

1 **Viral infection and transmission in a large, well-traced outbreak caused by the**  
2 **SARS-CoV-2 Delta variant**

3

4 Baisheng Li<sup>1,2#</sup>, Aiping Deng<sup>1,2#</sup>, Kuibiao Li<sup>3#</sup>, Yao Hu<sup>1,2#</sup>, Zhencui Li<sup>1,2#</sup>, Qianling  
5 Xiong<sup>1,2,4</sup>, Zhe Liu<sup>1,2</sup>, Qianfang Guo<sup>1,2</sup>, Lirong Zou<sup>1,2</sup>, Huan Zhang<sup>1,2</sup>, Meng Zhang<sup>1,2</sup>,  
6 Fangzhu Ouyang<sup>1,2</sup>, Juan Su<sup>1,2</sup>, Wenzhe Su<sup>3</sup>, Jing Xu<sup>1,2</sup>, Huifang Lin<sup>1,2,4</sup>, Jing Sun<sup>1,2,4</sup>,  
7 Jinju Peng<sup>1,2,4</sup>, Huiming Jiang<sup>1,2,4</sup>, Pingping Zhou<sup>1,2,4</sup>, Ting Hu<sup>1,2</sup>, Min Luo<sup>1,2</sup>, Yingtao  
8 Zhang<sup>1,2</sup>, Huanying Zheng<sup>1,2</sup>, Jianpeng Xiao<sup>1,2,4</sup>, Tao Liu<sup>1,2,4</sup>, Rongfei Che<sup>1,2</sup>, Hanri  
9 Zeng<sup>1,2</sup>, Zhonghua Zheng<sup>1,2</sup>, Yushi Huang<sup>1,2</sup>, Jianxiang Yu<sup>1,2</sup>, Lina Yi<sup>1,2,4</sup>, Jie Wu<sup>1,2</sup>,  
10 Jingdiao Chen<sup>1,2</sup>, Haojie Zhong<sup>1,2</sup>, Xiaoling Deng<sup>1,2</sup>, Min Kang<sup>1,2</sup>, Oliver G. Pybus<sup>5</sup>,  
11 Matthew Hall<sup>6</sup>, Katrina A. Lythgoe<sup>6</sup>, Yan Li<sup>1,2\*</sup>, Jun Yuan<sup>3\*</sup>, Jianfeng He<sup>1,2\*</sup>, Jing  
12 Lu<sup>1,2,4\*</sup>

13 <sup>1</sup>Guangdong Provincial Center for Disease Control and Prevention, Guangzhou,  
14 Guangdong, China

15 <sup>2</sup>Guangdong Workstation for Emerging Infectious Disease Control and Prevention,  
16 Chinese Academy of Medical Sciences, Guangzhou, Guangdong, China

17 <sup>3</sup>Guangzhou Center for Disease Control and Prevention, Guangzhou, Guangdong,  
18 China

19 <sup>4</sup>Guangdong Provincial Institution of Public Health, Guangzhou, Guangdong, China

20 <sup>5</sup>Department of Zoology, University of Oxford, Oxford OX1 3SZ, UK

21 <sup>6</sup>Big Data Institute, Nuffield Department of Medicine, University of Oxford, Old  
22 Road Campus, Oxford OX3 7LF, UK

23 #Joint first authors.

24 \*Corresponding authors: Yan Li, Jun Yuan, Jianfeng He, Jing Lu

25

26

27 **Summary**

28 We report the first local transmission of the SARS-CoV-2 Delta variant in  
29 mainland China. All 167 infections could be traced back to the first index case. Daily  
30 sequential PCR testing of the quarantined subjects indicated that the viral loads of  
31 Delta infections, when they first become PCR+, were on average ~1000 times greater  
32 compared to A/B lineage infections during initial epidemic wave in China in early  
33 2020, suggesting potentially faster viral replication and greater infectiousness of Delta  
34 during early infection. We performed high-quality sequencing on samples from 126  
35 individuals. Reliable epidemiological data meant that, for 111 transmission events, the  
36 donor and recipient cases were known. The estimated transmission bottleneck size  
37 was 1-3 virions with most minor intra-host single nucleotide variants (iSNVs) failing  
38 to transmit to the recipients. However, transmission heterogeneity of SARS-CoV-2  
39 was also observed. The transmission of minor iSNVs resulted in at least 4 of the 30  
40 substitutions identified in the outbreak, highlighting the contribution of intra-host  
41 variants to population level viral diversity during rapid spread. Disease control  
42 activities, such as the frequency of population testing, quarantine during  
43 pre-symptomatic infection, and level of virus genomic surveillance should be adjusted  
44 in order to account for the increasing prevalence of the Delta variant worldwide.

45

46 During the global spread of the COVID-19, genetic variants of the SARS-CoV-2  
47 virus have emerged. Some variants have increased transmissibility or could exhibit an  
48 increased propensity for escape from host immunity, and therefore pose an increased  
49 risk to global public health<sup>1-3</sup>. An emerging genetic lineage, B.1.617, has gained  
50 global attention and has been dominant in the largest outbreak of COVID-19 in India  
51 since March 2021. One descendent lineage, B.1.617.2, which carries spike protein  
52 mutations L452R, T478K and P681R, accounts for ~28% sequenced cases in India  
53 and has rapidly replaced other lineages to become dominant in multiple regions and  
54 countries (<https://outbreak.info/>)<sup>4</sup>. Lineage B.1.617.2 has been labeled a variant of  
55 concern (VOC) and given the name Delta  
56 (<https://www.who.int/activities/tracking-SARS-CoV-2-variants>). Data on the  
57 virological profile of the Delta VOC is urgently needed.

58 On May 21, 2021 the first local infection of the Delta variant in Guangzhou,  
59 Guangdong, China was identified. As of the early epidemic in China in January 2020<sup>5</sup>,  
60 a suite of comprehensive interventions have been implemented to limit transmission,  
61 including population screening, active contact tracing, and centralized  
62 quarantine/isolation. However, in contrast to the limited level of onward transmission  
63 observed in Guangdong in early 2020<sup>5</sup>, successive generations of virus transmission  
64 were observed in the 2021 outbreak of the Delta variant in the region. Here, we  
65 investigated epidemiological and genetic data from the well-traced outbreak in  
66 Guangdong in order to characterize the virological and transmission profiles of the  
67 Delta variant. We discuss how intervention strategies may need to be adjusted to cope  
68 with the virological properties of this emerging variant.

69

70 **Results**

71 A total of 167 local infections were identified during the outbreak, starting with  
72 the first index case identified on May 21, 2021 and ending with the last case reported  
73 on June 18, 2021 (Figure 1a). All cases could be epidemiologically or genetically  
74 traced back to the first index case (Figure 1b). One notable epidemiologic feature of  
75 the Delta variant is a shorter serial interval compared with to infection with early  
76 Wuhan-like strains or other VOC variants<sup>6-8</sup>. However, critical parameters before the  
77 illness onset remain poorly known, including when the viruses can be first detected in  
78 a subject after exposure, and how infectious infected individuals are.

79 We investigated the data from the quarantined subjects in this outbreak and  
80 compared it to data from the early 2020 epidemic caused by A/B genetic clade (Pango  
81 nomenclature<sup>9</sup>) strains. The centrally-quarantined subjects were the close contacts of  
82 confirmed cases. Once a new infection was identified, his/her close contacts were  
83 immediately traced, centrally isolated, and underwent daily PCR testing. The dataset  
84 from quarantined subjects allowed us to determine the time interval in the infected  
85 subjects between exposure and when viral loads were first detectable by PCR. The  
86 exact exposure time for the intra-family transmissions was difficult to pinpoint, hence  
87 we removed intra-family transmission pairs from our time interval analysis. Our  
88 results revealed that the time interval from exposure to the first PCR+ test in the  
89 quarantined population was 6.00 days (IQR 5.00-8.00) during the 2020 epidemic  
90 (n=29; peak at 5.61 days) and 4.00 days (IQR 3.00-5.00) in the 2021 Delta epidemic  
91 (n=34; peak at 3.71 days; Figure 1c).

92 We next evaluated viral load measurements at the time when SARS-CoV-2 was  
93 first detected by PCR in each subject. The relative viral loads of cases infected with  
94 the Delta variant (n=62, Ct =24.00 for the *ORF1ab* gene, IQR 19.00~29.00) were  
95 1260 times higher than those for the 2020 infections with clade 19A/19B viruses  
96 (n=63, Ct = 34.31 for *ORF1ab* gene, IQR 31.00~36.00) on the day when viruses were  
97 first detected (Figure 1d). We hypothesized a higher within-host growth rate of the

98 Delta variant, which led to the higher observed viral loads once viral nucleotides  
99 exceeded the PCR detection threshold (Figure 1e). Similar to results reported by  
100 Roman *et.al.*, we found that samples with Ct > 30 ( $<6 \times 10^5$  copies/mL viruses) did not  
101 yield an infectious isolate in-vitro. For the Delta variant infections, 80.65% of  
102 samples contained  $>6 \times 10^5$  copies/mL in oropharyngeal swabs when the viruses were  
103 first detected, compared to 19.05% of samples from clade 19A/19B infections. These  
104 data indicate that the Delta variant could be more infectious during the early stage of  
105 the infection (Figure 1e).

106 Individuals undergo a latent period after infection, during which viral titers are  
107 too low to be detected. As viral proliferation continues within host, the viral load will  
108 eventually reach detectable levels and the individual will become infectious. Knowing  
109 when an infected person can transmit is essential for designing intervention strategies  
110 that break chains of transmission. However, infectiousness is difficult to measure  
111 from clinical investigations since  $>50\%$  of transmission occurs during the  
112 pre-symptomatic phase<sup>10</sup>. Our investigation of quarantined subjects suggests that, for  
113 the Delta variant, the time window from exposure to the detection of virus was  $\sim 3.7$   
114 days, and infections presented a higher transmission risk when the virus was first  
115 detected compared to earlier circulating viral lineages. Consequently, the provincial  
116 government required people leaving Guangzhou city from airports, train stations and  
117 shuttle bus stations to show proof of a negative COVID-19 test within 72 hours on  
118 June 6 and this was shortened to 48 hours on June 7. In contrast, the comparable time  
119 window implemented in the 2020 epidemic was seven days.

## 120 **Transmission bottleneck and the association between minor iSNVs transmissions** 121 **and viral population diversity**

122 The non-pharmaceutical interventions in Guangdong mainly focus on  
123 epidemiological investigation, contact tracing and mass testing. Approximately 30

124 million PCR tests were performed between May 26, 2021 and June 8, 2021. The  
125 intense testing and screening of high-risk populations makes cryptic transmissions  
126 unlikely. Nearly all the infections we identified could be connected epidemiologically,  
127 either through evidence of direct contact, or indirectly (staying in or visiting the same  
128 area) (Figure 1b). In addition, all sequences could be genetically traced back to the  
129 index case. This provided a unique opportunity for us to characterize virus  
130 transmission dynamics at a finer scale, particularly the extent to which virus genetic  
131 diversity is transmitted among hosts. Whole-genome deep sequencing was performed  
132 on all identified infections, and 126 high-quality viral genomes (coverage>95%) were  
133 obtained, comprising 75% of identified infections in the outbreak (Figure 1a).

134 Phylogenetic analysis was performed by combining the virus genomes we  
135 obtained from the Delta outbreak with genomes from 346 imported cases; the latter  
136 represent travelers to Guangdong during March 2020 to June 2021 who arrived from  
137 66 different source countries. We also included a set of reference sequences,  
138 comprising 50 genomes randomly selected from each of 13 defined NextStrain clades  
139 (<https://nextstrain.org/>) and the notified VOCs (Alpha, Beta, Gamma, Delta). The  
140 viral lineage distribution of the imported cases was approximately representative of  
141 the SARS-CoV-2 genetic lineages that were circulating at that time at the global scale.  
142 These importations pose a challenge for disease control and prevention in Guangdong,  
143 China (Figure 2a).

144 Viral phylogenies of the Guangzhou outbreak were inferred using the assembled  
145 consensus sequence of each sample, which was generated by choosing the  
146 majority-frequency nucleotide (>50%) at each position. All Guangzhou outbreak  
147 sequences segregated into a single cluster (Figure 2a). Compared with the index case  
148 (5137) of the outbreak, 30 substitutions were identified among 125 cases during the  
149 26-days long outbreak (Figure 2b). The most genetically-divergent outbreak sequence  
150 contained four nucleotide differences from the index case sample. To understand how

151 these variants emerged, grew and finally fixed during the epidemic (and during the  
152 SARS-CoV-2 pandemic more generally), we estimated within-host virus diversity for  
153 each sample by mapping polymorphic sites against the consensus genome of the  
154 index case (XG5137\_GZ\_2021/5/21), thereby generating a list of intra-host  
155 single-nucleotide variants (iSNVs). Minor iSNVs were called by setting 3% as the  
156 threshold for minor allele frequency, in order to exclude potential PCR and  
157 sequencing errors <sup>11-13</sup>. For 126 high-quality sequences, most samples harbored 3  
158 iSNVs (median) which is consistent with other reported levels (Supplemental Figure)  
159 <sup>11,12</sup>.

160 We calculated the transmission bottleneck size among  
161 epidemiologically-confirmed transmission pairs. Contact tracing and epidemiological  
162 investigation enabled us assign 111 donor-recipient transmission pairs with a high  
163 degree of confidence. Of these, the donor had one or more iSNVs above the variant  
164 calling threshold of 3% in 74 transmission pairs (Table S1), enabling estimation of the  
165 transmission bottleneck size,  $N_b$ , using the beta-binomial method <sup>14</sup>. The maximum  
166 likelihood estimate for  $N_b$  was one for 65 out of these 74 transmission pairs, and two  
167 or three for the remaining 9 transmission pairs (Figure 2C). Uncertainty in the  $N_b$   
168 estimate was large for some transmission pairs, with the 95% confidence interval  
169 ranging from 1 to ~500 or more, suggesting for some pairs the sequencing data was  
170 not sufficiently informative. Our data suggest the transmission bottleneck of  
171 SARS-CoV-2 is very narrow in general, consistent with the previous household  
172 transmission studies <sup>11,15</sup>. The transmission bottleneck size influences the extent to  
173 which within-host diversity contributes to viral diversity at the population scale. The  
174 stringent transmission bottleneck of SARS-CoV-2 suggests the substitutions we  
175 observed in Guangdong outbreak (and SARS-CoV-2 pandemic more generally)  
176 largely resulted from de-novo mutations appearing within individuals.

177 Although the transmission bottleneck of SARS-CoV-2 is narrow in general, it

178 may be not constant and could be impacted by both viral and host factors. To  
179 investigate the contribution of the transmission of minor iSNV to population-level  
180 diversity, we identified the sequences with minor iSNVs and the sequences in which  
181 the derived nucleotide state was fixed. Notably, sequences exhibited minor intra-host  
182 single nucleotide variants (iSNVs) at 10 of the 30 variant sites (positions that varied  
183 from the sequence of the first index case) (Figure 2b). The direct (transmission pair 61,  
184 1, 2, 3) and indirect (from case 6190 to 6486) epidemiological links were observed  
185 between the hosts with the minor iSNVs and their potential recipients with these  
186 iSNVs fixed (Figure 2C). Therefore, at least three fixed substitutions in this outbreak  
187 could be traced to the direct transmission of minor iSNVs, and one substitution was  
188 from a suspicious transmission chain. It is also noteworthy that the transmission pairs  
189 with 5137 as the donor had a relatively higher estimated  $N_b$ , suggesting heterogeneity  
190 in iSNV transmission (Figure 2c). The differences in bottleneck size are possibly due  
191 to the different transmission route or exposure doses, as has been observed for  
192 influenza<sup>16</sup>. The case 5137 presented a high viral load (Ct value of 17.6, approximate  
193  $2 \times 10^9$  copies/mL in oropharyngeal swabs) 2 days after their direct contact with the  
194 cases 5645 and 5571. The high viral loads, direct contacts and relatively high  
195 frequency of the iSNVs (4% for T21673C and 47% for C27086T) may have enabled  
196 the successful transmission of iSNVs to the recipients (Figure 2C). Taken together,  
197 our observations suggest that the transmission bottleneck of SARS-CoV-2 is stringent  
198 in general, with most donor iSNVs not found in the recipients. However, transmission  
199 of minor iSNVs, with their fixation in the recipient host, resulted in at least some of  
200 the substitutions that accumulated during the outbreak.

201 In this study, we characterized a large transmission chain that originated from the  
202 first local infection of the SARS-CoV-2 Delta variant in mainland China. We find  
203 evidence for a potentially higher viral replication rate of the Delta variant, as viral  
204 loads in Delta infections are ~1000 times higher than those for clade 19A/19B  
205 infections on the day of the first PCR+ test. This suggests that infectiousness of Delta



206 variant during the early stage of infection is likely to be higher. Consequently, the  
207 frequency of population screening should be optimized<sup>17</sup>. If Delta infections are  
208 indeed more infectious during the pre-symptomatic phase, then timely quarantine  
209 (before clinical onset or PCR screening) for suspected cases or for close contacts  
210 becomes more important. Although the transmission bottleneck of SARS-CoV-2 is  
211 narrow in general, heterogenicity of minor iSNV transmission is observed and  
212 explains some of the fixed substitutions observed in the virus population during the  
213 outbreak. In some settings, the advantageous iSNVs that are present at a low  
214 frequency could rise and become fixed in the one generation of transmission, and  
215 further predominance in the virus population if the epidemic is not well contained.

216

## 217 **Methods**

### 218 *Ethics*

219 This study was approved by the institutional ethics committee of the Guangdong  
220 Provincial Center for Disease Control and Prevention (GDCDC). Written consent was  
221 obtained from patients or their guardian(s) when samples were collected. Patients  
222 were informed about the surveillance before providing written consent, and data  
223 directly related to disease control were collected and anonymized for analysis.

### 224 *Sample collection, clinical surveillance and epidemiological data*

225 Since the first local SARS-CoV-2 infection reported on May 21 in the capital city of  
226 Guangdong, the enhanced surveillance was performed by Guangdong CDC and local  
227 CDCs to detect suspected infections. Epidemiological investigations had been done  
228 on all confirmed cases. Population screening were performed by third-party detection  
229 institutions. Once virus positive samples were confirmed by local CDCs or other  
230 institutions, the samples were required to send to Guangdong CDC in 24 hours. To  
231 make the results comparable, in Guangdong CDC, the real-time reverse transcription

232 PCR (RT-PCR) were performed by using the same commercial kit (DaAn Gene) and  
233 RT-PCR machine (CFX96) as the previous studies<sup>5,18</sup>. The exposure history for  
234 positive cases and their close contacts were obtained through an interview, public  
235 video monitoring systems and cell phone apps, *etc.* Information regarding the  
236 demographic and geographic distribution of SARS-CoV-2 cases can be found at the  
237 website of Health Commission of Guangdong Province  
238 (<http://wsjkw.gd.gov.cn/xxgzbd/fk/yqtb/>). The surge population screening test ensure  
239 all possible infections were identified and 111 donor-recipient transmission pairs were  
240 assigned with very high confidence. All transmission pairs met the following rules: 1.  
241 The recipient was the close contract of the donor and had a clear and direct  
242 epidemiological link to the donor; 2. The recipient did not have any contacts with  
243 other identified cases.

#### 244 ***Virus amplification and sequencing***

245 Total RNAs were extracted from oropharyngeal swab samples by using QIAamp Viral  
246 RNA Mini Kit (Qiagen, Cat. No. 52904). Virus genomes were generated by two  
247 different approaches, (i) using commercial sequencing kit of BGI (ATOPlex  
248 1000021625) and sequencing on the BGI MGISEQ-2000 (n=25), and (ii) using  
249 version 3 of the ARTIC COVID-19 multiplex PCR primers  
250 (<https://artic.network/ncov-2019>) for genome amplification, followed by library  
251 construction with Illumina Nextera XT DNA Library Preparation Kit and sequencing  
252 with PE150 (n=63) or SE100 (n=38) on Illumina Miniseq. We report only  
253 high-quality genome sequences for which we were able to generate >95% genome  
254 coverage.

#### 255 ***Sequence analysis***

256 The bioinformatics pipeline for BGI platform  
257 ([https://github.com/MGI-tech-bioinformatics/SARS-CoV-2\\_Multi-PCR\\_v1.0](https://github.com/MGI-tech-bioinformatics/SARS-CoV-2_Multi-PCR_v1.0)) was  
258 used to generate consensus sequences and call single nucleotide variants relative to

259 the reference sequence. For sequence data from Miniseq, the raw data were first  
260 quality controlled (QC) using fastp<sup>19</sup> to trim artificial sequences (adapters), to cut  
261 low-quality bases (quality scores  $< 20$ ). PCR primers were trimmed by using  
262 cutadapt version 3.1<sup>20</sup> or other published method<sup>21</sup>. Since all infections could be  
263 traced back to the first index case, the cleaned reads of each sample were mapped  
264 against the genome of the first index case (5137\_GZ\_2021/5/21) using BWA 0.7.17<sup>22</sup>.  
265 The consensus sequences were determined with iVar 1.2.1<sup>23</sup>, taking the most common  
266 base as the consensus (allele frequency  $> 50\%$ ). An N was placed at positions along  
267 the reference with the sequencing depth fewer  $\leq 10$ . The surge population screening  
268 test ensure all possible infections were identified and through the contact tracing the  
269 donor-recipient transmission pairs could be assigned with high confidence. To  
270 characterize the viral transmission in these pairs, we identified iSNVs relative to the  
271 reference genome (XG5137\_GZ\_2021/5/21) for each sequence with iVar 1.2.1 using  
272 the following parameters: alternated frequency at a SNV site  $\geq 3\%$ ; total sequencing  
273 depth at SNV site  $\geq 100$ ; sequencing depth for the variant allele  $\geq 10$ ; iVar  
274 PASS=TRUE. We exclude the head and tail sequences of viral genome  
275 (corresponding to the positions 1 to 100 and 29803 to 29903 in Wuhan-Hu-1  
276 reference genome) due to the lower sequencing coverage for most samples in the  
277 analysis and the 7 “highly shared” iSNV sites (1959, 4091, 21987, 24404, 28448,  
278 28389, 29681) possibly due to the contamination of the primer sequences or mapping  
279 errors<sup>11</sup>. To infer the iSNVs transmission in 74 donor-recipient pairs, all sites with  $\geq 3\%$   
280 minor allele frequency in the assumed donor were used in the analysis. In the  
281 recipient, all reads at these sites were considered, with a variant calling threshold of 3%  
282 using the beta-binomial method of Sobel Leonard *et.al*<sup>14</sup>. The nextstrain pipeline<sup>24</sup>  
283 was used to analyze and visualize the genetic distribution of SARS-CoV-2 infections  
284 and its dynamic change in Guangdong between January 2020 and June 2021.  
285 Maximum likelihood (ML) tree was estimated with phyml<sup>25</sup> using the HKY+Q4  
286 substitution model with gamma-distributed rate variation<sup>26</sup>. The branch length was

287 recalculated as the number of mutations to the reference sequence of the first index  
288 case. The tree was visualized with R package of ggtree<sup>27</sup>.

### 289 ***Data availability***

290 All sequencing reads after primer trimming and mapped to the reference sequence  
291 (the sequences of the first index case, XG5137\_GZ\_2021/5/21) have been submitted  
292 to the National Genomics Data Center (<https://bigd.big.ac.cn/>) with submission  
293 number CRA004571. The generated consensus sequences were submitted with  
294 accession number GWHBDIM01000000 – GWHBDNH01000000.

### 295 **Code availability**

296 The pipeline for sequencing data analysis was deposit in  
297 <https://github.com/Jinglu1982/Delta-variant-outbreak-in-GZ>. Code to implement the  
298 beta-binomial method is publicly available<sup>14</sup>.

299

### 300 **Acknowledgements**

301 We gratefully acknowledge the efforts of China national CDCs, Guangdong local  
302 CDCs, hospitals, and the third-party detection institutions in epidemiological  
303 investigations, sample collection, and detection. This work was supported by grants  
304 from Science and Technology Planning Project of Guangdong (2018B020207006),  
305 the Key Research and Development Program of Guangdong Province  
306 (2019B111103001), and Guangdong Workstation for Emerging infectious Disease  
307 Control and Prevention, Chinese Academy of Medical Sciences (2020-PT330-004).

308

### 309 **Competing interests**

310 The views expressed in this article are those of the authors and not necessarily those  
311 of the Guangdong Provincial Center for Diseases Control and Prevention, or the  
312 Guangdong Provincial Institute of Public Health.

313

314

315

316 **References**

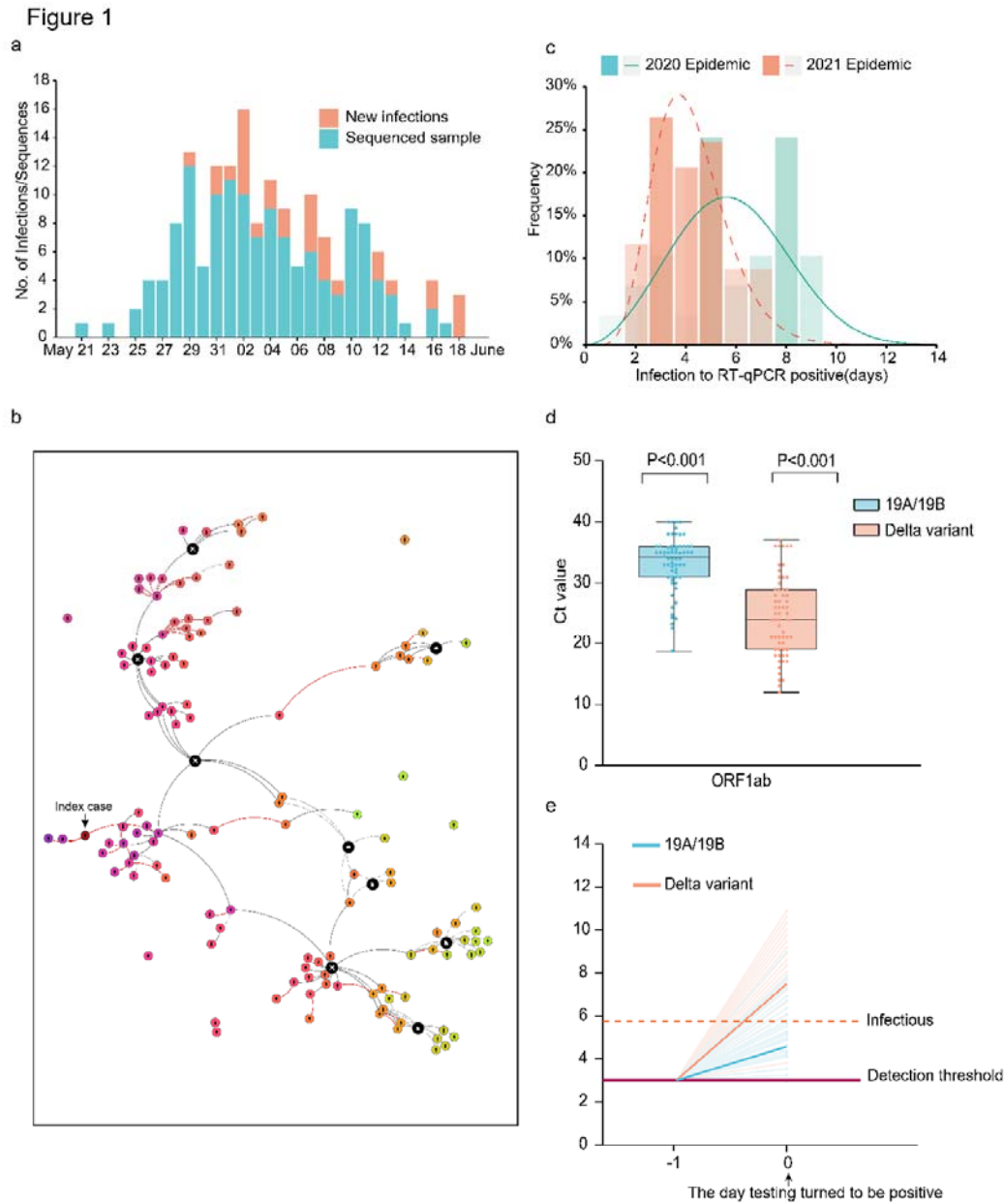
- 317 1. Faria, N. R. *et al.* Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in  
318 Manaus, Brazil. *Science* (2021) doi:10.1126/science.abh2644.
- 319 2. Davies, N. G. *et al.* Estimated transmissibility and impact of SARS-CoV-2  
320 lineage B.1.1.7 in England. *Science* **372**, (2021).
- 321 3. Volz, E. *et al.* Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in  
322 England. *Nature* **593**, 266–269 (2021).
- 323 4. outbreak.info.
- 324 5. Lu, J. *et al.* Genomic Epidemiology of SARS-CoV-2 in Guangdong Province,  
325 China. *Cell* S0092867420304864 (2020) doi:10.1016/j.cell.2020.04.023.
- 326 6. Vöhringer, H. S. *et al.* Genomic reconstruction of the SARS-CoV-2 epidemic  
327 across England from September 2020 to May 2021.  
328 <http://medrxiv.org/lookup/doi/10.1101/2021.05.22.21257633> (2021)  
329 doi:10.1101/2021.05.22.21257633.
- 330 7. Pung, R., Mak, T. M., CMMID COVID-19 working group, Kucharski, A. J. &  
331 Lee, V. J. *Serial intervals observed in SARS-CoV-2 B.1.617.2 variant cases.*  
332 <http://medrxiv.org/lookup/doi/10.1101/2021.06.04.21258205> (2021)  
333 doi:10.1101/2021.06.04.21258205.
- 334 8. Zhang, M. *et al.* Transmission Dynamics of an Outbreak of the COVID-19 Delta  
335 Variant B.1.617.2 — Guangdong Province, China, May–June 2021. *China CDC Wkly.*  
336 **3**, 584–586 (2021).
- 337 9. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages  
338 to assist genomic epidemiology. *Nat. Microbiol.* **5**, 1403–1407 (2020).
- 339 10. Hu, S. *et al.* Infectivity, susceptibility, and risk factors associated with  
340 SARS-CoV-2 transmission under intensive contact tracing in Hunan, China. *Nat.*  
341 *Commun.* **12**, 1–11 (2021).
- 342 11. Lythgoe, K. A. *et al.* SARS-CoV-2 within-host diversity and transmission.  
343 *Science* **372**, eabg0821 (2021).
- 344 12. Valesano, A. L. *et al.* Temporal dynamics of SARS-CoV-2 mutation  
345 accumulation within and across infected hosts. *PLOS Pathog.* **17**, e1009499 (2021).
- 346 13. Poon, L. L. M. *et al.* Quantifying influenza virus diversity and transmission in  
347 humans. *Nat. Genet.* **48**, 195–200 (2016).
- 348 14. Leonard, A. S., Weissman, D. B., Greenbaum, B., Ghedin, E. & Koelle, K.  
349 Transmission Bottleneck Size Estimation from Pathogen Deep-Sequencing Data, with  
350 an Application to Human Influenza A Virus. *J. Virol.* **91**, (2017).
- 351 15. Martin, M. A. & Koelle, K. Reanalysis of deep-sequencing data from Austria  
352 points towards a small SARS-COV-2 transmission bottleneck on the order of one to  
353 three virions. *bioRxiv* 2021.02.22.432096 (2021) doi:10.1101/2021.02.22.432096.
- 354 16. Varble, A. *et al.* Influenza A virus transmission bottlenecks are defined by  
355 infection route and recipient host. *Cell Host Microbe* **16**, 691–700 (2014).
- 356 17. Larremore, D. B. *et al.* Test sensitivity is secondary to frequency and turnaround  
357 time for COVID-19 screening. *Sci. Adv.* **7**, eabd5393 (2021).

- 358 18. Liu, T. *et al.* Risk factors associated with COVID-19 infection: a retrospective  
359 cohort study based on contacts tracing. *Emerg. Microbes Infect.* **9**, 1546–1553 (2020).
- 360 19. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ  
361 preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
- 362 20. Martin, M. Cutadapt removes adapter sequences from high-throughput  
363 sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
- 364 21. Itokawa, K., Sekizuka, T., Hashino, M., Tanaka, R. & Kuroda, M. Disentangling  
365 primer interactions improves SARS-CoV-2 genome sequencing by multiplex tiling  
366 PCR. *PLOS ONE* **15**, e0239403 (2020).
- 367 22. Li, H. & Durbin, R. Fast and accurate short read alignment with  
368 Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.* **25**, 1754–1760 (2009).
- 369 23. Grubaugh, N. D. *et al.* An amplicon-based sequencing framework for accurately  
370 measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol.* **20**, 8  
371 (2019).
- 372 24. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution.  
373 *Bioinformatics* **34**, 4121–4123 (2018).
- 374 25. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood  
375 phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321  
376 (2010).
- 377 26. Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences  
378 with variable rates over sites: approximate methods. *J. Mol. Evol.* **39**, 306–314  
379 (1994).
- 380 27. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree: an r package for  
381 visualization and annotation of phylogenetic trees with their covariates and other  
382 associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).

383

384

385 Figures



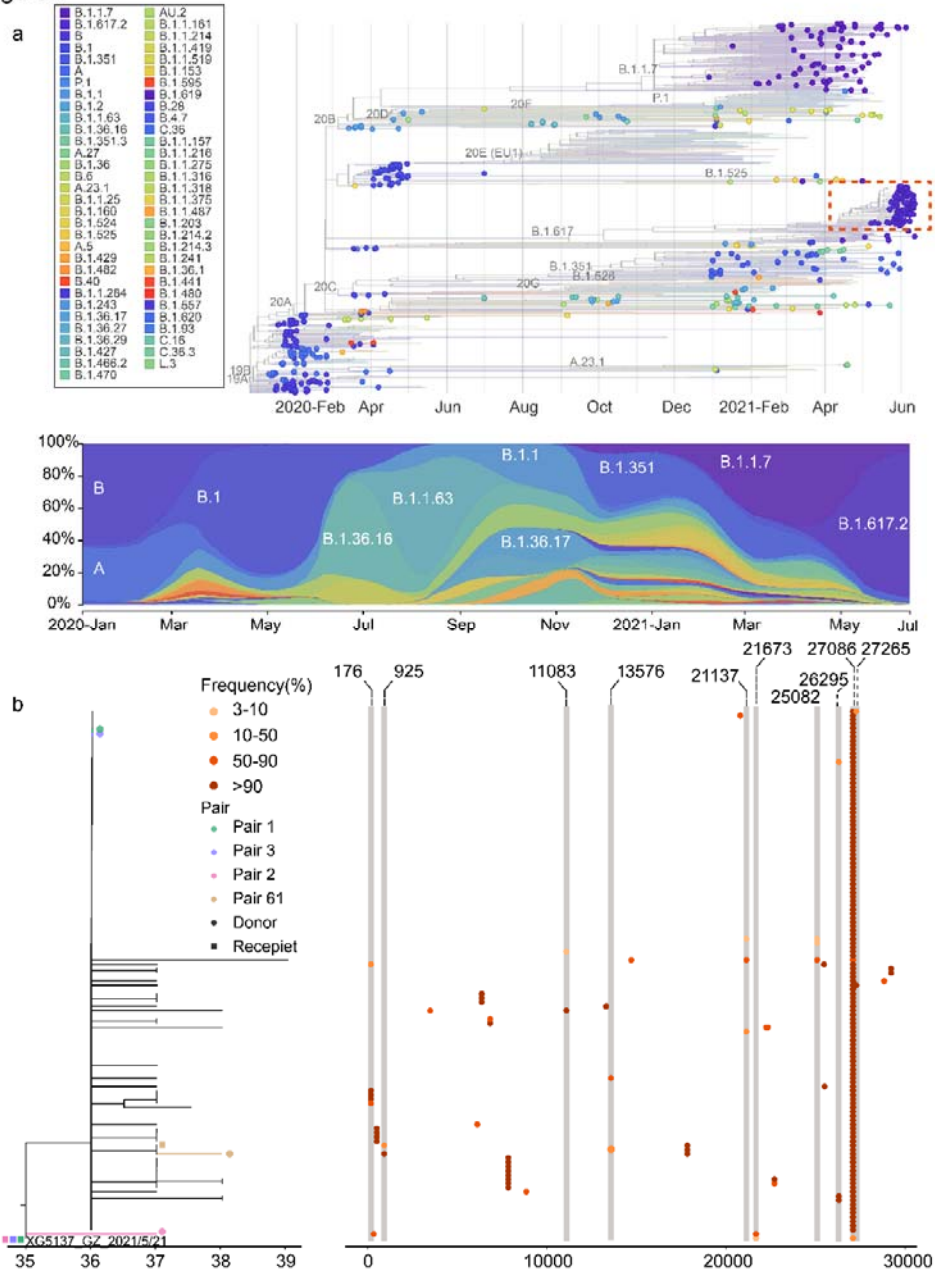
386

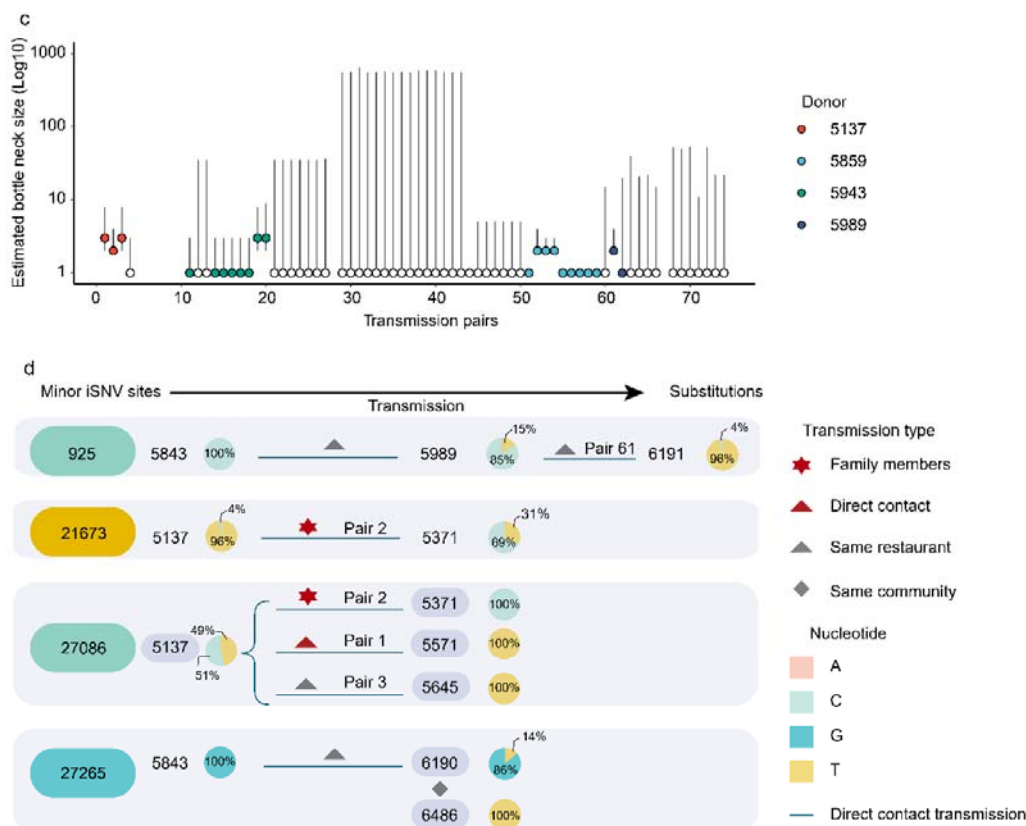
387 Figure 1: Summary of the epidemiology and early detection of the Delta  
388 SARS-CoV-2 variant in Guangdong. (a) Time series of 167 laboratory-confirmed  
389 infections originating from the first index case on May 21 2021. Daily numbers of  
390 new infections are shown in red and samples with high-quality sequences



391 (coverage>95%) are shown in blue. (b) The Delta variant transmission in the  
392 Guangzhou outbreak. The transmission relationship between 126 sequenced cases  
393 were indicated with solid lines (high confidence) or the dash lines (unsure). The  
394 interactive version showing summary statistics of all cases could be found at  
395 <https://viz.vslashr.com/guangdongcdc/>. (c) Estimate of the time interval between  
396 exposure and time of the first RT-PCR positive test in quarantined subjects. The  
397 curves show the best-fitting distributions of the interval durations for Delta variant  
398 cases (n=34) and for 19A/19B clade cases (n=29). Bars show the histograms of  
399 estimated intervals durations (days). (d) Ct values of the first PCR+ test in  
400 quarantined subjects, for the Delta variant infections (n=62) and for previous  
401 19A/19B clade strains infections (n=63). Dots represent Ct values for RT-PCR of the  
402 ORF1ab gene (left) and N gene (right). Box plots indicate the median and  
403 interquartile range (IQR); the whiskers represent the maximum and minimum values.  
404 (e) Schematic of the relation between the viral growth rate and the relative viral loads  
405 on the day viruses were first detected (Day 0). The viral load of A/B clade infections  
406 and of the Delta variant infections on Day 0 were measured. The horizontal dashed  
407 line in purple represents the detection threshold of RT-PCR testing; the dashed line in  
408 red represents the lower limit above which infectious viruses could be potentially  
409 isolated.

Figure2





411

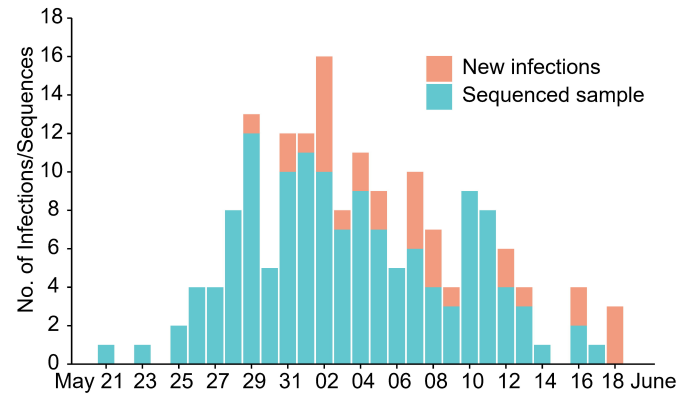
412 Figure 2: Viral phylogenies and transmission dynamics of the Guangzhou outbreak. (a)  
 413 A time resolved phylogenetic tree was estimated using the NextStrain pipeline and  
 414 includes (i) Guangdong sequences collected from local infections and imported cases,  
 415 January 2020 – June 2021, and (ii) reference sequences from different genetic  
 416 lineages. The sequences from the Guangdong Delta variant outbreak (May 21,2021 –  
 417 June 18, 2021) are highlighted with a red box. The changing frequencies of

418 SARS-CoV-2 lineages identified in Guangdong (most of which are imported) are  
419 shown in the lower panel (b) Maximum likelihood tree of 126 sampled sequences of  
420 the Guangzhou outbreak. SNV frequencies (%) across the virus genome are marked  
421 with colored dots (right hand panel). (c) Estimated bottleneck size in 66  
422 donor-recipient transmission pairs calculated using the exact beta-binomial method  
423 described in <sup>14</sup>. There were 8 transmission pairs with extremely large confidence  
424 interval (range from 1 to more than 1000) of estimated bottleneck size were removed.  
425 (d) Minor iSNVs transmission resulted in the diversity of viral population. The pie  
426 charts show the frequency of iSNVs. Arrows show the direction of transmission for  
427 those pairs of cases for which this is known with high confidence.

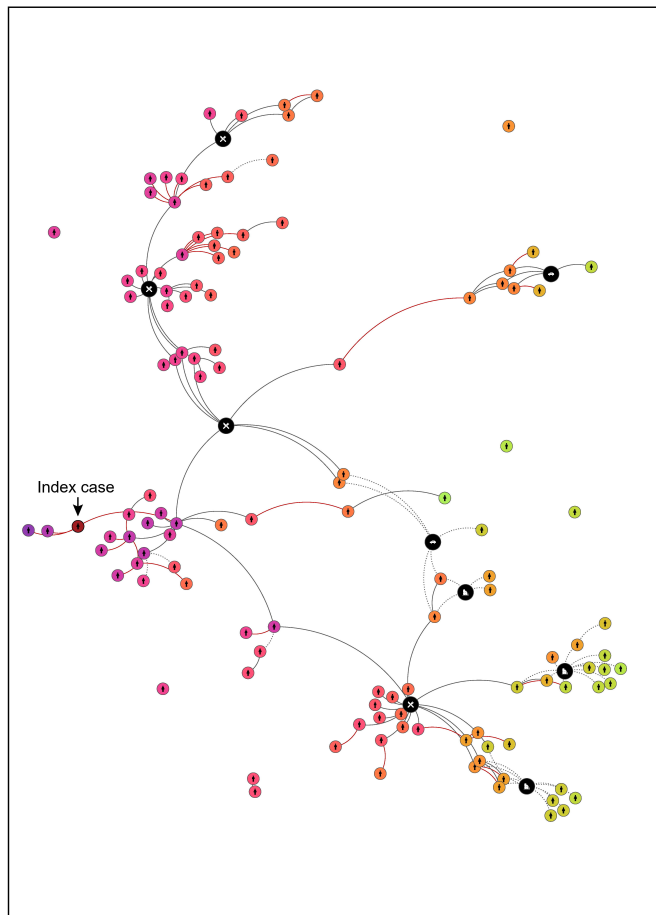
428

# Figure 1

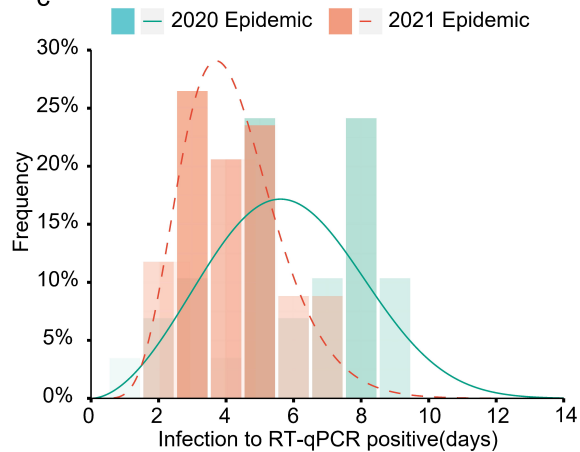
a



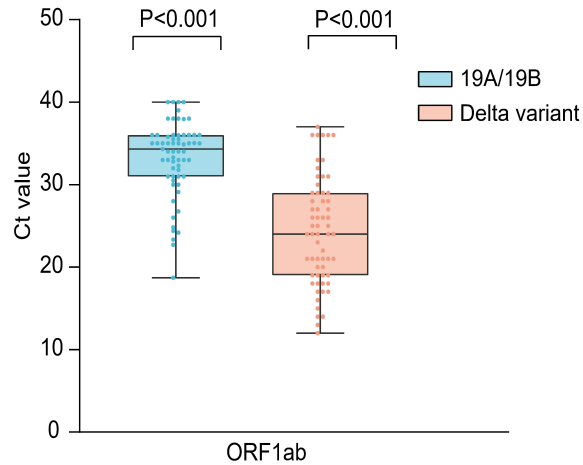
b



c



d



e

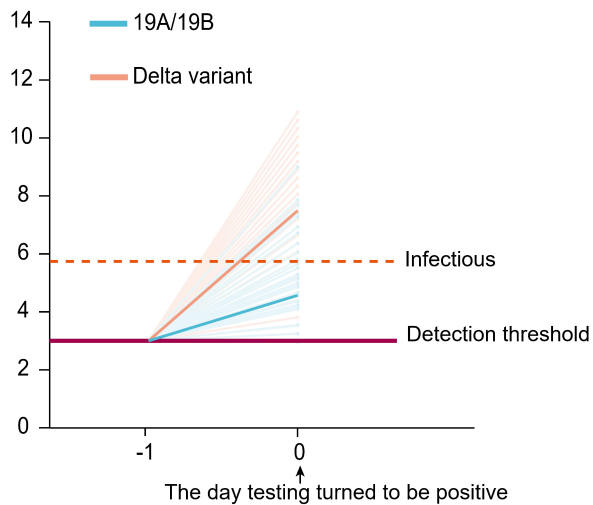
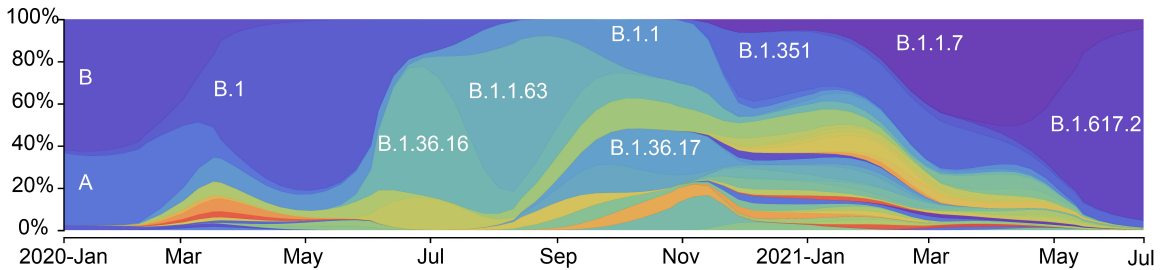
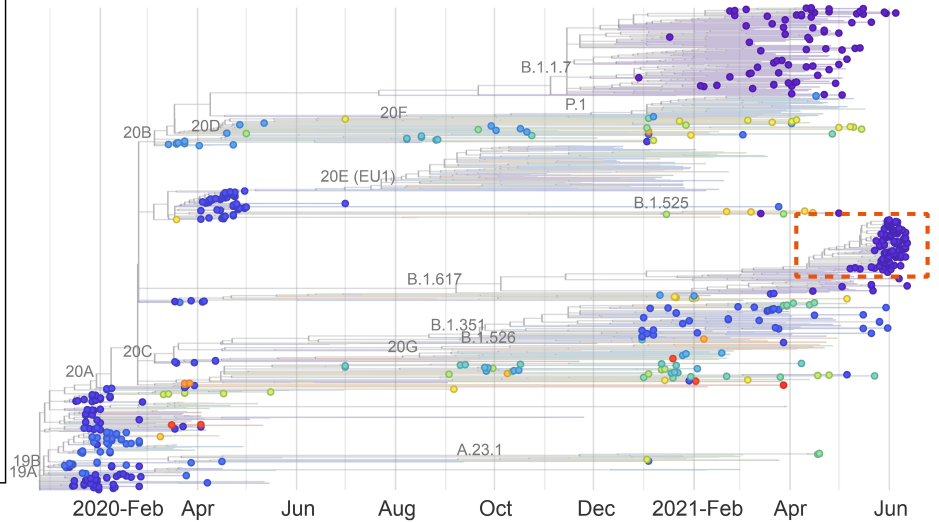
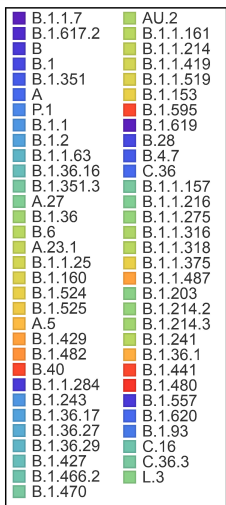
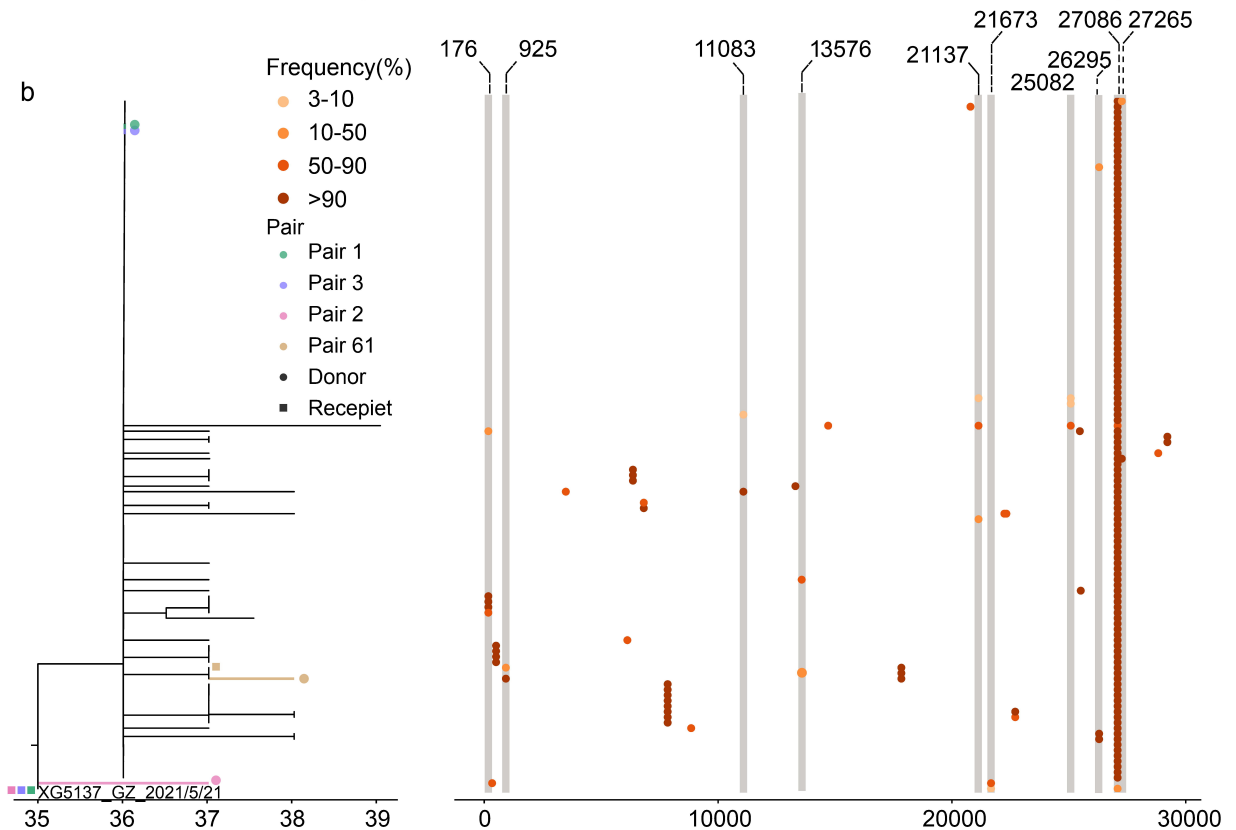


Figure2

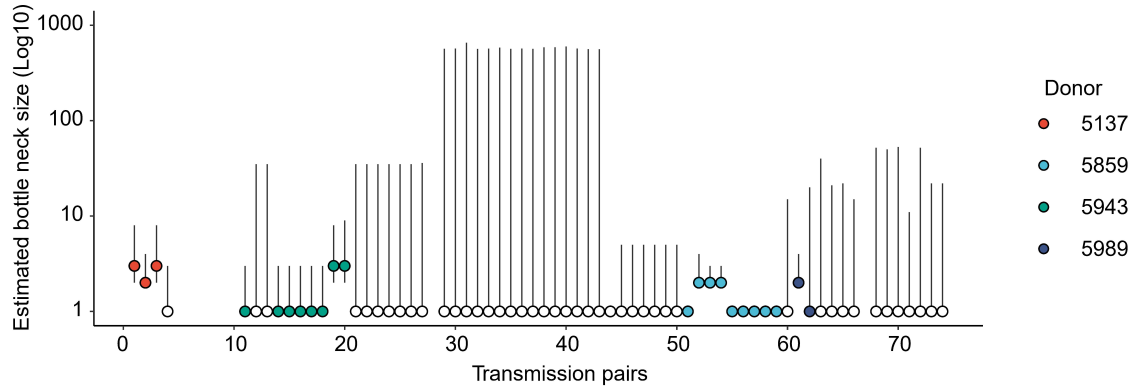
a



b



c



d

