Both Simulation and Sequencing Data Reveal Multiple SARS-CoV-2 Variants Coinfection in COVID-19 Pandemic

3

4 Yinhu Li^{1,#}, Yiqi Jiang^{1,#}, Zhengtu Li^{2,#}, Yonghan Yu^{1,#}, Jiaxing Chen^{1,3}, Wenlong
5 Jia¹, Yen Kaow Ng⁴, Feng Ye^{2,*}, Bairong Shen^{5,*}, Shuai Cheng Li^{1,*}

6

7 ¹ Department of Computer Science, City University of Hong Kong, Hong Kong 999077, China

8 ² State Key Laboratory of Respiratory Disease, National Clinical Research Center for Respiratory
9 Disease, Guangzhou Institute of Respiratory Health, the First Affiliated Hospital of Guangzhou
10 Medical University, Guangzhou 510120, China

11 ³ Department of Computer Science, Hong Kong Baptist University, Hong Kong 999077, China

12 ⁴ Department of Computer Science, Faculty of Information and Communication Technology,
13 Universiti Tunku Abdul Rahman, Kajang 43000, Malaysia

14 ⁵ Institutes for Systems Genetics, Frontiers Science Center for Disease-related Molecular Network,
15 West China Hospital, Sichuan University, Chengdu 610041, China

16

17 [#]**Equal contribution.**

18 ^{*}**Corresponding authors.**

19 E-mail: shuaicli@cityu.edu.hk (Li SC), bairong.shen@scu.edu.cn (Shen B),
20 tu276025@gird.cn (Ye F).

21

22 **Abstract**

23 SARS-CoV-2 is a single-stranded RNA betacoronavirus with a high mutation rate.
24 The rapidly emerged SARS-CoV-2 variants could increase the transmissibility,
25 aggravate the severity, and even fade the vaccine protection. Although the
26 coinfections of SARS-CoV-2 with other respiratory pathogens have been reported,
27 whether multiple SARS-CoV-2 variants coinfection exists remains controversial. This
28 study collected 12,986 and 4,113 SARS-CoV-2 genomes from the GISAID database
29 on May 11, 2020 (GISAID20May11) and April 1, 2021 (GISAID21Apr1),
30 respectively. With the single-nucleotide variants (SNV) and network clique analysis,
31 we constructed the single-nucleotide polymorphism (SNP) coexistence networks and
32 noted the SNP number of the maximal clique as the coinfection index. The
33 coinfection indices of GISAID20May11 and GISAID21Apr1 datasets were 16 and 34,
34 respectively. Simulating the transmission routes and the mutation accumulations, we
35 discovered the linear relationship between the coinfection index and the coinfecting
36 variant number. Based on the linear relationship, we deduced that the COVID-19
37 cases in the GISAID20May11 and GISAID21Apr1 datasets were coinfecting with 2.20
38 and 3.42 SARS-CoV-2 variants on average. Additionally, we performed Nanopore
39 sequencing on 42 COVID-19 patients to explore the virus mutational characteristics.
40 We found the heterozygous SNPs in 41 COVID-19 cases, which support the
41 coinfection of SARS-CoV-2 variants and challenge the accuracy of phylogenetic
42 analysis. In conclusion, our findings reported the coinfection of SARS-CoV-2 variants
43 in COVID-19 patients, demonstrated the increased coinfecting variants number in the
44 epidemic, and provided clues for the prolonged viral shedding and severe symptoms
45 in some cases.

46
47
48 **KEYWORDS:** SARS-CoV-2 variant coinfection; Viral transmission simulation;
49 Coinfection index; Heterozygous SNPs

50

51 Introduction

52 Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has infected more
 53 than 176.5 million persons, with more than 3.8 million deaths at the time of preparing
 54 this manuscript [1, 2]. The virus is an enveloped and single-stranded RNA
 55 betacoronavirus of 30k base-pairs, which belongs to the family Coronaviridae [1].
 56 Since the year 2000, we have witnessed and experienced three highly widespread
 57 pathogenic coronaviruses in human populations, and the other two are severe acute
 58 respiratory syndrome (SARS)-CoV in 2002-2003, and Middle East Respiratory
 59 Syndrome (MERS)-CoV in 2012 [3]. All three viruses can lead to acute respiratory
 60 distress syndrome (ARDS) in the human hosts, which may cause pulmonary fibrosis
 61 and lead to permanent lung function reduction or death [4]. Although with lower
 62 mortality rates than SARS-CoV and MERS-CoV, SARS-CoV-2 could invade host
 63 cells by binding to the ACE2 on the host cell surface and cause rapid spread among
 64 people [5].

65 To address the challenges, researchers conducted various studies to explore the
 66 genomic sequences of SARS-CoV-2 [6-8]. Qianqian Li *et al.* have analyzed 13,406
 67 spike sequences of SARS-COV-2 variants in the GISAID database and divided the
 68 SARS-CoV-2 variants into seven evolutionary groups using neutralizing monoclonal
 69 antibodies [6]. Correspondingly, the Centers for Disease Control and Prevention also
 70 reported the new emerged SARS-CoV-2 variants that circulating globally, including
 71 B.1.1.7 lineage in the United Kingdom, B.1.351 lineage in Nelson Mandela Bay and
 72 South Africa, P.1 lineage in Japan and Brazil, B.1.429 lineage in the United States,
 73 *etc.* [9]. From Pengfei Wang *et al.*'s study, we learned that the extensive mutations in
 74 the spike protein of B.1.1.7 and B.1.351 variants could enhance their resistance to the
 75 neutralization by convalescent and post-vaccination sera. These reports enforce the
 76 notion that the newly emerged SARS-CoV-2 variants would increase the viral
 77 transmissibility and disease severity and reduce the protective ability of vaccines [10].

78 Besides the rapidly emerged SARS-CoV-2 variants, previous studies also reported
 79 the coinfection of SARS-CoV-2 with other respiratory pathogens [11, 12]. David Kim

80 and his colleagues found that 116 COVID-19 patients were also positive for other
81 microbial pathogens, such as influenza A/B, respiratory syncytial virus, human
82 metapneumovirus, and *Chlamydia pneumoniae* [11]. Also, the reinfection with
83 different SARS-CoV-2 variants in a COVID-19 patient has been reported. Richard L
84 Tillett *et al.* presented a COVID-19 patient who tested positive for SARS-CoV-2 on
85 April 2020 and was reinfected by a different SARS-CoV-2 variant on June 2020 [13].
86 The astonishing discovery was hard to explain why previous exposure to
87 SARS-CoV-2 failed to provide immunity protection to the patient. Since coinfection
88 is prevalent in viral infections [14-16], the studies inspire us to explore whether
89 coinfection of multiple SARS-CoV-2 variants exists in COVID-19 patients, providing
90 clues for prolonged viral shedding time and severe symptom [17].

91 Here, we collected 12,986 SARS-CoV-2 genomic sequences from the GISAID
92 database on May 11, 2020, constructed single-nucleotide polymorphisms (SNP)
93 coexistence network, and found a maximal clique of 16 coexisted loci. By simulating
94 the SNVs accumulation with SARS-CoV-2 transmission, we discovered 2.20
95 averaged coinfecting variants in the COVID-19 patients with the coinfection index. To
96 validate the methods and results, we extracted 4,113 additional genomes from the
97 GISAID database on April 1, 2021, and discovered an increased coinfecting variants
98 number of 3.42. Then, we performed Nanopore sequencing on the sputum samples
99 from 42 COVID-19 patients and found the heterozygous SNPs on some loci of the
100 SARS-CoV-2 genome, confirming the multiple variants coinfection. Hence, our study
101 proposed a computational simulating method to detect the number of the coinfecting
102 variants in COVID-19 patients, confirmed the coinfection of multiple SARS-CoV-2
103 variants, and implied the increased coinfecting variants in the epidemic.

104

105

106 **Materials and methods**

107 **Ethics Statement**

108 The First Affiliated Hospital approved this study of Guangzhou Medical University,
109 and the sample and data collection procedures were conducted following the
110 principles expressed in the Declaration of Helsinki. All patients provided written
111 informed consent and volunteered to receive investigation for scientific research.

112 **GISAID datasets and mutation detection**

113 This study collected SARS-CoV-2 genomic sequences from the GISAID database
114 (<https://www.gisaid.org/>) and divided them into two genomic datasets according to
115 their releasing date: For the 12,986 SARS-CoV-2 genomic sequences published
116 before May 11, 2020, we noted them as GISAID20May11 dataset; For the 4,113
117 SARS-CoV-2 genomic sequences posted on April 1, 2021, we noted them as
118 GISAID21Apr1 dataset. All genomes in these two datasets were tagged as complete
119 (>29,000 bp) and high coverage (<1% Ns with <0.05% unique amino acid mutation)
120 in the GISAID. We adopted MUMmer (version 3.23) to obtain the SNVs of the
121 SARS-CoV-2 genomes [25]. Each SARS-CoV-2 genome is aligned with the
122 SARS-CoV-2 reference genome (MN908947.3) to obtain the homology region using
123 the nucmer function with the default parameters [25]. Then we got the SNPs matrix
124 from the alignment results with show-SNPs function [25] and prepared for the SNV
125 clique analysis.

126 **SNP coexistence network and clique analysis**

127 To evaluate the complexity of SNPs co-occurrences within the GISAID dataset, we
128 applied single-nucleotide variant (SNV) clique analysis by in-house scripts. Firstly,
129 we considered a pair of SNPs from two different loci as complex if it occurred in at
130 least one variant of the GISAID datasets. However, a complex paired-loci is hard to
131 be explained in phylogeny, and it may happen by chance. Therefore, to remove such a
132 possibility, we performed an analysis based on SNV cliques instead.

133 After obtaining all SNPs, we checked the alleles at every locus of the
134 SARS-CoV-2 genome. Over 92% of the SNPs loci (5,671/6,178) had two alleles.
135 Focusing on the loci with two alleles, we removed the SNPs loci with three or four
136 alleles. We labeled the major allele of SNP locus as R and the minor allele as A. Thus,
137 it had four possible genetic combinations for every pair of two SNPs loci: RR, RA,

138 AR, AA. We recognized each SNP locus as a vertex and created an edge between a
139 loci pair only if all four genetic combinations existed in at least one assembly genome
140 within the GISAID dataset (Figure 1A). We obtained the maximal clique from the
141 network. Based on the cliques, we can tell whether the SARS-CoV-2 coinfection
142 exists since the existence of a large clique will be intractable to explain using
143 phylogeny.

144 **Prediction of coinfecting variant number based on the simulation**

145 With the SNVs in the collected genomic sequences, we predicted the coinfecting
146 variant numbers by simulations with the mutation rate (r) and the average variant
147 number (w). In previous reports, the estimated mutation rate of SARS-CoV-2 by
148 several groups ranged from 2.88×10^{-6} to 3.45×10^{-6} substitutions per site per day
149 [26-28]. However, the obtained SNVs number distribution curve in our test does not
150 fit the distribution curve from the real data set with a mutation rate of 3.0×10^{-6}
151 (Supplementary Figure 1). The mutation rate we used has four values, which are
152 1.5×10^{-6} , 2×10^{-6} , 2.5×10^{-6} , and 3×10^{-6} . The average variant number in the
153 simulation with 15 values ranged from 1.2 to 4, with an interval of 0.2. The
154 distribution of variant numbers in all samples conformed to Poisson distribution with
155 λ equals the average variant number.

156 **Sample collection**

157 To confirm the coinfection of SARS-CoV-2 variants, we performed RT-PCR on the
158 sputum samples collected from COVID-19 patients. Forty-two patients were recruited
159 from the First Affiliated Hospital of Guangzhou Medical University and Guangdong
160 Second Provincial General Hospital, China (Supplementary Table 1). The sputum
161 samples from the patients were inactivated under 56°C for 30 minutes following
162 WHO and Chinese guidelines [29-31]. The specimens were stored at 4°C until ready
163 for shipment to the Guangdong Centers for Disease Control and Prevention.

164 **Nanopore sequencing**

165 For the samples, we extracted the total RNA from the samples according to the
166 protocol of RNA isolation kit (RNAqueous Total RNA isolation Kit, Invitrogen,
167 China), and determined the RNA concentration by Qubit (ThermoFisher Scientific,

China). Based on two pools of primers (98 pairs of primers in total) (Supplementary Table 2), the entire genomic sequence of SARS-CoV-2 was amplified segmentally by reverse transcription. Then, the libraries were built by adding the adapter and barcode to the amplified genomic fragments with a Nanopore library construction kit (EXP-FLP002-XL, Flow Cell Priming Kit XL, YILIMART, China). The samples were sequenced on the MinIon sequencing platform (Oxford Nanopore Technologies, U.K.).

Nanopore data filtration

MinIon sequencer generated Fast5 format data, which was converted into fastq format with guppy basecaller (version 3.0.3). By applying NanoFilt (version 1.7.0) [32], we performed data filtration on the raw fastq data with the following criteria: the read lengths should be longer than 100 bp after removing the adapter sequences overall quality of reads should be higher than 10. Furthermore, due to the random connection of multiplex RT-PCR amplicons, the chimeric reads should be processed to avoid false identification of virus recombination or host integration. Therefore, we positioned the primers on the sequencing reads to identify the chimeric reads, split the identified chimeric reads into segments corresponding to PCR amplicons, and retained the final reads by aligning the segments to the viral genome (Supplementary Figure 2). This method allowed us to salvage a huge amount of sequencing data, leading to more accurate alignment and higher coverage.

Mutation detection with Nanopore data

We aligned the filtered and segmented reads to the SARS-CoV-2 reference genome (MN908947.3) with Minimap2 by applying the default parameters for Oxford Nanopore reads [33]. The aligned PCR amplicons were separated according to the corresponding primer pool. With the separated alignment results, the genomic variations with average quality larger than ten were called with bcftools (version 1.8) [34]. Mutations with less than ten supported reads were filtered. To reduce the PCR amplification effects, we also filtered the variations within ten bp upstream or downstream of the primer region within the corresponding primer pool. The filtered mutations for different primer pools were then merged as the final mutations. The

198 final mutations were annotated by in-house software based on the gene information in
199 the SARS-CoV-2 reference genome.

200

201

202 **Results**

203 **Discovery of the 16-SNV-clique with the GISAID20May11 dataset**

204 The GISAID20May11 dataset contains 12,986 SARS-CoV-2 genomes published
205 between December 30, 2019, and May 11, 2020. After filtering 1,804 duplicated
206 sequences, we aligned the rest of 11,182 viral genomes to the SARS-CoV-2 reference
207 genome to obtain SNVs. Then, we removed three viral genomes with over 1,000
208 SNVs and obtained 11,179 genomes for the following-up analysis. With 57,548 SNVs
209 on 6,178 SNPs loci, we performed SNP clique analysis (**Figure 1A**) and constructed
210 the SNP coexistence networks with 1,150 vertices and 8,003 edges. Among the
211 networks, we discovered the maximal clique with 16 coexisted loci (**Figure 1B**). With
212 the result, we deduced that some SARS-CoV-2 assembly genomes were mixed
213 sequences of multiple coinfecting variants, except the incredible-fast mutation.

214 **Coinfection index to determine the SARS-CoV-2 variant number in a sample**

215 We selected the maximal clique from the SNP coexistence networks and noted its size
216 as the coinfection index. We further determine the average coinfecting variants number
217 with computational simulations. By simulating the transmission route tree of
218 COVID-19, we traced the virus transmission among the infected individuals. Based
219 on the publishing date of the sequences, we selected the sequences at the same
220 transmission period as the simulated sequences and calculated the coinfection index
221 using SNP clique analysis. Using different mutation rates and the average variant
222 number in the simulation, we could obtain a chart of the average variant number
223 against the coinfection index under a specific mutation rate (**Figure 2A**). During
224 transmission, the variants in a sample at the child node were randomly inherited from
225 the sample at the parent node. The variants would generate new SNVs based on a
226 given simulated mutation rate (**Figure 2B**). In the simulation, we proposed two

227 methods of how the coinfecting variants in a sample construct their assembly genome.
 228 The first method randomly selected a variant from the coinfecting multiple variants in
 229 the sample, and reported the SNVs in this variant. The second method (the mixed
 230 method) generated an assembly genome, which was a mixture of all variants. We split
 231 the genome as windows with a fixed size of 100 bp for the second method, and each
 232 window comes from a randomly selected variant in the sample. Using these two
 233 methods, we obtained the SNVs in the assembly genome (**Figure 2C**).

234 After plotting the coinfection index against the average variant number, we got
 235 two regression lines between them (**Figure 3A**). With the results, we noticed that only
 236 the regression line based on the mixed method could achieve a coinfection index of 16
 237 for the GISAID20May11 dataset. With the regression lines of these two methods, we
 238 concluded that the 16-SNV clique from the GISAID20May11 dataset should result
 239 from coinfection, and the assembly genome comes from the mixed sequencing data of
 240 the coinfecting variants.

241 Then, we determined the averaged variant number in the GISAID20May11
 242 dataset with the coinfection index line. We performed regression analysis between
 243 averaged variant number and coinfection index and discovered the significant linear
 244 relationship between them with method 2 (F-statistic p-value < 2.2e-16, adjusted
 245 R-squared = 0.79, Figure 3A). According to the obtained fitting equation, we deduced
 246 that the corresponding average variant number was 2.20 when the coinfection index
 247 was 16 (Figure 3A).

248 **Coinfection index increased along with the COVID-19 pandemic**

249 With the GISAID20May20 dataset, we obtained a maximal clique with 34 coexisted
 250 SNPs from 140,348 SNVs on 6,415 SNPs loci (**Figure 1C**). Then, we constructed the
 251 coinfection index curve with the GISAID21Apr1 dataset and determined the average
 252 variants number in this dataset. The genomes of GISAID21Apr1 were sampled from
 253 five different continents. Europe provided primary samples as 3,023 samples were
 254 from Europe, and the rest 1,047 samples were from North America, 27, 12, and 4
 255 samples were from Asia, South America, and Oceania, respectively. While, we found
 256 28 SNPs existed in over 3,000 samples, which reveals those samples should have the

257 same or related ancestor. We altered the simulated procedure since we assumed those
258 samples had the same ancestor to fit the SNVs distribution in the GISAID21Apr1
259 dataset. The regressed linear of the coinfection index and the average number of
260 variants showed a significant linear relationship (F-statistic p-value < 2.2e-16,
261 adjusted R-squared = 0.69, **Figure 3B**). The fitting equation revealed the average
262 stain number of 3.42 in the GISAID21Apr1 dataset. The pandemic of COVID-19
263 made the virus could transfer between continents and increased the coinfection of
264 different variants.

265 **Sequencing data statistics for the 42 COVID-19 patients**

266 For the 42 COVID-19 patients enrolled from the First Affiliated Hospital of
267 Guangzhou Medical University and Guangdong Second Provincial General Hospital,
268 we performed Nanopore sequencing on their sputum samples for SARS-CoV-2
269 genome acquirement and mutation detection (Supplementary Table 1). After
270 sequencing on the multiple-PCR products, a total of 7,877,736 clean reads were
271 generated, with an average of $187,565 \pm 143,719.55$ (Mean $\pm$ SD) reads per sample
272 (**Figure 4**). To eliminate the chimeric reads formed by the unintended random
273 connection of multiplex PCR amplicons, we developed a software tool named
274 CovProfile [18] (Supplementary Figure 2) and perform data filtration and detect the
275 mutations in SARS-CoV-2 variants. Then we discovered that the chimeric reads were
276 making up 1.69% of total sequencing reads. Aligning the clean reads to the
277 SARS-CoV-2 genome and human transcriptome, we discovered that the ratio of
278 primary aligned sequence ranged from 3.86% to 99.74% on the SARS-CoV-2 genome
279 and ranged from 0.13% to 70.5% on the human transcriptome database (Figure 4).
280 Moreover, the SARS-CoV-2 genomic coverage reached over 99.7% with >1800x
281 depth in each sample, ensuring adequate data volume for SNP calling (Supplementary
282 Figure 3).

283 **Identification of heterozygous SNPs on SARS-CoV-2 genome**

284 After aligning the filtered data to the SARS-CoV-2 genome, we detected the
285 mutations of SARS-CoV-2 in the 42 samples (**Figure 5**). Based on these mutations,

we discovered a total of 115 SNPs in all samples, and 108 of them located on the genetic regions, including genes ORF1ab, S, ORF3a, N, M, ORF6, ORF8, and ORF10 (Supplementary Table 3). Furthermore, we discovered the heterozygous SNPs in 41 of the enrolled samples (Figure 5). Since each locus contained only one genotype in a viral genome, the heterozygous SNPs indicated that each host was infected with two variants at least. Moreover, twenty heterozygous SNPs existed in over two samples, such as C865T, A1430T, C8782T, etc (Supplementary Table 3). Notably, we also discovered that 14 samples contained two genotyped SNPs on loci 8,782 and 28,144 simultaneously, which were significant SNPs identified in recent phylogenetic analysis. Meanwhile, we did not find creditable InDels (Insertions and Deletions), structural variations, or viral-host recombination.

Discussion

SARS-CoV-2 posed a significant threat to human lives, and recent studies have reported the rapidly emerged variants and their impact on clinical severity and vaccine protection [7, 9, 19]. In this study, we aimed to detect whether the coinfection of multiple SARS-CoV-2 variants exists in COVID-19 patients, which might associate with frequent homologous recombination and greater clinical severity. This study performed the SNP coexistence network analysis to detect the "coinfection index" based on the maximal clique in the collected GISAID datasets and constructed the relationship between the coinfection index and the average variant number. We deciphered the number of coinfecting variants for SARS-CoV-2 in hosts with the linear regression between the coinfection index and the average variant number. With the GISAID20May20 and GISAID21Apr1 datasets, we discovered that the number of the coinfecting variants increased from 2.20 to 3.42 in the COVID-19 patients. Considering the rapidly emerged SARS-CoV-2 variants worldwide, we hypothesized that the coinfecting variants in hosts would aggravate the clinical severity, increase the change of viral recombination, and posed a greater threat to us [20]. Moreover, the

coinfection index can be applied to other viruses in hosts. Although the coinfection explained the large clique detected in the SNP coexistence networks in the collected datasets, the discoveries still need to be verified experimentally.

To verify the coinfection of multiple SARS-CoV-2 variants, we performed Nanopore sequencing on 42 COVID-19 patients and implemented CovProfile for the sequencing data processing and the genomic mutation detection [18]. Our results confirmed the reliability of the multiplex RT-PCR method in identifying SARS-CoV-2 and discovered the recurrent heterozygous SNPs on 41 of 42 samples. Moreover, we found two genotyped SNPs on loci 8,782 and 28,144 in fourteen patients. Since loci 8,782 and 28,144 were important for SARS-CoV-2 phylogenetic analysis [21], the finding has crucial impacts on the evolution derivation of SARS-CoV-2, as the heterogeneous loci might cause mis-links during viral genomic assembly. Corresponding to the simulation results, the discoveries of heterozygous SNPs confirmed the multiple variants coinfection in the COVID-19 patients.

The discovery of SARS-CoV-2 variants coinfection provided explanations for the severe clinical symptoms in some COVID-19 patients and significantly impacted the application of vaccines [9, 22, 23]. Since vaccines were developed referencing a specific SARS-CoV-2 variant, the infection of variants limited the protection afforded by vaccines [9]. For instance, SARS-CoV-2 B.1.351 variant, which is widely spread in Nelson Mandela Bay and South Africa, can evade the immune response stimulated by the vaccines and greatly reduce the vaccine's protective effect on the population [19]. Moreover, Nicole Pedro et al. also discovered the coinfection of dual SARS-CoV-2 variants in a severity COVID-19 patient in Portugal, which supported our discoveries [17]. Therefore, the coinfection of multiple SARS-CoV-2 variants raised another challenge, and we need to stay alert in the battle against the COVID-19 epidemic.

Although the findings implied the coinfection of multiple SARS-CoV-2 variants in patients from the perspectives of algorithm derivation and mutation detection, this study still has several limitations. In the simulation, we assumed that the first submitted sequence was the source of all SARS-COV-2 variants. While, in the

345 pandemic, the first infective SARS-COV-2 variant should emerge long before being
 346 discovered. The study by Giovanni Apolone *et al.* proposed that SARS-CoV-2
 347 RBD-specific antibodies can already be detected in the serum samples of Italian
 348 cohorts collected in March 2019, indicating that the source variants of all currently
 349 sequenced variants should appear earlier before [24]. Determining the virus's origin is
 350 difficult, so we chose an exact time point during the simulation, but it does not affect
 351 our conclusions on the coinfection of multiple variants in hosts. Moreover, there was
 352 no guarantee considering the quality of the viral variants submitted to GISAID, which
 353 might influence the accuracy and potential phylogenetic study. Last but not least, the
 354 discovered heterozygous SNPs need to be verified with biological duplication, and we
 355 should identify the coinfecting viral lineages in the COVID-19 patients in future study.

356 In conclusion, our study proposed a computational simulating approach to
 357 decipher the number of the coinfecting variants, declared the coinfection of multiple
 358 SARS-CoV-2 variants in COVID-19 patients, and reported the increased coinfecting
 359 variants in the COVID-19 epidemic, reminding us of the threats brought by the
 360 SARS-CoV-2 infection.

361

362

363

364

365 **Data availability**

366 CovProfile is an open-source collaborative initiative available in the GitHub
367 repository (<https://gitlab.deepomics.org/yyh/covprofile>). All other code is available
368 from the authors upon reasonable request. The Nanopore sequencing data in this paper
369 have been deposited in the Genome Sequence Archive in BIG Data Center, Beijing
370 Institute of Genomics (BIG), Chinese Academy of Sciences, under BioProject
371 PRJCA002503 with accession ID CRA002522 (<https://bigd.big.ac.cn/gsa>).

372

373 **Authors' contributions**

374 S.C.L., B.S. and F.Y. proposed the simulation approach and supervised the project.
375 Y.L. and Z.L. performed the samples collection and Nanopore data analysis. Y.J. and
376 Y.Y. collected the public data and optimized the algorithms in simulation. J.C., W.J.
377 and Y.K.N. guided the analysis and optimized the graphs. Y.L., Y.J., Z.L. and Y.Y.
378 interpreted the results and wrote the manuscript. S.C.L., B.S. and F.Y. polished the
379 manuscript. All authors reviewed the article and approved the final manuscript.

380

381 **Competing interests**

382 The authors have declared no competing interests.

383

384 **Acknowledgments**

385 This work was supported by the Chengdu Science and Technology Project for
386 COVID-19 prevention and control [Grant no. 2020-YF05-00281-SN] awarded to B.S.
387 We would like to thank Beijing YuanShengKangTai (ProtoDNA) Genetech Co. Ltd.
388 for their helping in Nanopore sequencing, and we would also like to thank all the
389 doctors in the First Affiliated Hospital of Guangzhou Medical University, for their
390 assistance on the specimen and clinical data collection.

391

392

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.
- [2] Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med* 2020;382:727-33.
- [3] de Wit E, van Doremalen N, Falzarano D, Munster VJ. SARS and MERS: recent insights into emerging coronaviruses. *Nat Rev Microbiol* 2016;14:523-34.
- [4] Picchianti Diamanti A, Rosado MM, Pioli C, Sesti G, Lagana B. Cytokine Release Syndrome in COVID-19 Patients, A New Scenario for an Old Concern: The Fragile Balance between Infections and Autoimmunity. *Int J Mol Sci* 2020;21.
- [5] Cyranoski D. Profile of a killer: the complex biology powering the coronavirus pandemic. *Nature* 2020;581:22-6.
- [6] Li Q, Wu J, Nie J, Zhang L, Hao H, Liu S, et al. The Impact of Mutations in SARS-CoV-2 Spike on Viral Infectivity and Antigenicity. *Cell* 2020;182:1284-94 e9.
- [7] Hu J, Peng P, Wang K, Fang L, Luo FY, Jin AS, et al. Emerging SARS-CoV-2 variants reduce neutralization sensitivity to convalescent sera and monoclonal antibodies. *Cell Mol Immunol* 2021.
- [8] Hourdel V, Kwasiborski A, Baliere C, Matheus S, Batejat CF, Manuguerra JC, et al. Rapid Genomic Characterization of SARS-CoV-2 by Direct Amplicon-Based Sequencing Through Comparison of MinION and Illumina iSeq100(TM) System. *Front Microbiol* 2020;11:571328.
- [9] John P. Moore PAO. SARS-CoV-2 Vaccines and the Growing Threat of Viral Variants. *JAMA* 2021.
- [10] Burioni R, Topol EJ. Assessing the human immune response to SARS-CoV-2 variants. *Nat Med* 2021.
- [11] Kim D, Quinn J, Pinsky B, Shah NH, Brown I. Rates of Co-infection Between SARS-CoV-2 and Other Respiratory Pathogens. *JAMA* 2020;323:2085-6.
- [12] Kondo Y, Miyazaki S, Yamashita R, Ikeda T. Coinfection with SARS-CoV-2 and influenza A virus. *BMJ Case Rep* 2020;13.

422 [13] Tillett RL, Sevinsky JR, Hartley PD, Kerwin H, Crawford N, Gorzalski A, et al.
423 Genomic evidence for reinfection with SARS-CoV-2: a case study. *Lancet Infect Dis*
424 2021;21:52-8.

425 [14] Teweldemedhin M, Asres N, Gebreyesus H, Asgedom SW. Tuberculosis-Human
426 Immunodeficiency Virus (HIV) co-infection in Ethiopia: a systematic review and
427 meta-analysis. *BMC Infect Dis* 2018;18:676.

428 [15] Furuya-Kanamori L, Liang S, Milinovich G, Soares Magalhaes RJ, Clements
429 AC, Hu W, et al. Co-distribution and co-infection of chikungunya and dengue viruses.
430 *BMC Infect Dis* 2016;16:84.

431 [16] Villamil-Gomez WE, Gonzalez-Camargo O, Rodriguez-Ayubi J, Zapata-Serpa
432 D, Rodriguez-Morales AJ. Dengue, chikungunya and Zika co-infection in a patient
433 from Colombia. *J Infect Public Health* 2016;9:684-6.

434 [17] Pedro N, Silva CN, Magalhaes AC, Cavadas B, Rocha AM, Moreira AC, et al.
435 Dynamics of a Dual SARS-CoV-2 Lineage Co-Infection on a Prolonged Viral
436 Shedding COVID-19 Case: Insights into Clinical Severity and Disease Duration.
437 *Microorganisms* 2021;9.

438 [18] Yonghan Yu ZL, Yinhu Li, Le Yu, Wenlong Jia, Yiqi Jiang, Feng Ye, Shuai
439 Cheng Li. CovProfile: profiling the viral genome and gene expressions of
440 SARS-COV-2. *bioRxiv* 2020.

441 [19] Zhou D, Dejnirattisai W, Supasa P, Liu C, Mentzer AJ, Ginn HM, et al. Evidence
442 of escape of SARS-CoV-2 variant B.1.351 from natural and vaccine-induced sera.
443 *Cell* 2021;184:2348-61 e6.

444 [20] Gao Y, He S, Tian W, Li D, An M, Zhao B, et al. First complete-genome
445 documentation of HIV-1 intersubtype superinfection with transmissions of diverse
446 recombinants over time to five recipients. *PLoS Pathog* 2021;17:e1009258.

447 [21] Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of
448 SARS-CoV-2 genomes. *Proc Natl Acad Sci U S A* 2020;117:9241-3.

449 [22] Williams TC, Burgers WA. SARS-CoV-2 evolution and vaccines: cause for
450 concern? *Lancet Respir Med* 2021.

451 [23] Dong Y, Dai T, Wei Y, Zhang L, Zheng M, Zhou F. A systematic review of
452 SARS-CoV-2 vaccine candidates. *Signal Transduct Target Ther* 2020;5:237.

453 [24] Apolone G, Montomoli E, Manenti A, Boeri M, Sabia F, Hyseni I, et al.
454 Unexpected detection of SARS-CoV-2 antibodies in the prepandemic period in Italy.
455 *Tumori* 2020:300891620974755.

456 [25] Delcher AL, Phillippy A, Carlton J, Salzberg SL. Fast algorithms for large-scale
457 genome alignment and comparison. *Nucleic Acids Res* 2002;30:2478-83.

458 [26] Trevor Bedford RN, James Hadfield, Emma Hodcroft, Misja Ilcisin, Nicola
459 Muller (2020), 'Genomic analysis of nCoV spread. Situation report 2020-01-23.',
460 ResearchWorks Archive.

461 [27] Baric RS, Yount B, Hensley L, Peel SA, Chen W. Episodic evolution mediates
462 interspecies transfer of a murine coronavirus. *J Virol* 1997;71:1946-55.

463 [28] Security JHCfH (2020), 'SARS-CoV-2 Genetics'.

464 [29] Li T. Diagnosis and clinical management of severe acute respiratory syndrome
465 Coronavirus 2 (SARS-CoV-2) infection: an operational recommendation of Peking
466 Union Medical College Hospital (V2.0). *Emerg Microbes Infect* 2020;9:582-5.

467 [30] Prevention CCfDCa. COVID-19 outbreak report2020.

468 [31] Prevention CCfDCa. The guideline of diagnosis and treatment of COVID-19 (the
469 seventh edition)2020.

470 [32] De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack:
471 visualizing and processing long-read sequencing data. *Bioinformatics*
472 2018;34:2666-9.

473 [33] Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*
474 2018;34:3094-100.

475 [34] Li H. A statistical framework for SNP calling, mutation discovery, association
476 mapping and population genetical parameter estimation from sequencing data.
477 *Bioinformatics* 2011;27:2987-93.

478

479

480 **Figure legends**

481 **Figure 1. The workflow of SNP clique analysis and maximal clique in the** 482 **collected GISAID datasets.**

483 **A)** First, we construct the SNP coexisted network from the SNP matrix. Every SNP
484 locus is a vertex, and we add an edge between a loci pair if they have all four major
485 genotypes. We then extract the maximal clique from the network. **B)** The maximal
486 16-SNV-clique was found in the GISAID20May11 dataset with 11,179 SARS-CoV-2
487 genomes. **C)** In the GISAID21Apr1 dataset, the 4,113 SARS-CoV-2 genomes
488 contained the maximal clique of size 34.

489 **Figure 2. The simulation flowchart of viral SNVs in samples.**

490 **A)** We simulated the transmitted route based on known epidemiological information
491 of SARS-CoV-2, and construct the transmission tree. Then we select the sequenced
492 samples based on their releasing date in GISAID database. **B)** Variants number in
493 all samples fit the Poisson distribution with λ equals the average variant number. In a
494 single transmission branch, variants in child nodes are randomly inherited from the
495 parent sample. For every child variant, we generated new SNVs with the period
496 mutation rate. **C)** We simulated two possible assembling situations of samples with
497 multiple variants coinfection and acquired the SNVs list of all samples as the output.

498 **Figure 3. The regression of variant number and the coinfection index.**

499 **A)** The distribution of coinfection index with different average variant numbers in the
500 GISAID20May11 dataset. Method 2 exhibited the linear regression relationship
501 between the coinfection index and average variant number, and the generated formula
502 suggested the mixed variants of the assembly genome in the dataset. **B)** The linear
503 relationship between coinfection index with different average variant numbers in the
504 GISAID21Apr1 dataset. With method 2, the average variant number was 3.4 when the
505 coinfection index was 34.

506 **Figure 4. Statistics of Nanopore sequencing data for the 42 COVID-19 samples.**

507 After the low-quality filtration, we aligned the sequencing data to the SARS-CoV-2
508 genome and human transcriptome, respectively. The histograms in red and green

509 represent the reads number aligned to the SARS-CoV-2 genome and human
510 transcriptome.

511 **Figure 5. SNP distributions in 42 samples gathered from COVID-19 patients.**

512 The alternate alleles were shown in red, while the reference and mutated alleles were
513 in green and red, respectively.

514

515

516 **Supplementary material**

517 **Supplementary Figure 1. The distribution of samples with different SNVs in**
518 **GISAID20May11 dataset and the simulation under different mutation rates.**

519 We had 15 possible average numbers of variants and ten duplicates for each pair of
520 mutation rate and the average variant number. We plotted sample number in all
521 simulations and regress samples number VS number of SNVs of all simulations with
522 specific mutation rate, and the 95% CI region showed in grey.

523 **Supplementary Figure 2. The procedure of chimeric reads identification and**
524 **reads splicing.**

525 **Supplementary Figure 3. The coverage of depth of aligned data in the 42**
526 **COVID-19 samples.**

527 The X coordinate stands for the location of SARS-CoV-2 genome, and the Y
528 coordinate stands for the sequencing depth. The bars with red, yellow, green, pink,
529 brown, light green, purple and dark brown colors stand for the genetic regions of
530 ORF1ab, S, ORF3a, M, ORF6, ORF8, N and ORF10, respectively.

531 **Supplementary Table 1. Physical information of the 42 enrolled COVID-19**
532 **patients.**

533 **Supplementary Table 2. Primers applied for RT-PCR amplification of**
534 **SARS-CoV-2.**

535 **Supplementary Table 3. SNP distributions on 42 COVID-19 patients.**

536

A) Workflow of SNP clique analysis

1) Raw SNPs matrix

	SNP1	SNP2	SNP3	...	SNPm
Sample1	A	C	T		G
Sample2	T	C	T		A
Sample3	T	C	T		G
Sample4	T	G	C		A
...					
Samplen	A	C	C		A

2) Genotypes counts between two SNPs

	SNP1	SNP2	SNP3	...	SNPm
SNP1	-	AC TC 650 443 AG TG 12 4	AT TT 810 234 AC TC 34 0		AA TA 721 311 AG TG 15 2
SNP2	-	-			CA GA 1120 36 CG GG 3 0
...					
SNPm	-	-	-		-

Genotypes
counts

SNPs pair with all four
combination genotypes

3) Construct SNP coexist network: add edge if
two SNPs have all four combinations of genotypes.

4) Find maximal clique and output maximal
clique size.

B) Maximal clique of size 16 in GISAID20May20

C) Maximal clique of size 34 in GISAID21Apr1

A) Simulate transmit route

GISAID collection date information

	Collection date
Sample1	2020-04-08
Sample2	2020-01-27
Sample3	2020-03-18
Sample4	2020-03-20
Sample5	2020-04-13
...	

$t_0 = 2019-12-30$

	Collection date	Δt	Period
Sample1	2020-04-08	102	11
Sample2	2020-01-27	31	4
Sample3	2020-03-18	82	9
Sample4	2020-03-20	84	9
Sample5	2020-04-13	107	11
...

B) SNVs accumulated in strains

C) Two possible assembly genome

A

GISAID20May20

B

GISAID21Apr1

Reads number

● Alternate allele loci

● Reference allele loci