the major strain to be circulated, if P23 was infected by all the observed viral strains before the
 127 transmission.

128 The transmission bottlenecks among intra-household pairs were estimated using a beta binomial
 129 model, which was designed to allow some temporal stochastic dynamics of viral population in the
 130 recipient [10]. Here, we defined the donor and recipient within the intra-household pairs according to
 131 their dates of the first symptom onset. The estimated bottleneck sizes were 6 (P03 and P11) and 8
 132 (P23 and P24) for the two intra-household pairs (**Table S5**). This result is consistent with the patterns
 133 observed in many animal viruses and human respiratory viruses [11,12], while the only study reporting

134a loose bottleneck among human respiratory viral infections [13] was argued as the generic
135consequence of shared iSNVs caused by read mapping artefacts [14]. The relatively narrow
136transmission bottleneck sizes is expected to increase the variance of viral variants being circulated
137between transmission pairs [15]. Even after successful transmission, virions carrying the minor
138variants are likely to be purged out due to the frequent stochastic dynamics within the respiratory tract
139[7], which is also consistent with the low diversity and instable iSNV observed among the RT samples.

140 The observed narrow transmission bottleneck suggests that, in general, only a few virions
141successfully enter host cells and eventually cause infection. Although the number of transmitted
142virions is sparse, they can easily replicate in the respiratory tract, given the observed viral replication
143activities in all the RT sample types and the high host-cell receptor binding affinity of SARS-CoV-2
144[16]. The narrow transmission bottleneck also indicate that instant hand hygiene and mask-wearing
145might be particular effective in blocking the transmission chain of SARS-CoV-2.

146 In summary, we have characterized and compared SARS-CoV-2 populations of patients within and
147across households using both metatranscriptomic and hybrid-capture based techniques. Our work
148showed an active viral replication activity in the human respiratory tract and the co-existence of
149genetically distinct viruses within the same host. The inter-host comparison among viral populations
150further revealed a narrow transmission bottleneck between patients from the same households,
151suggesting a dominated role of stochastic dynamics in both inter-host and intra-host evolution. The
152present work enhanced our understanding of SARS-CoV-2 virus transmission and shed light on the
153integration of genomic and epidemiological in the control of this virus.

154 **Materials and methods**

155 **Patient and Ethics statement**

156 Respiratory tract (RT) samples, including nasal swabs, throat swabs, sputum, were collected from 13
157 COVID-19 patients during the early outbreak of the pandemic (from January 25 to February 10 of
158 2020). Those patients were hospitalized at the first affiliated hospital of Guangzhou Medical University
159 (10 patients), the fifth affiliated hospital of Sun Yat-sen University (1 patient), Qingyuan People's
160 Hospital (1 patient) and Yangjiang People's Hospital (1 patient). The research plan was assessed and
161 approved by the Ethics Committee of each hospital. All the privacy information was anonymized.

162

163 **Dataset description**

164 Public consensus sequences were downloaded from GISAID.

165

166 **Real-time RT-qPCR and sequencing**

167 RNA was extracted from the clinical RT samples using QIAamp Viral RNA Mini Kit (Qiagen, Hilden,
168 Germany), which was then tested for SARS-CoV-2 using Real-time RT-qPCR. Human DNA was
169 removed using DNase I and RNA concentration was measured using Qubit RNA HS Assay Kit
170 (Thermo Fisher Scientific, Waltham, MA, USA). After human DNA-depletion, the samples were RNA
171 purified and then subjected to double-stranded DNA library construction using the MGIEasy RNA
172 Library preparation reagent set (MGI, Shenzhen, China) following the method used in the previous
173 study [17]. Possible contamination during experimental processing was tracked using human breast
174 cell lines (Michigan Cancer Foundation-7). The constructed libraries were converted to DNA nanoballs
175 (DNBs) and then sequenced on the DNBSEQ-T7 platform (MGI, Shenzhen, China), generating
176 paired-end short reads with 100bp in length. Most samples were also sequenced using hybrid
177 capture-based enrichment approach that was described in previous study [17]. Briefly, the SARS-
178 CoV-2 genomic content was enriched from the double-stranded DNA libraries using the 2019-
179 nCoVirus DNA/RNA Capture Panel (BOKE, Jiangsu, China). The enriched SARS-CoV-2 genomic
180 contents were converted to DNBs and then sequenced on the MGISEQ-2000 platform, generating
181 paired-end short reads with 100bp in length.

182

183 **Data filtering**

184 Read data from both metatranscriptomic and hybrid capture based sequencing were filtered following
185 the steps described in the previous research [17]. In brief, short read data were mapped to a database
186 that contains major coronaviridae genomes. Low-quality, adaptor contaminations, duplications, and
187 low-complexity within the mapped reads were removed to generate the high quality coronaviridae-like
188 short read data.

189

190 **Profiling of sub-genomic messenger RNA (sgmRNAs)**

191 Coronaviridae-like short reads were mapped to the reference genome (EPI_ISL_402119) using the
192 aligner HISAT2 [18]. Reads spanning the transcription regulatory sequence (TRS) sites of both leader
193 region and the coding genes (S gene, ORF3a, 6, 7a, 8, E, M and N gene) were selected to represent
194 the sgmRNAs. The junction sites were predicted using RegTools junctions extract [19]. The ratio of
195 sgmRNA reads to the viral genomic RNA reads (sgmRNA ratio) was used to estimate the relative
196 transcription activity of SARS-CoV-2.

197

198 **Detection of intra-host variants**

199 We defined an intra-host single nucleotide variant (iSNV) as the co-existence of an alternative allele
200 and the reference allele at the same genomic position within the same sample. To identify iSNV sites,
201 paired-end metatranscriptomic coronaviridae-like short read data were mapped to the reference
202 genome (EPI_ISL_402119) using BWA aln (v.0.7.16) with default parameters [20]. The duplicated
203 reads were detected and marked using Picard MarkDuplicates (v. 2.10.10)
204 (<http://broadinstitute.github.io/picard>). Nucleotide composition of each genomic position was
205 characterized from the read mapping results using pysamstats (v. 1.1.2)
206 (<https://github.com/alimanfoo/pysamstats>). The variable sites of each sample were identified using the
207 variant caller LoFreq with default filters and the cut-off of 5% minor allele frequency. After filtering the
208 sites with more than one alternative allele, the rest sites were regarded as iSNV sites. All the iSNVs
209 with less than five metatranscriptomic reads were verified using the hybrid capture data (at least two
210 reads). The identified iSNVs were then annotated using the SnpEff (v.2.0.5) with default settings [21].

211

212 Genetic distance

213 The genetic distance between sample pairs was calculated using L1-norm distance, as defined by the
214 following formula. The L1-norm distance (D) between sample pairs is calculated by summing the
215 distance of all the variable loci (N). The distance on each variable locus is calculated between vectors
216 (p and q for each sample) of possible base frequencies ($n=4$).

$$217 D = \sum_{k=1}^N \sum_{i=1}^n |p_i - q_i|$$

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

$$221 d(p, q) = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}$$

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

224

225 Beta binomial model of bottleneck size estimation

226 A beta-binomial model was used to estimate bottleneck sizes between donors and recipients. Here,
227 the bottleneck size represents the number of virions that pass into the recipient and finally shape the
228 sequenced viral population. The patient with the earlier symptom onset date was defined as the
229 donor, while the other was defined as the recipient. The maximum-likelihood estimates (MLE) of
230 bottleneck sizes were estimated within 95% confidence intervals.

231References

2321. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with
233 human respiratory disease in China. *Nature*. 2020. doi:10.1038/s41586-020-2008-3
2342. Zhou P, Yang X Lou, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak
235 associated with a new coronavirus of probable bat origin. *Nature*. 2020. doi:10.1038/s41586-
236 020-2012-7
2373. Smith EC, Sexton NR, Denison MR. Thinking Outside the Triangle: Replication Fidelity of the
238 Largest RNA Viruses. *Annu Rev Virol*. 2014;1: 111–132. doi:10.1146/annurev-virology-
239 031413-085507
2404. Shen Z. Genomic diversity of SARS-CoV-2 in Coronavirus Disease 2019 patients. *bioRxiv*.
241 2019; 1–27.
2425. Butler DJ, Mozsary C, Meydan C, Danko D, Fook J, Rosiene J, et al. Shotgun Transcriptome
243 and Isothermal Profiling of SARS-CoV-2 Infection Reveals Unique Host Responses, Viral
244 Diversification, and Drug Interactions. *bioRxiv*. 2020. doi:10.1101/2020.04.20.048066
2456. Fraser C, Crook D, Peto T, Andersson M, Jeffries K, Eyre D, et al. Shared SARS-CoV-2
246 diversity suggests localised transmission of minority variants. 2020.
2477. Wang Y, Wang D, Zhang L, Sun W, Zhang Z, Chen W, et al. Intra-host Variation and
248 Evolutionary Dynamics of SARS-CoV-2 Population in COVID-19 Patients. *bioRxiv*. 2020.
249 doi:10.1101/2020.05.20.103549
2508. Wölfel R, Corman VM, Guggemos W, Seilmaier M, Zange S, Müller MA, et al. Virological
251 assessment of hospitalized patients with COVID-2019. *Nature*. 2020. doi:10.1038/s41586-020-
252 2196-x
2539. Sola I, Almazán F, Zúñiga S, Enjuanes L. Continuous and Discontinuous RNA Synthesis in
254 Coronaviruses. *Annu Rev Virol*. 2015;2: 265–288. doi:10.1146/annurev-virology-100114-
255 055218
25610. Sobel Leonard A, Weissman DB, Greenbaum B, Ghedin E, Koelle K. Transmission Bottleneck
257 Size Estimation from Pathogen Deep-Sequencing Data, with an Application to Human
258 Influenza A Virus. *J Virol*. 2017;91: 0–19. doi:10.1128/jvi.00171-17
25911. Zwart MP, Elena SF. Matters of Size: Genetic Bottlenecks in Virus Infection and Their
260 Potential Impact on Evolution. *Annu Rev Virol*. 2015;2: 161–179. doi:10.1146/annurev-

261 virology-100114-055135

26212. McCrone JT, Woods RJ, Martin ET, Malosh RE, Monto AS, Luring AS. Stochastic processes
263 constrain the within and between host evolution of influenza virus. *Elife*. 2018;7: 1–19.
264 doi:10.7554/eLife.35962

26513. Poon LLM, Song T, Rosenfeld R, Lin X, Rogers MB, Zhou B, et al. Quantifying influenza virus
266 diversity and transmission in humans. *Nat Genet*. 2016;48: 195–200. doi:10.1038/ng.3479

26714. Xue KS, Bloom JD. Reconciling disparate estimates of viral genetic diversity during human
268 influenza infections. *Nat Genet*. 2019;51: 1298–1301. doi:10.1038/s41588-019-0349-3

26915. Xue KS, Moncla LH, Bedford T, Bloom JD. Within-Host Evolution of Human Influenza Virus.
270 *Trends Microbiol*. 2018;26: 781–793. doi:10.1016/j.tim.2018.02.007

27116. Shang J, Ye G, Shi K, Wan Y, Luo C, Aihara H, et al. Structural basis of receptor recognition
272 by SARS-CoV-2. *Nature*. 2020. doi:10.1038/s41586-020-2179-y

27317. Xiao M, Liu X, Ji J, Li M, Li J, Sun W, et al. Multiple approaches for massively parallel
274 sequencing of HCoV-19 genomes directly from clinical samples 2. 2020; 33.
275 doi:10.1101/2020.03.16.993584

27618. Kim D, Langmead B, Salzberg SL. HISAT: A fast spliced aligner with low memory
277 requirements. *Nat Methods*. 2015. doi:10.1038/nmeth.3317

27819. Feng Y-Y, Ramu A, Cotto KC, Skidmore ZL, Kunisaki J, Conrad DF, et al. RegTools:
279 Integrated analysis of genomic and transcriptomic data for discovery of splicing variants in
280 cancer. *bioRxiv*. 2018. doi:10.1101/436634

28120. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform.
282 *Bioinformatics*. 2010. doi:10.1093/bioinformatics/btp698

28321. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating
284 and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin)*. 2012.
285 doi:10.4161/fly.19695

286

287 **DATA AVAILABILITY**

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

290

291 **DISCLOSURE STATEMENT**

292 No conflict of interest was reported by the authors

293

294 **ACKNOWLEDGEMENTS**

295 This study was approved by the Health Commission of Guangdong Province to use patients'
296 specimen for this study. This study was funded by grants from The National Key Research and
297 Development Program of China (2018YFC1200100, 2018ZX10301403, 2018YFC1311900), the
298 emergency grants for prevention and control of SARS-CoV-2 of Ministry of Science and Technology
299 (2020YFC0841400), Guangdong province (2020B111107001, 2020B111108001, 2020B111109001,
300 2018B020207013, 2020B111112003), the Guangdong Province Basic and Applied Basic Research
301 Fund (2020A1515010911), Guangdong Science and Technology Foundation (2019B030316028),
302 Guangdong Provincial Key Laboratory of Genome Read and Write (2017B030301011) and
303 Guangzhou Medical University High-level University Innovation Team Training Program (Guangzhou
304 Medical University released [2017] No.159). This work was supported by the Shenzhen Municipal
305 Government of China Peacock Plan (KQTD2015033017150531). This work was supported by China
306 National GeneBank (CNGB). We thank the patients who took part in this study.

307

308 **AUTHOR CONTRIBUTIONS**

309 D.W., Y.X., J.L., W.Z. and J.Z. conceived the study, Y.W., L.Z., and Y.L. collected clinical specimen
310 and executed the experiments. D.W., W.S., X.C. and J.J. analyzed the data. All the authors
311 participated in discussion and result interpretation. D.W., Y.W., and Z.Z. wrote the manuscript. All
312 authors revised and approved the final version.

Figures

Figure 1. sub-genomic messenger RNAs (sgmRNAs)

a, The ratio of sgmRNA of each respiratory sample type (nasal, throat swabs and sputum). b, Correlation between the cycle threshold (Ct) of RT-qPCR and the days post symptoms onset. c, Correlation between estimated sgmRNA ratio and the days post symptoms onset. d, Correlation between estimated sgmRNA ratio and the cycle threshold of RT-qPCR.

320

321 Figure 2. Allele frequency changes of transmission pairs

322 **a**, Box plots showing the alternative allele frequency (AAF) distribution of synonymous and non-
323 synonymous intra-host variants. **b**, Box plots representing the L1-norm distance distribution among
324 sample pairs. Each dot represents the genetic distance between each sample pair. **c**, The AAF of
325 donor iSNVs in transmission pairs. Allele frequencies under 5% and over 95% were adjusted to 0%
326 and 1, respectively. **d**, Heatmap representing the alternative allele frequencies (AAFs) of consensus
327 and intra-host single nucleotide variants (iSNVs) of the two transmission pairs.

328 **SUPPLEMENTARY INFORMATION**

329 **Table S1. Demography and clinical outcomes of COVID-19 patients**

330 **Table S2. Summary of iSNVs**

331 **Table S3. Frequency of iSNVs**

332 **Table S4. Inter-host genetic distance (L1 and L2-norm)**

333 **Table S5. Bottleneck size of intra-household pairs**

334

335

336 **Figure S1. Transcription profile of sub-genomic messenger RNAs (sgmRNAs) of each patient.**

337

338

339 **Figure S2. Concordance between minor alternative allele frequencies (AAFs) derived from**
 340 **metagenomic and hybrid capture data.**

341

342 Figure S3. Alternative allele frequency (AAF) distribution of rare and common iSNVs

343

Figure

Figure

Figure

Figure