1   **Intra-host Variation and Evolutionary Dynamics of SARS-CoV-2**

2   **Population in COVID-19 Patients**

3

4   **Running Title:** SARS-CoV-2 Variation in COVID-19 Patients

5

6   Yanqun Wang[1#], Daxi Wang[2,3#], Lu Zhang[4#], Wanying Sun[2,3,5#], Zhaoyong Zhang[1#], Weijun

7   Chen[5,6#], Airu Zhu[1#], Yongbo Huang[1#], Fei Xiao[7], Jinxiu Yao[8], Mian Gan[1], Fang Li[1], Ling luo[1],

8   Xiaofang Huang[1], Yanjun Zhang[1], Sook-san Wong[1], Xinyi Cheng[2,9], Jingkai Ji[2,3,10], Zhihua Ou[2,3],

9   Minfeng Xiao[2,3], Min Li[2,3,5], Jiandong Li[2,3,5], Peidi Ren[2,3], Ziqing Deng[2,3], Huanzi Zhong[2,3],

10   Huanming Yang[2,11], Jian Wang[2,11], Xun Xu[2,12], Tie Song[13], Chris Ka Pun Mok[1,14], Malik Peiris[1,14],

11   Nanshan Zhong[1], Jingxian Zhao[1*], Yimin Li[1*], Junhua Li[2,3,9*], Jincun Zhao[1, 4*]

12

13   [1]State Key Laboratory of Respiratory Disease, National Clinical Research Center for Respiratory

14   Disease, Guangzhou Institute of Respiratory Health, the First Affiliated Hospital of Guangzhou

15   Medical University, Guangzhou, Guangdong, 510120, China.

16   [2]BGI-Shenzhen, Shenzhen, 518083, China.

17   [3]Shenzhen Key Laboratory of Unknown Pathogen Identification, BGI-Shenzhen, Shenzhen,

18   518083, China.

19   [4]Institute of Infectious disease, Guangzhou Eighth People's Hospital of Guangzhou Medical

20   University, Guangzhou, Guangdong, 510060, China.

21   [5]BGI Education Center, University of Chinese Academy of Sciences, Shenzhen, 518083, China.

22   [6]BGI PathoGenesis Pharmaceutical Technology Co., Ltd , BGI-Shenzhen, Shenzhen, 518083,

23   China.

24   [7]Department of Infectious Diseases, Guangdong Provincial Key Laboratory of Biomedical

25   Imaging, Guangdong Provincial Engineering Research Center of Molecular Imaging. The Fifth

26   Affiliated Hospital, Sun Yat-sen University, Zhuhai, Guangdong, 519000, China.

27   [8]Yangjiang People's Hospital, Yangjiang, Guangdong, China.

28   [9]School of Biology and Biological Engineering, South China University of Technology,

29   Guangzhou, China.

30   [10]School of Future Technology, University of Chinese Academy of Sciences, Beijing 101408,

31   China.

32   [11]James D. Watson Institute of Genome Science, Hangzhou, 310008, China.

33   [12]Guangdong Provincial Key Laboratory of Genome Read and Write, BGI-Shenzhen, Shenzhen,

34   518120, China.

35   [13]Guangdong Provincial Center for Disease Control and Prevention, Guangzhou, Guangdong,

36   511430, China.

37   [14]The HKU–Pasteur Research Pole, School of Public Health, Li Ka Shing Faculty of Medicine,

38   the University of Hong Kong, Hong Kong SAR, 19406, China.

39

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

41

42   *Corresponding authors

43   Dr. Jincun Zhao, zhaojincun@gird.cn

44   Dr. Junhua Li, lijunhua@genomics.cn

45   Dr. Yimin Li, dryiminli@vip.163.com

46   Dr. Jingxian Zhao, zhaojingxian@gird.cn

47

48   **Keywords:** SARS-CoV-2, COVID-19, Intra-host, Variation

49 **ABSTRACT**

50 As of middle May 2020, the causative agent of COVID-19, SARS-CoV-2, has infected over 4

51 million people with more than 300 thousand death as official reports[1,2]. The key to

52 understanding the biology and virus-host interactions of SARS-CoV-2 requires the knowledge of

53 mutation and evolution of this virus at both inter- and intra-host levels. However, despite quite a

54 few polymorphic sites identified among SARS-CoV-2 populations, intra-host variant spectra and

55 their evolutionary dynamics remain mostly unknown. Here, using deep sequencing data, we

56 achieved and characterized consensus genomes and intra-host genomic variants from 32 serial

57 samples collected from eight patients with COVID-19. The 32 consensus genomes revealed the

58 coexistence of different genotypes within the same patient. We further identified 40 intra-host

59 single nucleotide variants (iSNVs). Most (30/40) iSNVs presented in single patient, while ten

60 iSNVs were found in at least two patients or identical to consensus variants. Comparison of

61 allele frequencies of the iSNVs revealed genetic divergence between intra-host populations of

62 the respiratory tract (RT) and gastrointestinal tract (GIT), mostly driven by bottleneck events

63 among intra-host transmissions. Nonetheless, we observed a maintained viral genetic diversity

64 within GIT, showing an increased population with accumulated mutations developed in the

65 tissue-specific environments. The iSNVs identified here not only show spatial divergence of

66 intra-host viral populations, but also provide new insights into the complex virus-host

67 interactions.

68

69 **MAIN**

70 From January 25 to February 10 in 2020, we collected a total of 62 serial clinical samples from

71 eight hospitalized patients (GZMU cohort) confirmed with SARS-CoV-2 infection using real-time

72 RT-qPCR (**Table S1**). All patients had direct contacts with confirmed cases during the early

73 stage of the outbreak. Most patients, except P15 and P62, had severe symptoms and received

74 mechanical ventilation in ICU, including the patient P01 who passed away eventually. The

3

75   patient P01 also showed much lower antibody (IgG and IgM) responses (**Table S1**) compared

76   to other patients. We then deep sequenced the 62 clinical samples using metatranscriptomic

77   and/or hybrid capture methods (**Table S1**). The numbers of SARS-CoV-2 reads per million

78   (SARS-CoV-2 RPM) among the metatranscriptomic data correlated well with the corresponding

79   RT-qPCR cycle threshold (Ct), reflecting a robust estimation of viral load (R = 0.71, *P* = 6.7e-11)

80   (**Fig. 1a**). The respiratory tract (RT: Nose, Sputum, Throat) and gastrointestinal tract (GIT: Anus,

81   Feces) samples showed higher SARS-CoV-2 RPMs compared to gastric mucosa and urine

82   samples (**Fig. 1b**). Furthermore, RT and GIT samples from two patients with mild symptoms

83   showed relatively low viral loads among their respective sample types. The data here may

84   reflect an active replication of SARS-CoV-2 in RT and GIT, especially in patients with severe

85   symptoms[3,4].

86   Here, using metatranscriptomic data, we obtained 32 consensus complete genomes

87   from the clinical samples with at least 60-fold sequence coverage (**Table S1 and Table S2**).

88   Comparing the assemblies to the reference sequence (GISAID accession: EPI_ISL_402119)

89   revealed 14 consensus variants (6 synonymous and 8 non-synonymous) located mostly in

90   ORF1ab, S and N genes (**Table S2**). Most of the consensus variants were also detected among

91   public sequences, including the widespread associated variants (C8782T and T28144C)

92   detected in four patients (P10, P13, P14 and P62). The novel consensus variant causes a

93   frameshift at the end of ORF8 in the patient P14, showing the phenotypic plasticity during the

94   evolution of SARS-CoV-2. Evolutionary relationships showed that the consensus SARS-CoV-2

95   genomes of the GZMU cohort belonged to distinct clades, including clades defined by T28144C

96   and A23403G, respectively (**Fig. 1c**). Remarkably, we observed distinct SARS-CoV-2 genomes

97   co-existed in the GIT samples of the patient (P08) with three nucleotide differences (**Fig. 1d and**

98   **Table S2**), suggesting independent replications of different SARS-CoV-2 genotypes within the

99   same host[5].

100   Although plenty of polymorphic sites were identified among SARS-CoV-2 populations,

101   intra-host variant spectra of closely related viral genomes are mostly disguised by the

102   consensus sequences. We firstly examined the reproducibility of our experimental procedures

103   for allele frequency identification. Only a minor difference of alternative allele frequencies (AAFs)

104   was observed among biological replicates of two selected samples (**Fig. S1**), showing that the

105   estimated population composition was marginally affected by independent experimental

106   procedures. To control false discovery rate, we applied a stringent approach to detect iSNVs.

107   The iSNVs were identified from the 32 samples using metatranscriptomic data and then verified

108   using hybrid capture data, which are available for most (27/32) samples (**Table S3 and Table**

109   **S4**).   Overall, we observed 1 to 23 iSNVs in six patients with a cut-off of 5% minor allele

110   frequency (**Fig. 2a and Fig. 2b**). When an iSNV was discovered in one patient, we reduced the

111   cut-off to 2% to detect that iSNV from the rest samples of the same patient (see methods). The

112   AAFs of iSNVs detected from the metagenomic data correlated well with those of the hybrid

113   capture data (Spearman's ρ = 0.99, P < 2.2e-16; **Fig. S2**). Furthermore, the numbers of the

114   observed iSNVs did not correlate with the sequencing coverage (**Fig. S3**), suggesting that the

115   coverage of metatrancriptomic data was sufficient to estimate intra-host variation in most

116   samples.

117   We further analyzed intra-host variation across genes for evidence of natural selection.

118   Overall, the 40 identified iSNV sites (10 synonymous iSNVs and 30 non-synonymous iSNVs)

119   distributed evenly across genomic regions (**Fig. 2c; Table S3**). High proportion of non-

120   synonymous iSNVs suggests that most iSNVs were either under frequent positive selection or

121   insufficient purifying selection. However, we did not observe significant difference in AAFs

122   between non-synonymous and synonymous iSNVs (**Fig. 2d**) and among codon positions (**Fig.**

123   **S4**), indicating a relaxed intra-host selection. It is likely that most of those non-synonymous

124   iSNVs will be removed by purifying selection and/or genetic drift in a longer timescale[6].

125   Nonetheless, the exact functional and evolutionary relevance of the intra-host variants remain to

126   be explored.

127   One central task when estimating intra-host variation is to identify the source of iSNVs.

128   Overall, the distribution of the iSNVs among samples does not correlate well with the consensus

129   SNPs (**Fig. 2a**). Samples carrying the same consensus SNPs generally had different iSNVs,

130   particularly in P01, P10 and P13. Here, we classified the iSNVs into i) rare iSNVs (30/40)

131   detected in a single patient, and ii) common iSNVs (10/40) detected in at least two patients

132   and/or identical to consensus variants. The ten common iSNVs did not show significant higher

133   AAFs than the rare iSNVs (**Fig. 2e**). Notably, the ten common iSNVs include two iSNVs

134   (G11083T and C21711T) exclusively detected in the GIT populations of P01, P08 and P10

135   (**Table S4**). Among the common iSNVs, G11083T is the most widespread consensus variant

136   distributed in multiple lineages of SARS-CoV-2, suggesting that it might derive from recurring

137   mutations on distinct strains rather than the mutation on a single ancestral strain. Interestingly,

138   although G11083T was detected as an intra-host variant in the GIT samples of three patients, it

139   was not detected in the corresponding RT samples, indicating a recurrent mutation of this loci,

140   especially in the GIT population. Interestingly, G11083T locate in a region encoding a predicted

141   T-cell epitope[7], suggesting that recurrent mutation may provide genetic plasticity to better adapt

142   against host defenses.

143   Using Shannon entropy, we observed a significantly higher genetic diversity within the

144   GIT samples than that of RT samples (Wilcoxon rank-sum test, P = 1.4e-05; **Fig. 3a and Table**

145   **S5**), reflecting an increased viral population size within the GIT samples. We further investigated

146   the genetic differentiation between the two places. Notably, no iSNVs was shared between RT

147   and GIT samples from the same patients, suggesting a clear genetic divergence among intra-

148   host viral populations. Here we used L1-norm distance to estimate genetic dissimilarity among

149   samples based on iSNVs and their AAFs and compared that between samples within and

150   among hosts (**Fig. 3b and Table S6**). As expected, genetic distances among samples from the

151 same host were smaller than those among inter-host samples (**Fig. 3b and Table S6**). Within

152 each host, the greatest genetic differentiation was observed among GIT samples and between

153 GIT and RT samples, while the differentiation among RT samples was relatively small. For

154 example, seven iSNVs were shared among the GIT samples of P01, while none of them was

155 observed in RT samples (**Fig. 2a**). It seems that the frequent genetic divergence between GIT

156 and RT populations is mostly driven by bottleneck events during distant intra-host transmissions.

157 However, the exact interaction mechanisms among intra-host populations require further

158 investigation.

159 Previous studies have revealed longitudinal evolution of intra-host populations in some

160 important RNA viruses[8-10]. We firstly compared the detected iSNVs among serial samples. All

161 the iSNVs of early GIT samples also presented in later GIT samples, while all the iSNVs

162 detected in RT samples disappeared in the following samples, suggesting that the viral genetic

163 diversity is better maintained in GIT. We further focused on the allele frequency dynamics of

164 GIT iSNVs of P01 and P08, respectively. Notably, most GIT iSNVs were remarkably stable and

165 showed continuous trends of AAFs across sampling dates. For example, within the GIT

166 population of P01, seven iSNVs showed continuous trends of allele frequency dynamics,

167 including four iSNVs with increased AAFs and two iSNVs with decreased AAFs across the three

168 sampling dates (**Fig. 4a**). Given their similar growth rates but distinct allele frequencies, it is

169 likely that more than two genetically related haplotypes co-existed in within P01. Similar patterns

170 were also observed in the GIT population of P08 (**Fig. 4b**). Notably, the dynamics of intra-host

171 variants changed the consensus allele (>50%) of three genomic loci (3160, 21711 and 28854)

172 of P08. Taken together, the iSNVs and their frequencies suggest that the viral populations in

173 GIT is more stable than those in RT. Nonetheless, in both P01 and P08, we observed increased

174 AAFs of C21711T and G11083T, suggesting that these two variants might be adaptively

175 selected, especially in the GIT. Whether viral adaptation is involved in the intra-host divergence

176 among distant populations warrants further investigation.

7

177   We further phased the proximal iSNVs into local haplotypes using paired-end mapped

178   reads (**Table S7**). Most minor haplotypes had one nucleotide difference from the dominant

179   haplotype of the same sample, suggesting that they might derive from the main strain of the

180   population. Nonetheless, we observed one exception in the GIT population of P01, covering the

181   variable sites of C21707T, C21711T and A21717G (**Fig. S5**). With the cut-off of 1%, one

182   dominant haplotype (T-C-A) and two minor haplotypes (T-T-A and T-T-G) were identified.

183   Despite that minor haplotype (T-T-A) was relatively stable (8%–10%), the proportion of the

184   dominant haplotype (T-C-A) decreased from 89% to 67%, while that of the minor haplotype (T-

185   T-G) increased from 2% to 22%. Based on the dynamics and nucleotide differences among

186   three haplotypes, we hypothesized that the minor haplotype (T-T-G) may derive from the

187   dominant haplotype (T-C-A) via the intermediate haplotype (T-T-A), showing a maintained

188   diversity within GIT population. More importantly, our observation supports that the mutated

189   viruses are capable to replicate and hence, accumulate more variants within GIT of the same

190   host, leading to an increased genetic diversity in the tissue specific environment.

191   Given the observations in patients with influenza[8], stochastic process is the dominant

192   factor driving the intra-host population dynamics, which is especially the case during distant

193   intra-host transmissions. For SARS-CoV-2, one possible intra-host transmission route is from

194   the respiratory tract to the gastrointestinal epithelia. During the intra-host transmission,

195   population composition may change dramatically through random sampling when a novel sub-

196   population was established from a small group of viruses of a larger population[13]. This is

197   supported by the genetic divergence of intra-host variants between RT and GIT populations.

198   The stochastic process between and within intra-host populations seems to also attenuate the

199   efficacy of intra-host purifying selection, as shown by the even distribution of AAFs among

200   synonymous and non-synonymous iSNVs. However, under the traditional genetic population

201   theories, novel founder populations are expected to have a low genetic variation due to the

202   subsampling from the original population. In contrast, viral populations in GIT showed a higher

8

203 genetic diversity than those in RT, reflecting a larger effective viral population size in the GIT.

204 This result is also consistent with the high viral load in GIT (**Fig. 1b**). During the viral replication,

205 both RT and GIT populations showed evidence of generating intra-host variants. Our findings

206 further demonstrated that those novel and/or recurrent intra-host variants are better maintained

207 within GIT, and hence, leading to a higher level of genetic diversity and potentially larger

208 effective population size in GIT. In contrast, the intra-host variants seemed to be less stable in

209 RT, probably associated with a more dramatic genetic drift in RT populations. Differences in

210 other factors, such as host-cell entry, immune responses and microbial communities among

211 tissue specific environments, may further drive the structuring among intra-host population. On

212 the other hand, those differences may also drive viral adaptation, given the two GIT specific

213 non-synonymous iSNVs observed in our study. However, it is still challenging to fully

214 disentangle the influences of stochastic processes and natural selection, considering the

215 frequent confounding genetic signals of these two processes.

216 Intra-host variants were identified in many RNA viruses[8,9,11-14]. Here, using deep

217 sequencing data of serial samples, we revealed the existence of intra-host variation within

218 COVID-19 patients, which is likely to be contributed by novel and/or recurring intra-host

219 mutations. Furthermore, our observation demonstrated a frequent genetic divergence between

220 GIT and RT samples, mostly driven by bottleneck events among intra-host transmissions.

221 Nonetheless, we observed a maintained viral genetic diversity within GIT, reflecting an

222 increased population with accumulated mutations developed in the tissue-specific environments.

223 Exact biological mechanisms of the intra-host population dynamics remain to be explored in

224 future. Our data presented here also reflects the evolutionary capacity of SARS-CoV-2 in

225 developing viral escape and drug resistance during infection. More broadly, these data provide

226 new insights into the complex virus-host interactions.

227

228

9

229   **METHODS**

230   **Patient enrollment and Ethics statement**

231   Eight pneumonia patients, referred as GZMU cohort, were confirmed with SARS-CoV-2

232   infection between January 25 to February 10 in 2020 and hospitalized at the first affiliated

233   hospital of Guangzhou Medical University (six patients), the fifth affiliated hospital of Sun Yat-

234   sen University (one patient), and Yangjiang People's Hospital (one patient). Serial samples

235   were collected, including nasal swabs, throat swabs, sputum, gastric mucosa, urine, plasma,

236   anal swabs and feces. The overall research plan was reviewed and approved by the Ethics

237   Committees of all the three hospitals. All the information regarding patients has been

238   anonymized.

239

240   **Real-time RT-qPCR and Metatranscriptomic sequencing**

241   A total of 62 serial clinical samples collected from eight patients with COVID-19 (**Table S1**) were

242   used for Real-time RT-qPCR. Clinical samples were subjected to RNA extraction using QIAamp

243   Viral RNA Mini Kit (Qiagen, Hilden, Germany). An in-house real-time RT-qPCR was performed

244   by targeting the SARS-CoV-2 RdRp and N gene regions (Zybio Inc.). Human DNA was

245   removed using DNase I and RNA concentration was measured using Qubit RNA HS Assay Kit

246   (Thermo Fisher Scientific, Waltham, MA, USA). DNA-depleted and purified RNA was used to

247   construct double-stranded DNA library using MGIEasy RNA Library preparation reagent set

248   (MGI, Shenzhen, China) following the protocol described in our previous study[15]. High

249   throughput sequencing of the constructed libraries was then carried out on the DNBSEQ-T7

250   platform (MGI, Shenzhen, China) to generate metatranscriptomic data of 100bp long paired-end

251   reads.

252

253   **Hybrid capture-based enrichment and sequencing**

254   For a subset of samples (**Table S1**), genomic content of SARS-CoV-2 was enriched from the

255   double-stranded DNA libraries mentioned above using the 2019-nCoVirus DNA/RNA Capture

256   Panel (BOKE, Jiangsu, China) as described in our previous study[15]. The SARS-CoV-2 content

257   enriched samples were used to construct DNA Nanoballs (DNBs) based libraries, which were

258   then sequenced using the same protocol described above.

259

260   **Data filtering and Genome assembly**

261   Data filtering was performed following the procedures described in previous research[15]. Briefly,

262   for both metatranscriptomic and hybrid capture data, sequence data of each sample were firstly

263   mapped to a pre-defined database comprising representative genomes of coronaviridae. The

264   mapped reads were then subject to the removal of low-quality, duplications, adaptor

265   contaminations and low-complexity to collect high quality coronaviridae-like reads. We also

266   compared the allele frequencies among the two data types when available, samples with

267   conflicted consensus alleles were removed. For the samples with 60-fold of metatranscriptomic

268   data, coronaviridae-like metatranscriptomic reads were used to generate consensus genomes

269   and identify intra-host variants. Full-length consensus genomes were generated from reads

270   mapped to the reference genome (GISAID accession: EPI_ISL_402119) using Pilon (v. 1.23)[16].

271   To prevent false discovery, base positions reporting an alternative allele with the following

272   conditions were masked as N: 1) sequencing coverage less than 5-fold; 2) sequencing

273   coverage less than 10-fold and the proportion of reads with the alternative allele less than 80%.

274   The collected coronaviridae-like reads were also de novo assembled using SPAdes (v. 3.14.0)

275   with default settings[17] with a maximum of 100-fold coverage of read data. Structural variations

276   between the de novo assemblies and consensus genomes, if any, were manually checked and

277   resolved based on read alignments. Nucleotide differences between the consensus sequences

278   and the reference genome were summarized into artificial Variant Call Format (VCF) files, which

279   were annotated using SnpEff (v.2.0.5)[18] with default settings.

280

Phylogenetic analysis

Available consensus sequences of SARS-CoV-2 (**Table S8**) were collected from GISAID database (https://www.gisaid.org/) on 5th April, 2020, after the removal of highly homologous sequences, 122 representative virus strains (**Table S8**) were used to infer evolutionary relationships with the assembled genomes. Within the GZMU cohort, only one genome was selected when more than one identical genome was achieved from the same patient. The assembled SARS-CoV-2 and selected representative genomes were aligned using MAFFT with default settings. A maximum likelihood (ML) tree was inferred using the software IQ-TREE (v.1.6.12)[19], with the best fit nucleotide substitution model selected by ModelFinder from the same software. The inferred ML tree was then visualized using the R package ggtree[20] (v.3.10). Major branches and the defining nucleotide mutations were manually labelled.

292

Summary of public consensus variants

All the consensus sequences of the public strains were aligned with the reference genome (GISAID accession: EPI_ISL_402119) using MAFFT (v.5.3)[21] with default settings. Nucleotide differences between the consensus sequences and the reference genome were summarized into an artificial VCF file, which was then were annotated using SnpEff (v.2.0.5) with default settings. The linkage disequilibrium among the identified consensus variants were estimated using VCFtools (v.0.1.16).

300

Calling of iSNVs

Here, an intra-host single nucleotide variant (iSNV) was defined as the alternative allele co-existed with the reference allele at identical genomic position within the same sample. To minimize false discovery, iSNVs were identified on samples with at least 60-fold mean

12

305   metatranscriptomic sequencing coverage and then verified using hybrid-capture data when

306   available.

307   First, paired-end metatranscriptomic reads were mapped to the reference genome

308   (GISAID accession: EPI_ISL_402119) using BWA aln (v.0.7.16) with default parameters[22].

309   Duplicated reads were marked using Picard MarkDuplicates (v. 2.10.10)

310   (http://broadinstitute.github.io/picard) with default settings. Base composition of each position

311   was summarized from the mapped reads using the software pysamstats (v. 1.1.2)

312   (https://github.com/alimanfoo/pysamstats), and then subject to iSNV site identification with

313   following criteria: 1) base quality larger than 20; 2) sequencing coverage of paired-end mapped

314   reads >= 10; 3) at least five reads support the minor allele 4) minor allele frequency >= 5%; 5)

315   strand bias ratio of reads with the minor allele and reads with major allele less than ten-fold. To

316   minimize false discoveries, sites with more than one alternative allele were filtered out.

317   Biological effects of the identified iSNVs were annotated using the SnpEff (v.2.0.5) with default

318   settings. Alternative allele frequencies (AAFs) at the identified iSNV sites were measured by the

319   proportion of paired-end mapped reads with alternative alleles. When an iSNV was detected in

320   one patient, the detection cut-off of that iSNV was reduced to 2% for the rest samples of the

321   same patient. Only the AAFs more than 2% with at least three reads were kept for the following

322   analyses. All the iSNVs were verified using hybrid capture data when available. At the iSNV

323   sites, the allele with higher frequency was defined as major allele, while one with less frequency

324   was defined as minor allele, regardless whether it is different from the reference allele. A

325   heatmap was generated to visualize the AAFs for all samples using the pheatmap package in R

326   (v.3.6.1). A subset of the identified iSNVs were validated by Sanger sequencing using the

327   protocol described in previous study[15].

328

329   **Statistics of iSNVs**

330   The distribution of iSNVs among genetic components and patients were summarized and

331   visualized using the Python package matplotlib (v.3.2.1). Alternative allele frequencies on all

332   the detected iSNV sites were compared among patients. To avoid oversampling, for the patient

333   with more than sample, only the median AAF among all samples of that patient was used for

334   comparison. Alternative allele frequencies among synonymous and nonsynonymous variants

335   and among codon positions were compared using Wilcoxon rank sum test and visualized

336   through boxplot using the R package *ggplot* (v.3.3.0). For the iSNVs detected in patient P01 and

337   P08, the dynamics of AAFs was visualized across time points using the R package *ggplot*

338   (v.3.3.0).

339

340   **Genetic diversity**

341   Genetic diversity of each sample was estimated using Shannon entropy based on the AAF of

342   each iSNV, assuming that all iSNVs are independent from each other.

$$H(x) = -\sum_{i}^{n} P(i)log_2 P(i)$$

343   where $P(i)$ is the AAF at variable site i.

344

345   **Genetic distance**

346   The genetic distance among samples was estimated using L1-norm distance in a pairwise

347   manner.

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

348   The L1-norm distance (D) between a pair of samples is the sum of distance across all the

349   variable sites (N). For each variable site, the distance is calculated between vectors (p and q for

350   each sample) comprising frequencies of all the four possible nucleotide bases ($n = 4$).

14

351

Haplotype reconstruction

Haplotypes of neighbor iSNV sites were reconstructed using mapped paired end reads.

354

DATA AVAILABILITY

Sequence data used in this study have been deposited in CNGB (https://db.cngb.org/) under

Project accession CNP0001004 and CNP0000997.

358

DISCLOSURE STATEMENT

No conflict of interest was reported by the authors.

361

ACKNOWLEDGEMENTS

376   Guangdong Provincial Key Laboratory of Genome Read and Write and China National

377   GeneBank at Shenzhen for providing sequencing service. We thank the patients who took part

378   in this study.

379

380   **AUTHOR CONTRIBUTIONS**

381   J.Z., J.L., Y.L and J.Z conceived the study, Y.W et al collected clinical specimen and executed

382   the experiments. D.W., W.S., X.C. and J.J. analyzed the data. All the authors participated in

383   discussion and result interpretation. D.W., Y.W., M.P. and J.Z. wrote the manuscript. All authors

384   revised and approved the final version.

385

386   **DISCLOSURE STATEMENT**

387   No conflict of interest was reported by the authors

388

389 **REFERENCES**

390　1　Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat
391　　　origin. *Nature* **579**, 270-273, doi:10.1038/s41586-020-2012-7 (2020).
392　2　Velavan, T. P. & Meyer, C. G. The COVID-19 epidemic. *Trop. Med. Int. Health* **25**, 278-
393　　　280, doi:10.1111/tmi.13383 (2020).
394　3　Wang, W. *et al.* Detection of SARS-CoV-2 in Different Types of Clinical Specimens.
395　　　*Jama*, doi:10.1001/jama.2020.3786 (2020).
396　4　Chen, W. *et al.* Detectable 2019-nCoV viral RNA in blood is a strong indicator for the
397　　　further clinical severity. *Emerging microbes & infections* **9**, 469-473,
398　　　doi:10.1080/22221751.2020.1732837 (2020).
399　5　Wölfel, R. *et al.* Virological assessment of hospitalized patients with COVID-2019.
400　　　*Nature*, doi:10.1038/s41586-020-2196-x (2020).
401　6　Domingo, E., Sheldon, J. & Perales, C. Viral quasispecies evolution. *Microbiol. Mol.*
402　　　*Biol. Rev.* **76**, 159-216, doi:10.1128/mmbr.05023-11 (2012).
403　7　Grifoni, A. *et al.* A Sequence Homology and Bioinformatic Approach Can Predict
404　　　Candidate Targets for Immune Responses to SARS-CoV-2. *Cell host & microbe* **27**, 671-
405　　　680.e672, doi:10.1016/j.chom.2020.03.002 (2020).
406　8　McCrone, J. T., Woods, R. J., Martin, E. T. & Malosh, R. E. Stochastic processes
407　　　constrain the within and between host evolution of influenza virus. **7**,
408　　　doi:10.7554/eLife.35962 (2018).
409　9　Bull, R. A. *et al.* Contribution of intra- and interhost dynamics to norovirus evolution.
410　　　*Journal of virology* **86**, 3219-3229, doi:10.1128/jvi.06712-11 (2012).
411　10　Orton, R. J. & Wright, C. F. Estimating viral bottleneck sizes for FMDV transmission
412　　　within and between hosts and implications for the rate of viral evolution. **10**, 20190066,
413　　　doi:10.1098/rsfs.2019.0066 (2020).
414　11　Gire, S. K. *et al.* Genomic surveillance elucidates Ebola virus origin and transmission
415　　　during the 2014 outbreak. *Science* **345**, 1369-1372, doi:10.1126/science.1259657 (2014).
416　12　Chen, C. *et al.* Phylogenomic analysis unravels evolution of yellow fever virus within
417　　　hosts. **12**, e0006738, doi:10.1371/journal.pntd.0006738 (2018).
418　13　Grubaugh, N. D. *et al.* Genetic Drift during Systemic Arbovirus Infection of Mosquito
419　　　Vectors Leads to Decreased Relative Fitness during Host Switching. *Cell host & microbe*
420　　　**19**, 481-492, doi:10.1016/j.chom.2016.03.002 (2016).
421　14　Andersen, K. G. *et al.* Clinical Sequencing Uncovers Origins and Evolution of Lassa
422　　　Virus. *Cell* **162**, 738-750, doi:10.1016/j.cell.2015.07.020 (2015).
423　15　Xiao, M. *et al.* Multiple approaches for massively parallel sequencing of HCoV-19
424　　　genomes directly from clinical samples. *bioRxiv*, 2020.2003.2016.993584,
425　　　doi:10.1101/2020.03.16.993584 (2020).
426　16　Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant
427　　　detection and genome assembly improvement. *PLoS One* **9**, e112963,
428　　　doi:10.1371/journal.pone.0112963 (2014).
429　17　Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to
430　　　single-cell sequencing. *J. Comput. Biol.* **19**, 455-477, doi:10.1089/cmb.2012.0021 (2012).
431　18　Cingolani, P. *et al.* A program for annotating and predicting the effects of single
432　　　nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster
433　　　strain w1118; iso-2; iso-3. *Fly* **6**, 80-92, doi:10.4161/fly.19695 (2012).

434   19   Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and
435   effective stochastic algorithm for estimating maximum-likelihood phylogenies.
436   *Molecular biology and evolution* **32**, 268-274, doi:10.1093/molbev/msu300 (2015).
437   20   Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree: an r package for
438   visualization and annotation of phylogenetic trees with their covariates and other
439   associated data. *Methods Ecol. Evol.* **8**, 28-36, doi:10.1111/2041-210x.12628 (2017).
440   21   Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy
441   of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511-518, doi:10.1093/nar/gki198
442   (2005).
443   22   Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler
444   transform. *Bioinformatics* **26**, 589-595, doi:10.1093/bioinformatics/btp698 (2010).
445
446

447   **FIGURE LEGEND**

448   **Figure 1. Sequence data from various sample types of patients with COVID-19**

449   **a,** SARS-CoV-2 RPM of meta-transcriptomic data plotted against RT–qPCR cycle threshold (Ct)

450   value for the clinical samples. **b,** Frequency distribution of samples based on SARS-CoV-2

451   reads per million (SARS-CoV-2 RPM). **c,** Maximum likelihood tree of consensus SARS-CoV-2

452   genomes using IQ-TREE (1,000 bootstrap replicates). Colors of dotted tips represent

453   geographic locations of samples. Node labels represent bootstrap values for each branch.

454   Nucleotide mutations that defines the branch were labelled outside the tree. **d,** Distribution of

455   consensus variants (in round circles) detected in GZMU cohort across the SARS-CoV-2

456   genome. Colors represent the biological effect of mutations. Non-synonymous variants are

457   denoted by green, synonymous variants by red, and frameshift by blue. EPI_ISL_402119 was

458   used as the reference sequence.

459

460   **Figure 2. Characteristics of iSNVs.**

461   **a,** Heatmap showing the alternative allele frequencies (AAFs) of intra-host single nucleotide

462   variants (iSNVs) and consensus variants among samples**.** The sample (e.g P01N0129) name

463   indicates patient number P01, sample type (N nosal swab, T throat swab, A anal swab, F feces,

464   S sputum) and collection date (01-27). **b,** The number of detected iSNVs per patient. **c,** Number

465   of iSNV sites among protein-encoding genes. **d,** Box plot showing the distribution of alternative

466   allele frequencies (AAFs) of non-synonymous and synonymous iSNVs. Each dot indicates the

467   median AAF among all the detected iSNVs of samples from same patient. **e,** Box plot showing

468   the distribution of AAFs of common and rare iSNVs. Each dot indicates the median AAF among

469   all the detected iSNVs of samples from same patient.

470

471   **Figure 3. Dynamics of iSNVs detected in SARS-CoV-2 infected patients.**

472   **a,** Box plot showing the distribution of genetic diversity among samples from gastrointestinal

473   tract (GIT) and respiratory tract (RT). **b,** Box plot showing the distribution of L1-norm distances

474   among samples from gastrointestinal tract (GIT) and respiratory tract (RT). Each dot represents

475   the genetic distance between a unique pair.

476

477   **Figure 4. Temporal dynamics of intra-host populations in patient P01 and P08.**

478   **a-b,** Alternative allele frequencies (AAFs) among sampling dates in patient P01 and P08. Days

479   post the first symptom date are shown in bracket. Combined iSNVs are the average frequency

480   of four similar iSNVs (A391T, A2275G, C25163A and T27817G). Colours represent different

481   iSNVs. Underlines represent common iSNVs.

482

483 **SUPPLEMENTARY INFORMATION**

484 **Figure S1. Correlation of estimated alternative allele frequencies between biological**

485 **replicates.**

486 **Figure S2. Correlation of estimated alternative allele frequencies between metagenomic**

487 **and hybrid capture data.**

488 **Figure S3. Correlation between sequencing depth and detected iSNVs.**

489 **Figure S4. Number of iSNV among three codon positions.**

490 **Figure S5. Haplotype frequency of proximal iSNVs within the gastrointestinal tract of the**

491 **patient P01**

492

493 **Table S1. Summary of clinical samples and patients with COVID-19**

494 **Table S2. Genomic information of 32 SARS-CoV-2 samples**

495 **Table S3. List of intra-host single nucleotide variants within 32 SARS-CoV-2 samples**

496 **Table S4. Allele frequency of iSNVs detected from metatranscriptomic and/or hybrid**

497 **capture data**

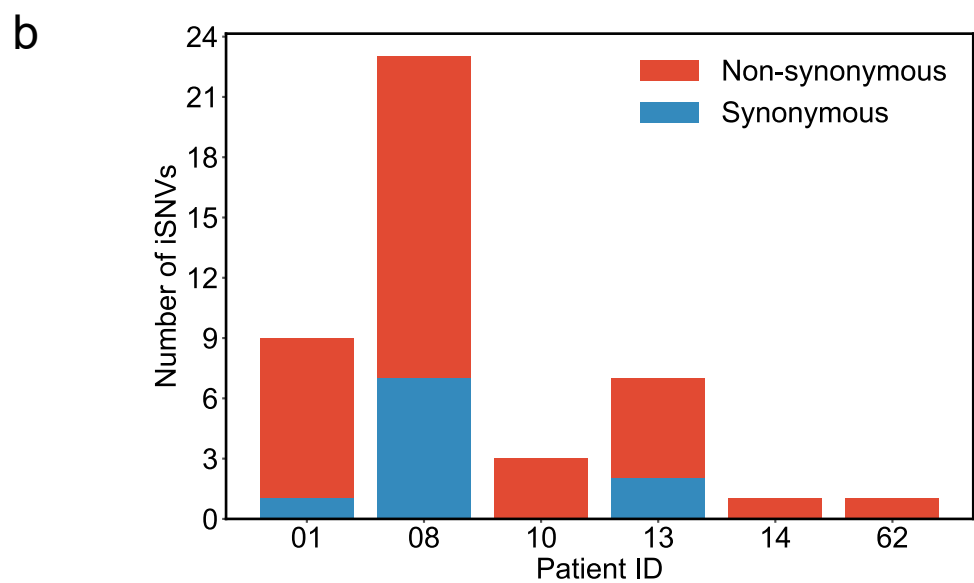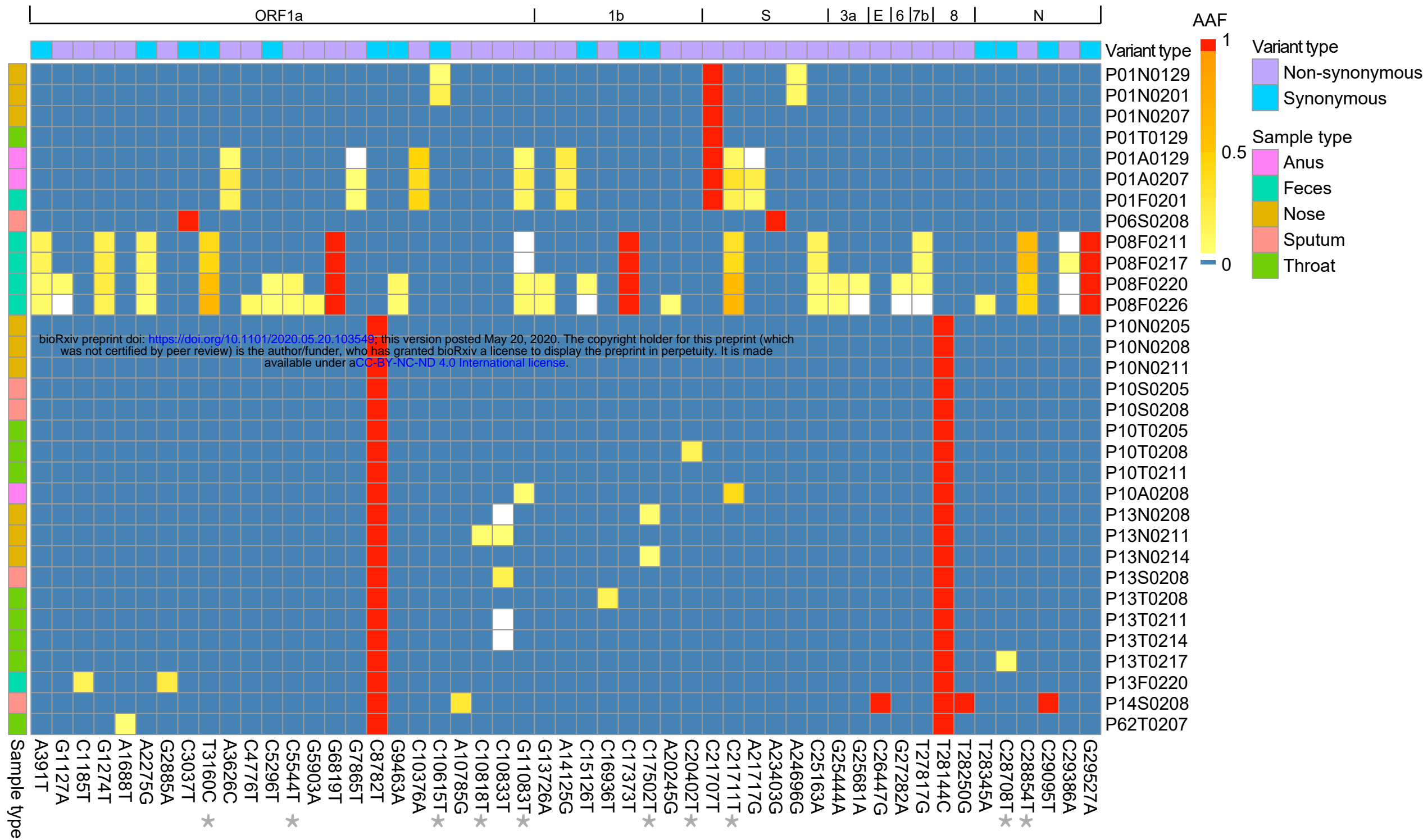498 **Table S5. Genetic diversity of 32 SARS-CoV-2 sampes**

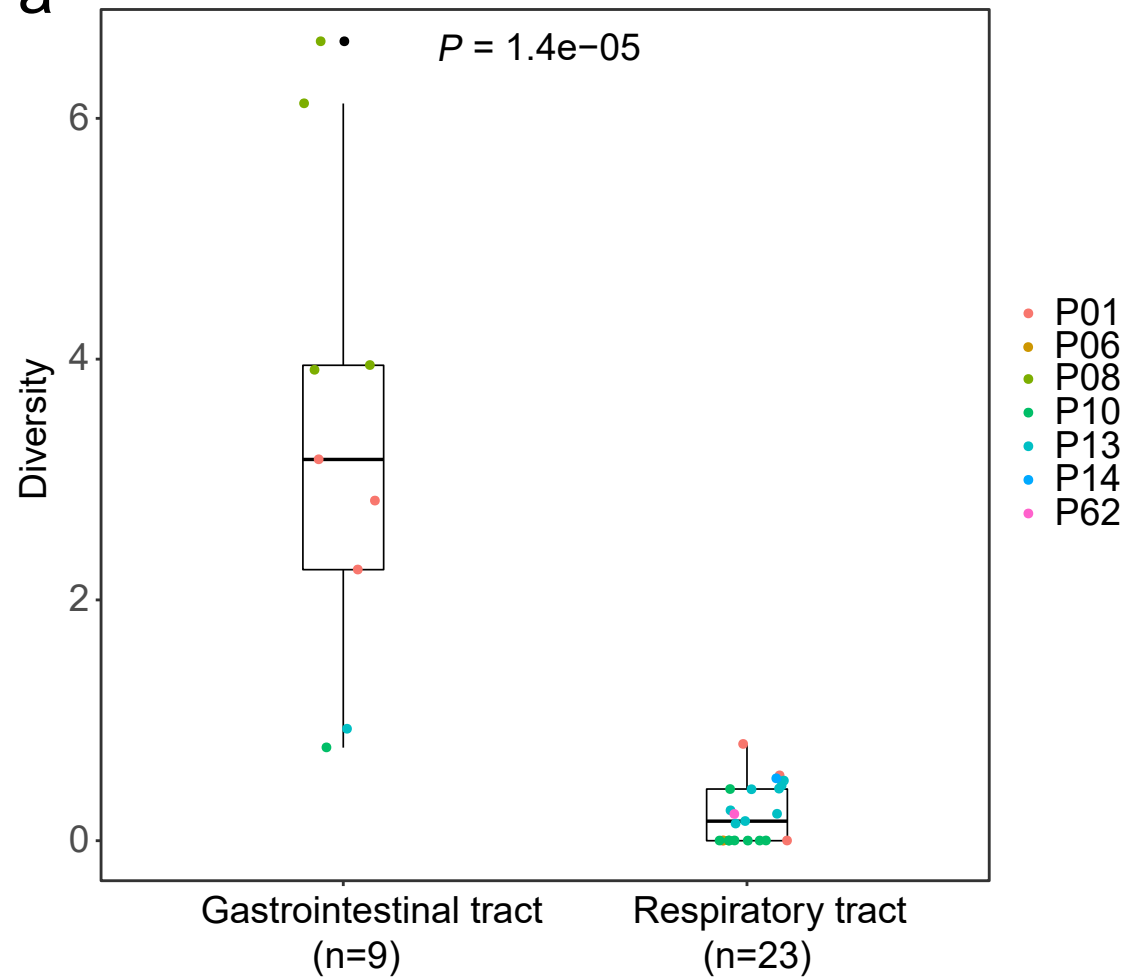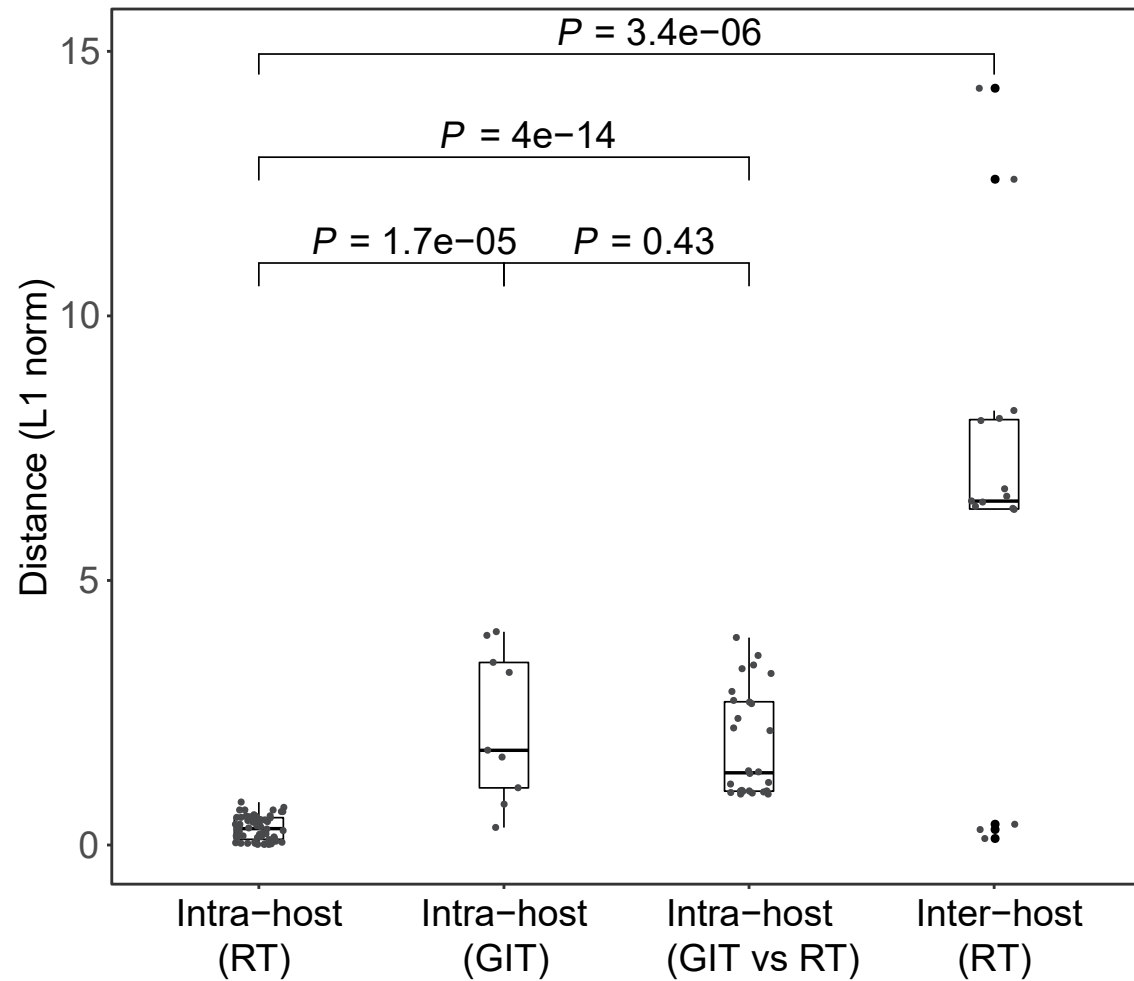499 **Table S6. Genetic distance between paired samples**

500 **Table S7. Frequency of proximal iSNVs using paired-end mapped reads**
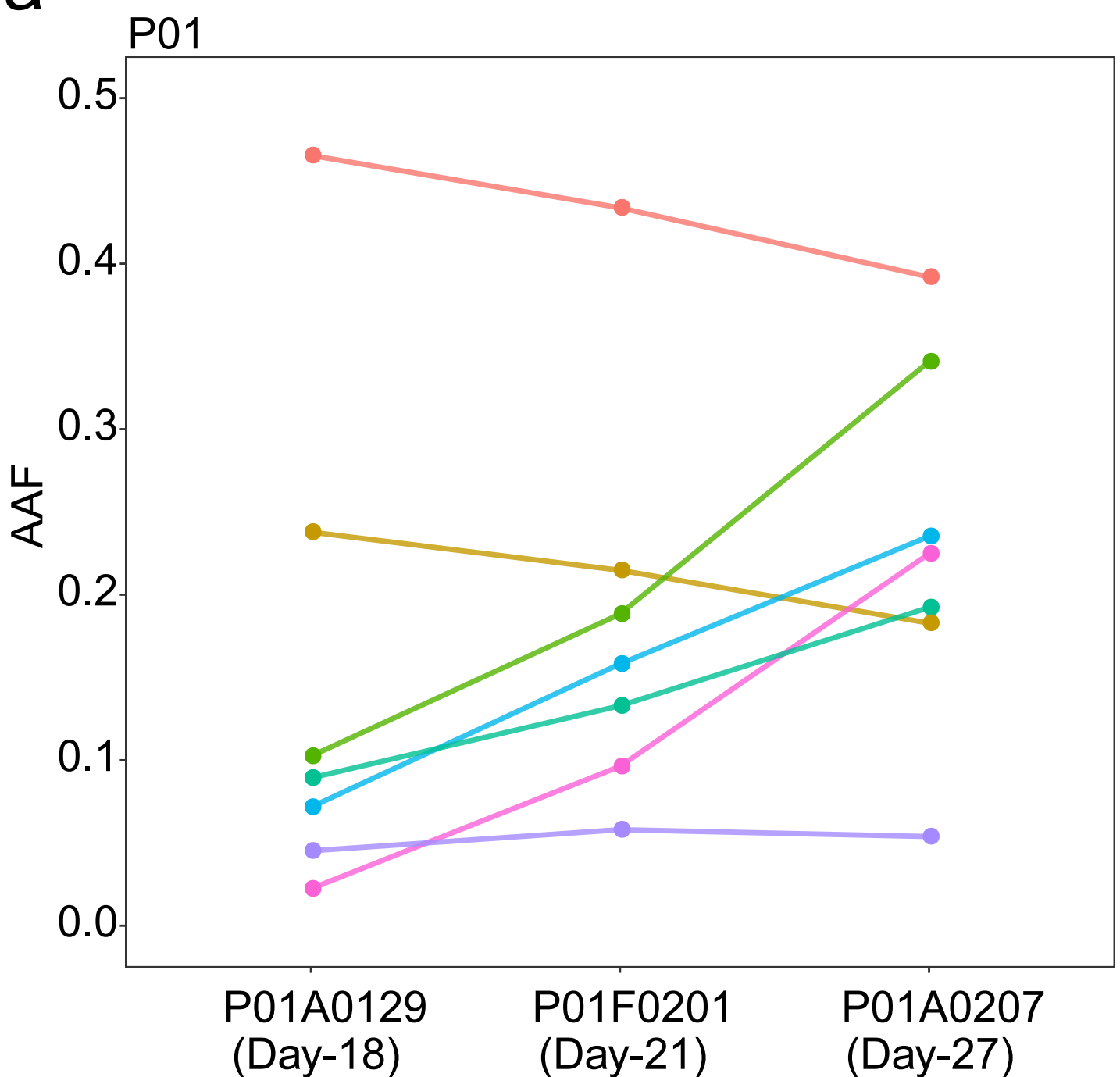
501 **Table S8. List of public genomes used for analysis**

a

P01

b
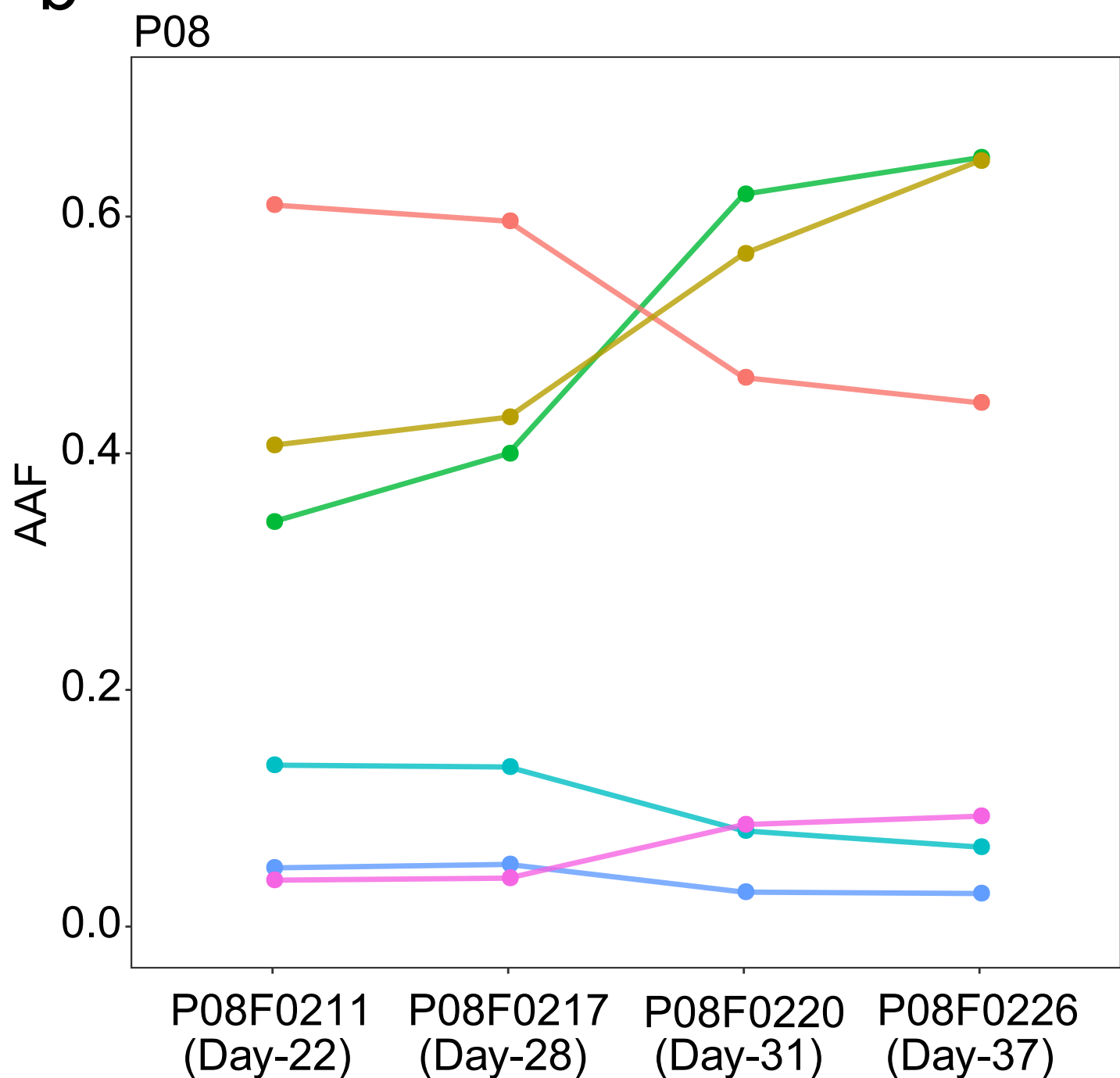
P08
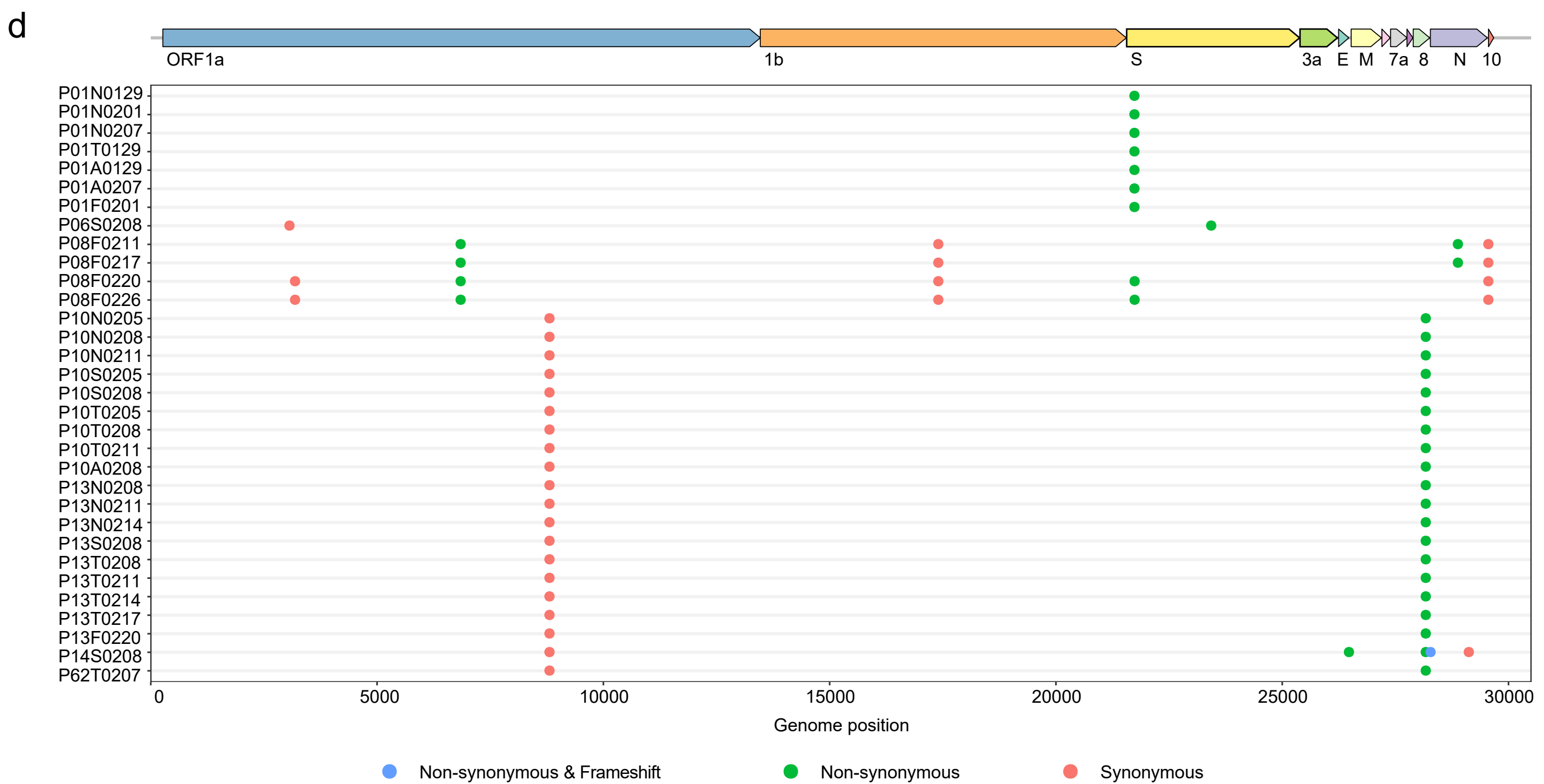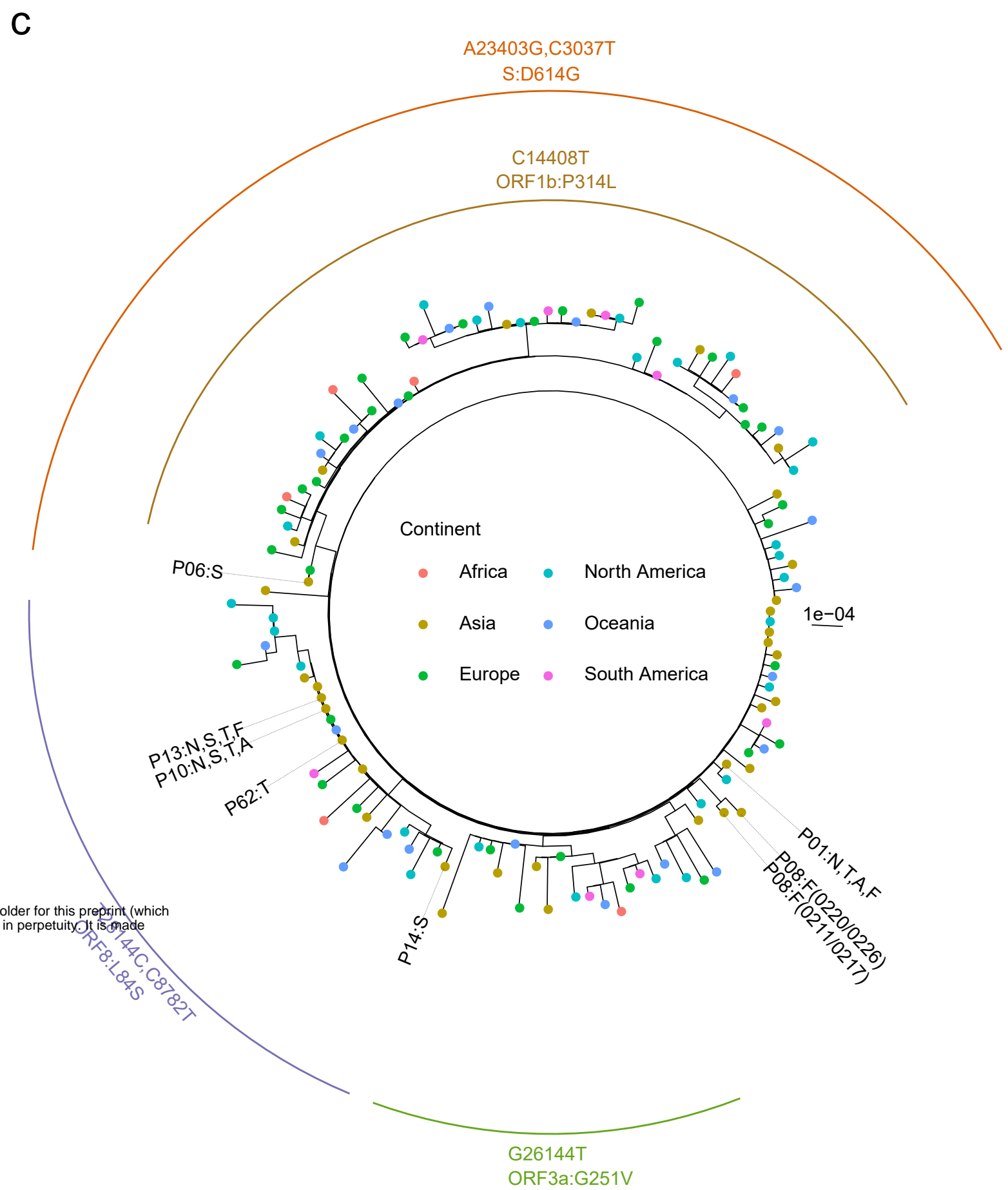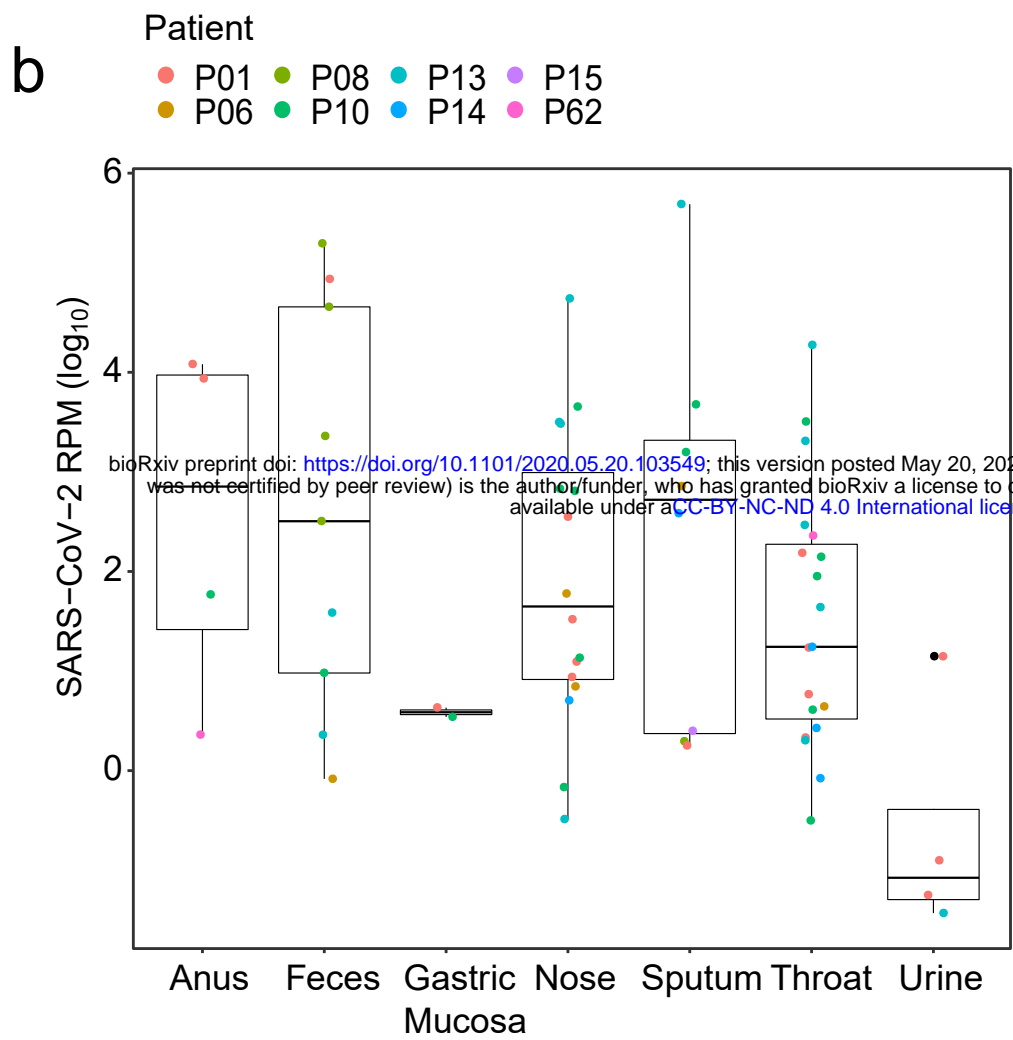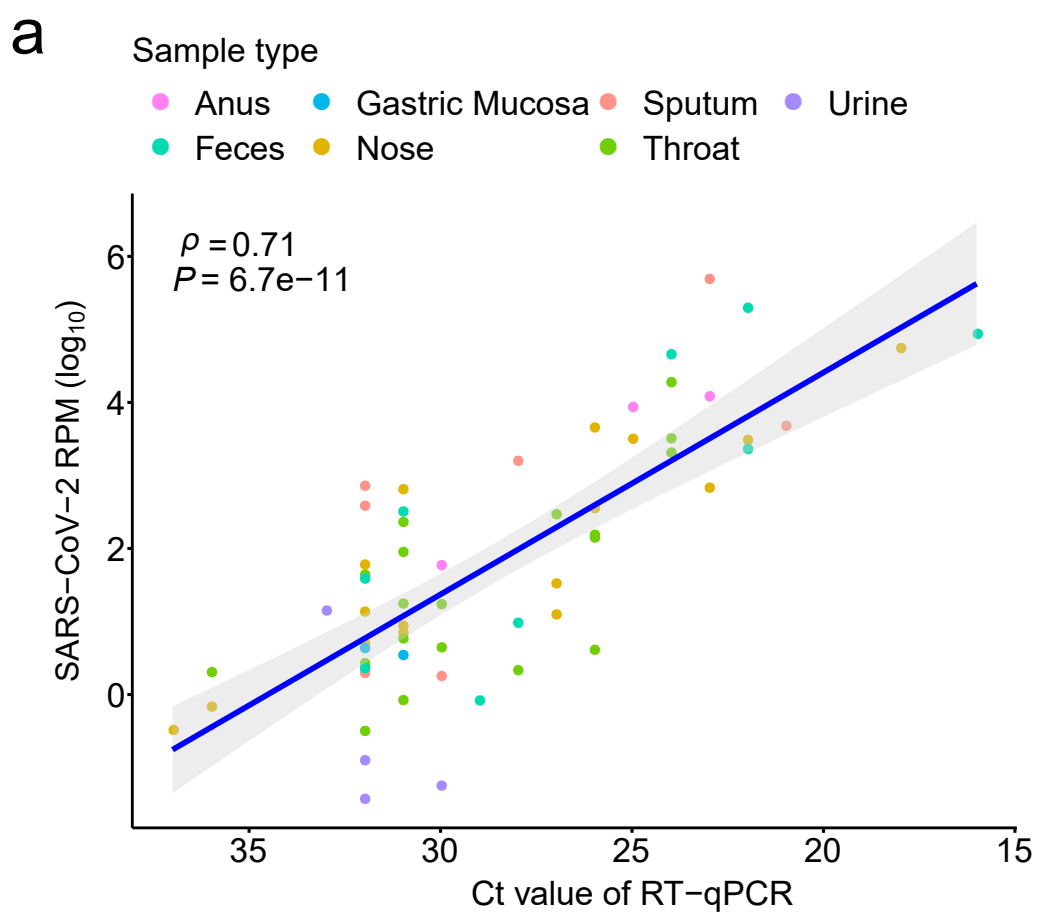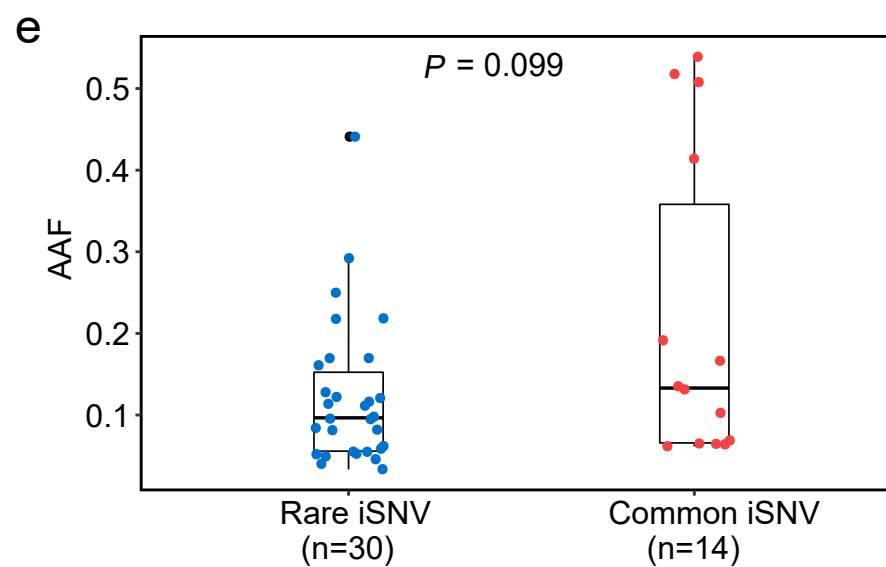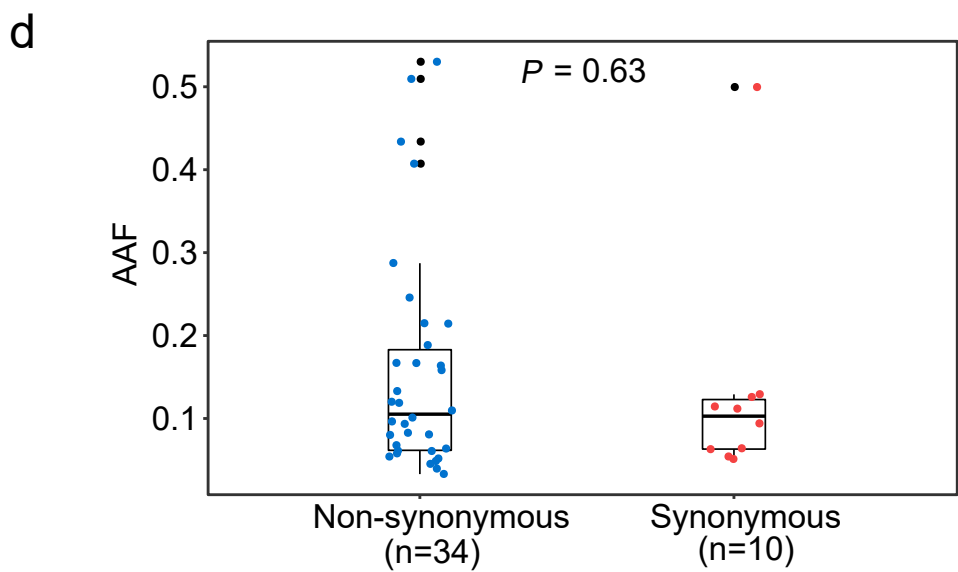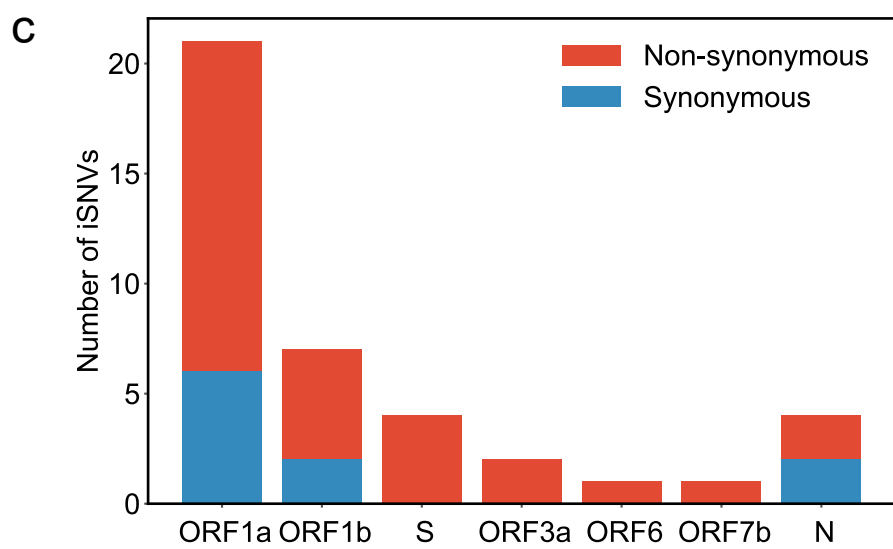
Legend for panel a:
- C10376A
- C21711T
- A3626C
- A21717G
- A14125G
- G11083T
- G7865T

Legend for panel b:
- C28854T
- C21711T
- C29386A
- T3160C
- Combined iSNVs
- G11083T