

1 **Comparative evaluation of the transmissibility of**
2 **SARS-CoV-2 variants of concern**

3 Liang Wang^{1,*}, Xavier Didelot², Yuhai Bi^{1,3*}, George F Gao^{1,3,*}

4 ¹CAS Key Laboratory of Pathogenic Microbiology and Immunology, Institute of
5 Microbiology, Center for Influenza Research and Early-warning (CASCIRE),
6 CAS-TWAS Center of Excellence for Emerging Infectious Diseases (CEEID),
7 Chinese Academy of Sciences, Beijing 100101, China

8 ²School of Life Sciences and Department of Statistics, University of Warwick,
9 Coventry CV47AL, United Kingdom

10 ³University of Chinese Academy of Sciences, Beijing 101408, China

11 * Correspondence: wangliang@im.ac.cn (L.W); beeyh@im.ac.cn (Y.B) ;
12 gaof@im.ac.cn (G.F.G)

13 **Keywords**

14 COVID-19, SARS-CoV-2, lineage, transmissibility

15

16 **Abstract**

17 Since the start of the SARS-CoV-2 pandemic in late 2019, several variants of concern
18 (VOC) have been reported, such as B.1.1.7, B.1.351, P.1, and B.1.617.2. The exact
19 reproduction number R_t for these VOCs is important to determine appropriate control
20 measures. Here, we estimated the transmissibility for VOCs and lineages of
21 SAR-CoV-2 based on genomic data and Bayesian inference under an epidemiological
22 model to infer the reproduction number (R_t). We analyzed data for multiple VOCs
23 from the same time period and countries, in order to compare their transmissibility
24 while controlling for geographical and temporal factors. The lineage B had a
25 significantly higher transmissibility than lineage A, and contributed to the global
26 pandemic to a large extent. In addition, all VOCs had increased transmissibility when
27 compared with other lineages in each country, indicating they are harder to control
28 and present a high risk to public health. All countries should formulate specific
29 prevention and control policies for these VOCs when they are detected to curve their
30 potential for large-scale spread.

31

32 **Introduction**

33 As the seventh coronavirus which could infect humans and then caused Coronavirus
34 Disease 2019 (COVID-19), SARS-CoV-2 (also known as 2019-nCoV, or HCoV-19)¹
35 was first identified in Wuhan China, in late 2019²⁻⁴. Within a few weeks,
36 SARS-CoV-2 spread all over the world, and caused a global pandemic⁵ declared by
37 the World Health Organization (WHO), which is the only pandemic caused by a
38 coronavirus to date. As of 8th June 2020, there are more than 172 million confirmed
39 cases from more than 200 countries with more than three million deaths⁶, posing a
40 global threat to public health. Furthermore, the global spread of COVID-19 has also
41 thoroughly taxed the medical systems and global economies.

42

43 The transmissibility of infectious diseases can be measured by the basic reproduction
44 number R_0 , which indicates how many secondary infectees could, on average, be
45 directly caused by one infector in a susceptible population. The higher the R_0 , the
46 higher the transmissibility of the infectious disease, which also means that the
47 infectious disease is more difficult to be controlled. By extension, the temporal
48 reproduction number R_t can be defined as the average number of secondary infections
49 at time t . Traditionally, R_t is estimated using epidemiological data, which can be either
50 individual contact tracing data or population-scale incidence data fitted with systems
51 of ordinary differential equations and that represents a population-level
52 epidemiological model⁷. However, getting unbiased datasets to apply these methods
53 can be challenging. Here we used an alternative which is to use sequencing data to

54 reconstruct a transmission tree which is informative about R_t . As previous described,
55 mutations in the genome of the SARS-CoV-2 have frequently occurred and
56 accumulated during the epidemic. Some of these mutations may have increased the
57 transmissibility, whereas the majority would likely have had no effect, but are still
58 useful to reconstruct transmission trees. The assessment of the effect of mutations on
59 transmissibility has been mainly based on non-human experimental animals (like
60 hamsters *etc*), and it is still controversial whether these conclusions apply to humans.
61 Besides, the timely adjustment of epidemic prevention and control strategies also
62 requires a rapid assessment of the impact of newly emerging important mutations
63 within pathogens' genomes on transmission. Furthermore, several types of
64 SARS-CoV-2 variants of concern (VOC) emerged during the pandemic, such as
65 B.1.1.7 (WHO label: Alpha), B.1.351 (WHO label: Beta), P.1 (WHO label: Gamma),
66 and B.1.617.2 (WHO label: Delta) *etc*. Under these circumstance, novel methods are
67 needed that can quickly evaluate the impact of mutations on transmissibility.

68

69 Here, we estimated R_t for different lineages of SARS-CoV-2 based on genomic data
70 and Bayesian inference under an epidemiological model, and then inferred the
71 offspring distribution. The mean of the offspring distribution is the temporal
72 reproductive number R_t , which depends on both the pathogen transmissibility and the
73 conditions in the host population (for example the proportion of immunized
74 individuals or the control measures in place). To account for this, we compared the R_t
75 of different lineages in the same country and during same periods to quantify the

76 difference transmissibility between different lineages, especially for the previous
77 described VOCs.

78

79 **Results**

80 **Lineage B has a higher transmissibility than lineage A**

81 Since only the United States and Australia contained sufficient numbers of viral
82 genomes from both lineage A and B during the early phase of the COVID-19
83 pandemic, we used data from these two countries to compare the transmissibility
84 between lineages A and B. The mean R_t for lineage A from Australia and USA were
85 estimated as 1.75 (95% credible intervals (CI) 1.43-2.11) and 1.74 (95% CI 1.61-1.89),
86 respectively (Figure 1A). However, the mean R_t for lineage B from Australia and USA
87 were estimated as 2.33 (95% CI 2.05-2.64) and 3.18 (95% CI 2.76-3.63), respectively
88 (Figure 1A). Firstly, the R_t of lineage B is significantly greater than that of lineage A,
89 indicating higher transmissibility of lineage B compared to lineage A. This might be
90 the reason why strains from lineage B rapidly became dominantly all over the world
91 (Figure 1B). Secondly, the R_t of lineage A from the two countries are very close,
92 however, the R_t of lineage B varied greatly between Australia and USA. We then
93 found that the composition of lineage was significantly different between the datasets
94 from these two countries (Figure 1C and D, $p < 0.01$, Fisher's exact test, two-sided).
95 We speculated that different sub-lineages within lineage B might have different
96 transmissibility and then tested the hypothesis by conducting further analysis. Since
97 the data from lineage A was limited, the evaluation of transmissibility for each

98 sub-lineage was mainly focused on those from lineage B and other emerging lineages
99 in the same country during the same periods. In order to reduce the amount of
100 calculation but at the same time be able to test the above hypothesis, only the
101 dominant lineages showing exponential growth in each country were selected to
102 perform the further analysis and comparison.

103

104 **B.1.1.7 has a higher transmissibility than other dominant lineages in UK**

105 The composition of lineages in UK is shown in Figure 2A. B.1.177 was the dominant
106 strain before 2021. We also found that the number of viral genomes from England far
107 exceeds that from other parts of the UK (Figure 2B). Besides, according to the
108 accumulation of number of viral genomes from each lineage in England, we could
109 find that only three lineages (B.1.177, B.1.1.37, B.1.1.7) grew exponentially after
110 October 2020 (Figure 2B). Taken together, only transmissibility of these three
111 lineages were evaluated during October 2020 to January 2021 in this study, so that the
112 impact of non-pharmaceutical interventions on the estimation of R_t will be consistent
113 for different lineages. The R_t for B.1.177, B.1.1.37, B.1.1.7 were estimated as 1.08 (95%
114 CI 1.072-1.09), 1.068 (95% CI 1.05-1.086), and 1.186 (95% CI 1.158-1.213) (Figure
115 2C). The B.1.177, B.1.1.37 had similar R_t which were both close to 1. However,
116 B.1.1.7 had a significantly higher transmissibility than these two lineages. We next
117 tested if the significantly high R_t could be affected by sampling bias. After five
118 independently repeated sampling and subsequent analysis, we found that all these R_t
119 for B.1.1.7 were close to each other, ranging from 1.178 to 1.194. Besides, all the 95%

120 credible intervals from repeated sampling also did not have any intersection with
121 those from lineage B.1.177 and B.1.1.37. Thus, the sampling bias had limited effect
122 on the estimation of R_t for each lineage. We also found that B.1.177 had a similar
123 transmissibility than B.1.1.37 (Student's t test, two-sided with Holm–Bonferroni
124 adjusted $p = 0.1$) (Figure 2D).

125

126 **Slightly higher transmissibility for B.1.351 than B.1.1.54 in South Africa**

127 The composition of lineages in South Africa is shown in Figure 3A. Lineage B.1.1.54
128 was the dominant strain before October 2020. Since then, the dominant strain in South
129 Africa was switched to lineage B.1.351 gradually. According to the accumulation of
130 number of viral genomes from each lineage in South Africa, we could find that only
131 lineage B.1.1.54 and B.1.351 grew exponentially after July 2020 (Figure 3B). In this
132 case, only transmissibility of these two lineages were evaluated during July 2020 to
133 February 2021 in this study, so that the impact of non-pharmaceutical interventions on
134 the estimation of R_t will be consistent for these two lineages. We could find the R_t for
135 B.1.351 and B.1.54 during July 2020 and February 2021 were estimated as 1.05 (95%
136 CI 1.044-1.065) and 1.02 (95% CI 1.011-1.034), respectively (Figure 3C). The
137 difference of transmissibility between B.1.351 and B.1.54 was also significant
138 (Student's t test, two-sided $p < 0.001$) (Figure 3D). In this case, isolates from B.1.351
139 had a slightly higher transmissibility than those from B.1.154.

140

141 **P.1 had a slightly higher transmissibility than P.2 in Brazil**

142 The composition of lineages in Brazil is shown in Figure 4A. Lineage B.1.1.33 and
143 B.1.1.28 were the dominated before January 2021. Both of them grew exponentially
144 after their first appearance in Brazil. However, their growth rate has slowed down
145 since July 2020. Since October 2020, two novel lineages (P.1 and P.2) had gradually
146 appeared and had shown exponential growth (Figure 4B). In this case, only
147 transmissibility of these two lineages (P.1 and P.2) were evaluated during December
148 2020 to February 2021 in this study, so that the impact of non-pharmaceutical
149 interventions on the estimation of R_t will be consistent for these two lineages. We
150 could find the R_t for P.1 and P.2 during December 2020 to February 2021 were
151 estimated as 1.07 (95% credible intervals 1.054-1.084) and 1.06 (95% credible
152 intervals 1.049-1.070) (Figure 4C), respectively. The difference of transmissibility
153 between P.1 and P.2 was also significant (Student's t test, two-sided $p=0.016$) (Figure
154 4D). In this case, isolates from P.1 had a slightly higher transmissibility than those
155 from P.2.

156

157 **B.1.617.2 has a higher transmissibility than other dominant lineages in India**

158 The top five dominant lineages and their corresponding proportion in India are shown
159 in Figure 5A. The B.1.306 was the dominated lineage in India. Since July 2020,
160 several other lineages, like B.1, B.1.36, B.1.36.29, emerged and grew exponentially in
161 India (Figure 5B). B.1.617.1 and B.1.617.2 were detected in India at late 2020, and
162 then they both grew exponentially in India (Figure 5B). Lineage B.1.617.2 has already
163 been considered as VOC by WHO. We also found lineage B.1, B.1.36, B.1.36.29,

164 B.1.617.1, B.1.617.2 grew exponentially after 1st January 2021. In order to reduce the
165 calculation, only data collected after 1st January 2021 were used to perform the further
166 analysis so that the impact of non-pharmaceutical interventions on the estimation of R_t
167 will be consistent for these lineages. In this case, only these five lineages were used to
168 estimate their R_t . The R_t was estimated as 1.013 (95% CI 1.006-1.021), 1.018 (95% CI
169 1.009-1.027), 1.019 (95% CI 1.010-1.027), 1.033 (95% CI 1.026-1.040), 1.123 (95%
170 CI 1.106-1.140) for B.1, B.1.36, B.1.36.29, B.1.617.1, B.1.617.2, respectively (Figure
171 5C). After 5 independently repeated sampling and followed analysis for each lineage,
172 we found that both B.1.617.1 and B.1.617.2 had significantly higher transmissibility
173 than B.1, B.1.36, and B.1.36.29 (all Student's t test, two-sided with Holm–Bonferroni
174 adjusted $p < 0.001$) (Figure 5D). Furthermore, B.1.617.2 also had a significantly higher
175 transmissibility than B.1.617.1 (Student's t test, two-sided with Holm–Bonferroni
176 adjusted $p < 0.001$). In addition, the transmissibility of both B.1.36, and B.1.36.29 is
177 significantly higher than that of B.1 (both Student's t test, two-sided with
178 Holm–Bonferroni adjusted $p < 0.001$) (Figure 5D). However, similar transmissibility
179 was found between B.1.36 and B.1.36.29 (Student's t test, two-sided with
180 Holm–Bonferroni adjusted $p = 0.057$) (Figure 5D).

181

182 **Discussion**

183 Assessing the transmissibility of pathogens is essential to tailor prevention and control
184 strategies. As the COVID-19 pandemic spread, several VOC have been found, such as
185 B.1.1.7 (WHO label: Alpha), B.1.351 (WHO label: Beta), P.1 (WHO label: Gamma),

186 and B.1.617.2 (WHO label: Delta) *etc.* The emergence of these VOCs has caused a
187 significant threat to public health. Since vaccination is the key to global containment
188 of the COVID-19 pandemic, a reduced vaccine efficacy against some VOCs would
189 increase the risk of infection in immunized individuals thereby increasing the
190 difficulty of containing the spread of the pandemic. For example, B.1.1.7 has been
191 documented to have reduced neutralization by original strain convalescent and
192 vaccine sera⁸⁻¹⁰. B.1.351 and P.1 also had reduced neutralization by mAbs and sera
193 induced by early SARS-CoV-2 isolates¹¹⁻¹³ and B.1.351 might also increase the risk of
194 infection in immunized individuals¹⁴. However, novel VOCs might emerge at any
195 time and anywhere in the future. In order to deal with a novel VOC, it is necessary to
196 quickly evaluate its transmissibility and use this as a basis to determine whether
197 prevention and control strategies need to be adjusted to control the epidemic. A
198 previous study had documented that B.1.1.7 has an advanced transmissibility
199 compared to other lineages circulating in UK (43%-90% with 95% confidence
200 intervals ranging from 38% to 130%)¹⁵. Another study illustrated that P.1 also had an
201 increased transmissibility by 54%-79% compared to non-P.1 lineage¹⁶. These results
202 show that different lineage can have different transmissibility. Together with the
203 changes in transmissibility for B.1.351 and B.1.617.2 which had not been previously
204 elucidated, here we estimated the lineage-specific transmissibility for each lineage,
205 especially for these VOCs.

206

207 The results show that lineage B has a significantly higher transmissibility than lineage

208 A (Figure 1A). Together with the fact that lineage B was the dominant types of
209 SARS-CoV-2 all over the world, it seems that the high transmissibility of lineage B
210 contributed to the global pandemic to a large extent. However, we also found that the
211 transmissibility for lineage B from Australia and USA differed significantly.
212 Considering the significantly different composition of sub-lineages among these two
213 countries, we speculated that different sub-lineage within lineage B would have
214 different transmissibility. We estimated the transmissibility of VOCs and the dominant
215 lineages with exponential growth during same period in each country, so that the
216 impact of non-pharmaceutical interventions on the estimation of R_t will be consistent
217 among different lineages. We estimated R_t for different lineages of SARS-CoV-2
218 based on genomic data and Bayesian inference under an epidemiological model, and
219 then inferred the mean of offspring distribution (R_t). Since limited variants among
220 each lineage would lead to uncertainty on phylogeny and the estimation of R_t was
221 solely based on dated-phylogenetic tree, it is necessary to assess how the phylogenetic
222 uncertainty affect the estimation of R_t . The estimation of R_t from random selected tree
223 from MCMC chain were always lower than for the MCC tree (Figure S1). As the
224 MCC tree is more accurate than to trees sampled in MCMC chains, this result
225 suggested that the uncertainty of the phylogeny would cause an underestimation of the
226 R_t . In this case, the use of MCC tree for estimation of R_t would reduce the impact of
227 phylogenetic uncertainty on the results as much as possible. In addition, the sampling
228 bias could also affect the phylogeny.

229

230 B.1.1.7 had a significant advance in transmissibility than B.1.37 and B.1.177 in UK.
231 The result was consistent with the previous report that B.1.1.7 had a higher
232 transmissibility than other lineages¹⁵. However, the increase in transmissibility of
233 B.1.1.7 estimated in this study was not as much as previous reported¹⁵, presumably
234 because the increase of transmissibility of B.1.1.7 was based on comparison to the
235 superimpose state of all other lineages for previous report. On the other hand, the
236 increase of transmissibility of B.1.1.7 was based on comparison to two other lineages
237 (B.1.1.37 and B.1.177) in the UK. Since B.1.1.37 and B.1.177 grew exponentially in
238 the UK, the transmissibility for these two lineages could be higher than those lineages
239 without exponentially growth. In this case, the increase of B.1.1.7 was not as much as
240 higher than previous report¹⁵. The transmissibility of B.1.1.7 has indeed increased and
241 is in line with the results from other reports, which further proves the accuracy of our
242 method. We also found that P.1 had a higher transmissibility than P.2 in Brazil, and
243 B.1.351 had a higher transmissibility than B.1.1.54 in South Africa. However, the
244 extent of increased transmissibility for P.1 and B.1.351 against to other dominant
245 lineages with exponential growth was not as much as for B.1.1.7. In India, we found
246 that both B.1.617.1 and B.1.617.2 had significant increase in transmissibility
247 compared to other lineages with exponential growth. Furthermore, B.1.617.2 also had
248 a significantly higher transmissibility than B.1.617.1. In addition, B.1.36 and
249 B.1.36.29 had similar transmissibility, both higher than B.1.

250

251 These results indicated that different lineages of SARS-CoV-2 have different

252 transmissibility, with some differences being more significant than others. The
253 transmissibility of four types of VOCs also increased to varying degrees. All countries
254 should formulate corresponding prevention and control policies for these VOCs to
255 avoid large-scale outbreaks in their countries.

256

257 **Methods**

258 **Data collection, selection and pre-processing**

259 The transmission could be significantly affected by the stringent prevention and
260 control strategies. Only data collected before the implementation of stringent
261 epidemic control measures in each country were used for lineages A and B, so as to
262 minimize the impact of prevention and control strategies on the estimation of R_t , and
263 reflect the real situation at the same time. SARS-COV-2 genomic sequences were
264 download from GISAID several times (data for estimating lineage A and B was
265 downloaded at 9th April 2020, data for UK was downloaded at 21st December 2020,
266 data for South Africa and Brazil was downloaded at 16th March 2021, data for India
267 was downloaded at 13th May 2021). Before estimating transmissibility of lineage A
268 and B during the early phase of COVID-19 pandemic, we first filtered data. Only
269 those viral genomes collected before the implementation of national
270 non-pharmaceutical interventions would be included in the analysis. In addition, those
271 countries that include lineage A and B, and the number of completely viral genomes
272 within each lineage ≥ 80 would be included in the subsequent analysis. Since only the
273 United States and Australia met the above criteria, the estimation of the

274 transmissibility of lineage A and B was only based on the data of these two countries.
275 The cut-off dates for the collection time in the USA and Australia are 20th and 25th
276 January 2020, respectively, as there were no nationwide epidemic prevention
277 measures were implemented before the date. Due to the high volume of genomic data
278 from sub-lineages in UK, South Africa, Brazil, and India, the amount of calculation
279 would be too large, especially for reconstruction of dated phylogeny. In this case, we
280 filtered and also sub-sampled the data for datasets from each sub-lineage. First, the
281 viral genomes of patients who had not had a history of international travel are retained,
282 according to their epidemiological data. Second, the viral genomes should also meet
283 the criteria as follow: length ≥ 29 KB, and the ratio of N in the genome $\leq 1\%$. Third,
284 based on the collection date, if more than 10 genomes were available in a specific date,
285 we randomly select 10 of them, otherwise all genomes would be included. Genomic
286 sequences were aligned using Mafft v7.310¹⁷. Then, we trimmed the uncertain regions
287 in 3' and 5' terminals and also masked 30 sites (Supplementary Table 1) that are
288 highly homoplastic and have no phylogenetic signal as previous noted
289 (<https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473>).

290

291 **Reconstruction of dated phylogeny**

292 As recombination could impact the evolutionary signal, we searched for
293 recombination events in these SARS-CoV-2 genomes using RDP4¹⁸. No evidence for
294 recombination was found in our dataset. We used jModelTest v2.1.6¹⁹ to find the best
295 substitution model for each dataset from different countries according to the Bayesian

296 Information Criterion. The best substitution model for each dataset were listed in
297 Supplementary Table 2. The list of genomic sequences used in this study were
298 provided in Supplementary Table 3. The list of genomic sequences used in this study
299 were openly shared via the GISAID initiative²⁰. We then used the Bayesian Markov
300 Chain Monte Carlo (MCMC) approach implemented in BEAST v1.10.4²¹ to derive a
301 dated phylogeny for SARS-CoV-2. Three replicate runs for each 100 million MCMC
302 steps, sampling parameters and trees every 10,000 steps. For data from lineage A and
303 B in USA and Australia during the early phase of COVID-19, the estimation of the
304 most appropriate combination of molecular clock and coalescent models for Bayesian
305 phylogenetic analysis was determined using both path-sampling and stepping-stone
306 models²². The model comparison result for datasets from lineage A and B in USA and
307 Australia were listed in Supplementary Table 4. In order to reduce the amount of
308 calculation, we assumed that data from sub-lineages followed a strict molecular clock
309 and with an exponential population growth tree prior, as genomic sequences used in
310 each dataset were all from the same sub-lineage and they all had an exponential
311 growth. Tracer 1.7.1²³ was then used to check the convergence of MCMC chain
312 (effective sample size >200) and to compute marginal posterior distributions of
313 parameters, after discarding 10% of the MCMC chain as burn-in. Bayesian evaluation
314 of temporal signal (BETS)²⁴ was used to evaluate the temporal signal in each dataset.
315 BETS relies on the comparison of marginal likelihoods of two models: the
316 heterochronous (with tip date) and isochronous (without tip date) models. Analyses
317 were performed with at least three independent replicates of 100 million MCMC steps

318 each, sampling parameters and trees every 10,000 steps with the best substitution
319 model and most appropriate combination of molecular clock and coalescent models
320 determined above for each dataset. The marginal likelihoods were estimated by PS.
321 The Bayes factor (BF) was then calculated based on the likelihoods of two models
322 (heterochronous and isochronous). If the $\log BF > 30$ (heterochronous model against
323 isochronous model), it indicated there was sufficient temporal signal in this dataset.
324 For dataset without sufficient temporal signal, we specified a clock rate following
325 uniform distribution ranging from 0.0004 to 0.0012 with a mean of 0.0008, otherwise
326 we specified a noninformative continuous-time Markov chain (CTMC) reference prior.
327 The log BF for each dataset was listed in Supplementary Table 5.

328

329 **Estimation of transmissibility using partially sampled viral genomic sequences**

330 As viral genomes were incompletely sampled and the pandemic is currently ongoing,
331 TransPhylo v1.4.4²⁵ was used to infer the transmission tree using the dated phylogeny
332 generated above as input. The generation time (i.e., the time gap from infection to
333 onward transmission, denoted as G) of COVID-19 was previously estimated as $7.5 \pm$
334 3.4 days²⁶ and we used these values to compute the shape and scale parameter of a
335 gamma distribution of G using the R package *epitrix*²⁷. This parameter was used when
336 estimating the transmissibility of lineages A and B in USA and Australia during the
337 early phase of COVID-19 pandemic. However, it was reported that the G was shorten
338 over time by nonpharmaceutical interventions²⁸. In this case, we used 4.8 ± 1.7 days²⁹
339 estimated by previous study as G when estimating the transmissibility of sub-lineages

340 in UK, South Africa and Brazil. The distribution of sampling time (*i.e.* the time gap
341 from infection to detection and sampling) was set equal to the distribution of
342 generation time. We performed the TransPhylo analysis with 100,000 iterations
343 estimating the the offspring distribution (which represents the number of secondary
344 cases caused by each infection). The R_t then could be inferred as the median of the
345 offspring distribution. All results were generated after discarding the first part of the
346 MCMC chains as burn-in. The MCMC mixing and convergence was assessed based
347 on the effective sample size of each parameter (>200) and by visual examination of
348 the MCMC traces. The effective sample size and value of R_t for each dataset was
349 listed in Supplementary Table 6.

350

351 **Evaluating the robustness of the estimation**

352 Since dated phylogeny was used to estimate the transmissibility for each lineage, we
353 should test whether and how the phylogenetic uncertainty and sampling bias affect the
354 estimation. We first tested how the phylogenetic uncertainty affect the result, because
355 only MCC tree was used to estimate the transmissibility. We used data from our
356 previous study³⁰. Ten dated phylogenetic trees were randomly selected from the
357 MCMC chains. The parameter setting was the same as previous study description. In
358 addition, the sampling bias was also a key factor affecting the phylogenetic
359 uncertainty. In order to test the robustness of the estimation of R_t , we also repeatedly
360 randomly sub-sampled the data five times for each dataset and then performed the
361 same analysis.

362 **References**

- 363 1. Jiang, S., *et al.* A distinct name is needed for the new coronavirus. *Lancet* **395**, 949
364 (2020).
- 365 2. Wang, C., Horby, P.W., Hayden, F.G. & Gao, G.F. A novel coronavirus outbreak of
366 global health concern. *Lancet* **395**, 470-473 (2020).
- 367 3. Wenjie, T., *et al.* A novel coronavirus genome identified in a cluster of pneumonia
368 cases — Wuhan, China 2019–2020. *China CDC Weekly* **2**, 61-62 (2020).
- 369 4. Zhu, N., *et al.* A novel coronavirus from patients with pneumonia in China, 2019. *N*
370 *Engl J Med* **382**, 727-733 (2020).
- 371 5. WHO. Coronavirus disease (COVID-2019) situation reports. (2020).
- 372 6. WHO. Coronavirus disease (COVID-19) Weekly epidemiological update and weekly
373 operational update. (2021).
- 374 7. Delamater, P.L., Street, E.J., Leslie, T.F., Yang, Y.T. & Jacobsen, K.H. Complexity of
375 the basic reproduction number (R_0). *Emerg Infect Dis* **25**, 1-4 (2019).
- 376 8. Supasa, P., *et al.* Reduced neutralization of SARS-CoV-2 B.1.1.7 variant by
377 convalescent and vaccine sera. *Cell* **184**, 2201-+ (2021).
- 378 9. Collier, D.A., *et al.* Sensitivity of SARS-CoV-2 B.1.1.7 to mRNA vaccine-elicited
379 antibodies. *Nature* **593**, 136-+ (2021).
- 380 10. Abu-Raddad, L.J., Chemaitelly, H., Butt, A.A. & COVID-19accination, N.S.G.
381 Effectiveness of the BNT162b2 COVID-19 vaccine against the B.1.1.7 and B.1.351
382 variants. *New Engl J Med* (2021).
- 383 11. Zhou, D.M., *et al.* Evidence of escape of SARS-CoV-2 variant B.1.351 from natural

- 384 and vaccine-induced sera. *Cell* **184**, 2348-+ (2021).
- 385 12. Hoffmann, M., *et al.* SARS-CoV-2 variants B.1.351 and P.1 escape from neutralizing
386 antibodies. *Cell* **184**, 2384-+ (2021).
- 387 13. Wang, P., *et al.* Increased resistance of SARS-CoV-2 variant P.1 to antibody
388 neutralization. *Cell Host Microbe* **29**, 747-751 e744 (2021).
- 389 14. Planas, D., *et al.* Sensitivity of infectious SARS-CoV-2 B.1.1.7 and B.1.351 variants to
390 neutralizing antibodies. *Nat Med* **27**, 917-+ (2021).
- 391 15. Davies, N.G., *et al.* Estimated transmissibility and impact of SARS-CoV-2 lineage
392 B.1.1.7 in England. *Science* **372**(2021).
- 393 16. Faria, N.R., *et al.* Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in
394 Manaus, Brazil. *Science* **372**, 815-821 (2021).
- 395 17. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid
396 multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**,
397 3059-3066 (2002).
- 398 18. Martin, D.P., Murrell, B., Golden, M., Khoosal, A. & Muhire, B. RDP4: detection and
399 analysis of recombination patterns in virus genomes. *Virus Evol* **1**, vev003 (2015).
- 400 19. Darriba, D., Taboada, G.L., Doallo, R. & Posada, D. jModelTest 2: more models, new
401 heuristics and parallel computing. *Nat Methods* **9**, 772 (2012).
- 402 20. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative
403 contribution to global health. *Glob Chall* **1**, 33-46 (2017).
- 404 21. Suchard, M.A., *et al.* Bayesian phylogenetic and phylodynamic data integration using
405 BEAST 1.10. *Virus Evolution* **4**(2018).

- 406 22. Baele, G., Li, W.L.S., Drummond, A.J., Suchard, M.A. & Lemey, P. Accurate model
407 selection of relaxed molecular clocks in Bayesian phylogenetics. *Mol Biol Evol* **30**,
408 239-243 (2013).
- 409 23. Rambaut, A., Drummond, A.J., Xie, D., Baele, G. & Suchard, M.A. Posterior
410 summarization in Bayesian phylogenetics using Tracer 1.7. *Syst Biol* **67**, 901-904
411 (2018).
- 412 24. Duchene, S., *et al.* Bayesian evaluation of temporal signal in measurably evolving
413 populations. *Mol Biol Evol* **37**, 3363-3379 (2020).
- 414 25. Didelot, X., Fraser, C., Gardy, J. & Colijn, C. Genomic Infectious Disease
415 Epidemiology in Partially Sampled and Ongoing Outbreaks. *Mol Biol Evol* **34**,
416 997-1007 (2017).
- 417 26. Li, Q., *et al.* Early transmission dynamics in Wuhan, China, of novel
418 coronavirus-infected pneumonia. *New Engl J Med* **382**, 1199-1207 (2020).
- 419 27. Jombart, T., Cori, A., Kamvar, Z.N. & Schumacher, D. epitrix: small helpers and tricks
420 for epidemics analysis. (2019).
- 421 28. Ali, S.T., *et al.* Serial interval of SARS-CoV-2 was shortened over time by
422 nonpharmaceutical interventions. *Science* **369**, 1106+ (2020).
- 423 29. Challen, R., Brooks-Pollock, E., Tsaneva-Atanasova, K. & Danon, L. Meta-analysis of
424 the SARS-CoV-2 serial interval and the impact of parameter uncertainty on the
425 COVID-19 reproduction number. *medRxiv*, 2020.2011.2017.20231548 (2020).
- 426 30. Wang, L., *et al.* Inference of person-to-person transmission of COVID-19 reveals
427 hidden super-spreading events during the early outbreak phase. *Nat Commun*

428 11(2020).

429

430 **Figure Legend**

431 Figure 1. Difference in transmissibility between lineages A and B.

432 A. The distribution of R_t for each lineage. The black line in each distribution
433 indicated the 95% CI.

434 B. The cumulative number of SARS-CoV-2 genomes for each lineage all over the
435 world.

436 C. The heatmap of number of viral genomes for each sub-lineage in lineage A.

437 D. The heatmap of number of viral genomes for each sub-lineage in lineage B.

438 Figure 2. Lineage B of SARS-CoV-2 has a higher transmissibility than lineage A.

439 A. The pie chart of SARS-CoV-2 lineage composition in UK. The circle size was
440 proportion to the number of SARS-CoV-2 genomes.

441 B. The cumulative number of SARS-CoV-2 genomes for each lineage in different
442 region in UK. The dash line indicated the earliest collection date of the data used
443 for estimating the transmissibility for each lineage.

444 C. The distribution of R_t for each lineage. The black line in each distribution
445 indicated the 95% CI.

446 D. The boxplot of repeated estimation of transmissibility by using 5 independent
447 re-sampling data for each lineage. Upper bound, center, and lower bound of box
448 represent the 75th percentile, the 50th percentile (median), and the 25th percentile,
449 respectively.

450 Figure 3. Difference in transmissibility for lineages in South Africa.

451 A. The donut chart of SARS-CoV-2 lineage composition in South Africa.

452 B. The cumulative number of SARS-CoV-2 genomes for each lineage in South
453 Africa. The dash line indicated the earliest collection date of the data used for
454 estimating the transmissibility for each lineage.

455 C. The distribution of R_t for each lineage. The black line in each distribution
456 indicated the 95% CI.

457 D. The boxplot of repeated estimation of transmissibility by using 5 independent
458 re-sampling data for each lineage. Upper bound, center, and lower bound of box
459 represent the 75th percentile, the 50th percentile (median), and the 25th percentile,
460 respectively.

461 Figure 4. Difference in transmissibility for lineages in Brazil.

462 A. The donut chart of SARS-CoV-2 lineage composition in Brazil.

463 B. The cumulative number of SARS-CoV-2 genomes for each lineage in Brazil. The
464 dash line indicated the earliest collection date of the data used for estimating the
465 transmissibility for each lineage.

466 C. The distribution of R_t for each lineage. The black line in each distribution
467 indicated the 95% CI.

468 D. The boxplot of repeated estimation of transmissibility by using 5 independent
469 re-sampling data for each lineage. Upper bound, center, and lower bound of box
470 represent the 75th percentile, the 50th percentile (median), and the 25th percentile,
471 respectively.

472 Figure 5. Difference in transmissibility for lineages in India.

473 A. The donut chart of SARS-CoV-2 lineage composition in India.

- 474 B. The cumulative number of SARS-CoV-2 genomes for each lineage in India. The
475 dash line indicated the earliest collection date of the data used for estimating the
476 transmissibility for each lineage.
- 477 C. The distribution of R_t for each lineage. The black line in each distribution
478 indicated the 95% CI.
- 479 D. The boxplot of repeated estimation of transmissibility by using 5 independent
480 re-sampling data for each lineage. Upper bound, center, and lower bound of box
481 represent the 75th percentile, the 50th percentile (median), and the 25th percentile,
482 respectively.
- 483

484 **Supplementary Information**

485 Figure S1. The 95% CI distribution of R_t using MCC tree and ten randomly selected
486 trees from the MCMC chains.

487 Supplementary Table 1. List of 30 masked sites in SARS-CoV-2 genome.

488 Supplementary Table 2. The best substitution model for dataset from each country.

489 Supplementary Table 3. The acknowledgement table of viral genomes used in this
490 study.

491 Supplementary Table 4. Log-marginal likelihood estimates from model selection by
492 using the path-sampling (PS) and stepping-stone (SS) approaches for lineage A and B.

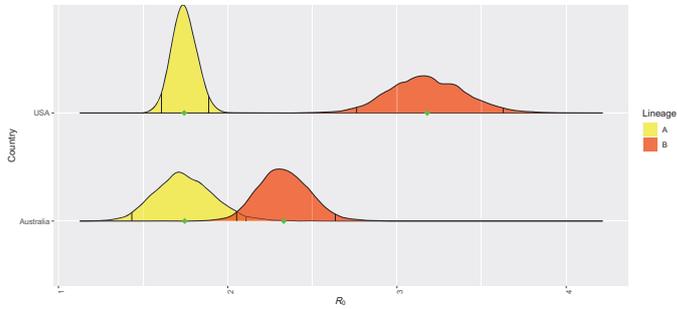
493 Supplementary Table 5. Bayesian evaluation for the temporal signal of dataset from
494 each dataset.

495 Supplementary Table 6. The estimation of R_t and corresponding effective size of each
496 dataset.

497

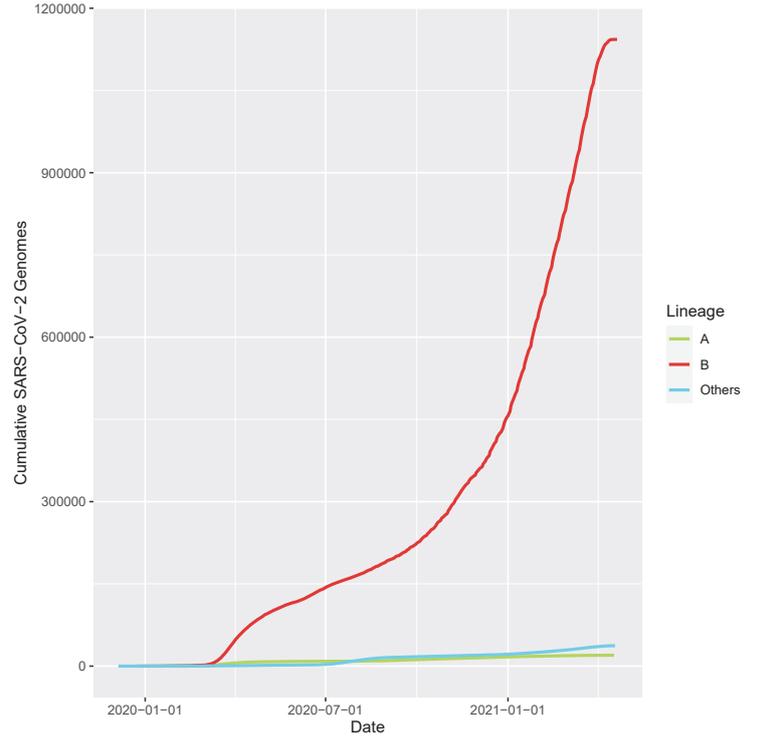
Figure 1

A

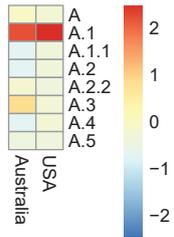


B

medRxiv preprint doi: <https://doi.org/10.1101/2021.06.25.21259565>; this version posted June 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY-NC-ND 4.0 International license](https://creativecommons.org/licenses/by-nc-nd/4.0/).



C



D

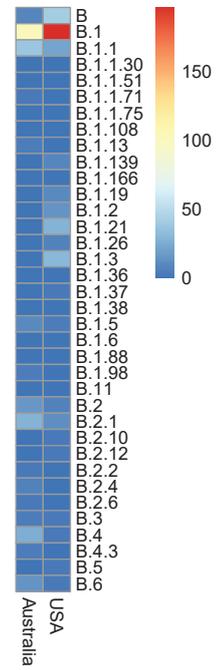
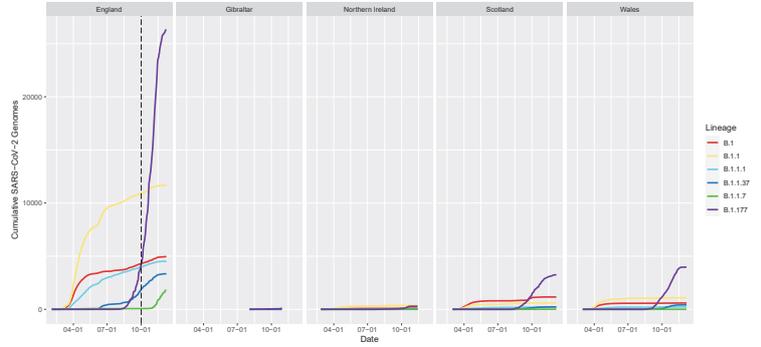
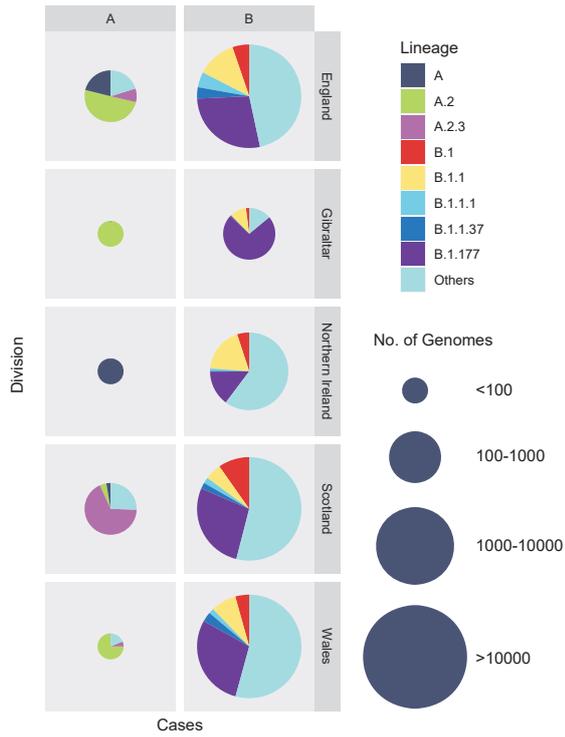
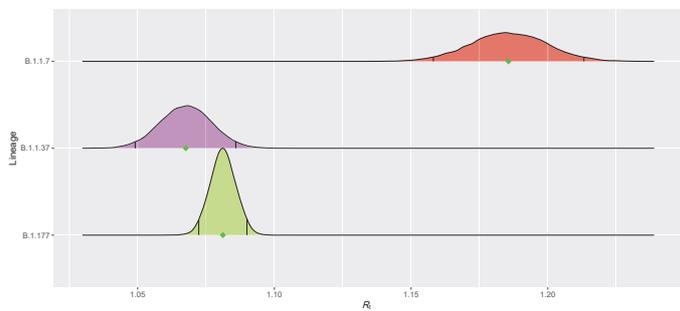


Figure 2

A medRxiv preprint doi: <https://doi.org/10.1101/2021.06.25.21259565>; this version posted June 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY-NC-ND 4.0 International license](https://creativecommons.org/licenses/by-nc-nd/4.0/).



C



D

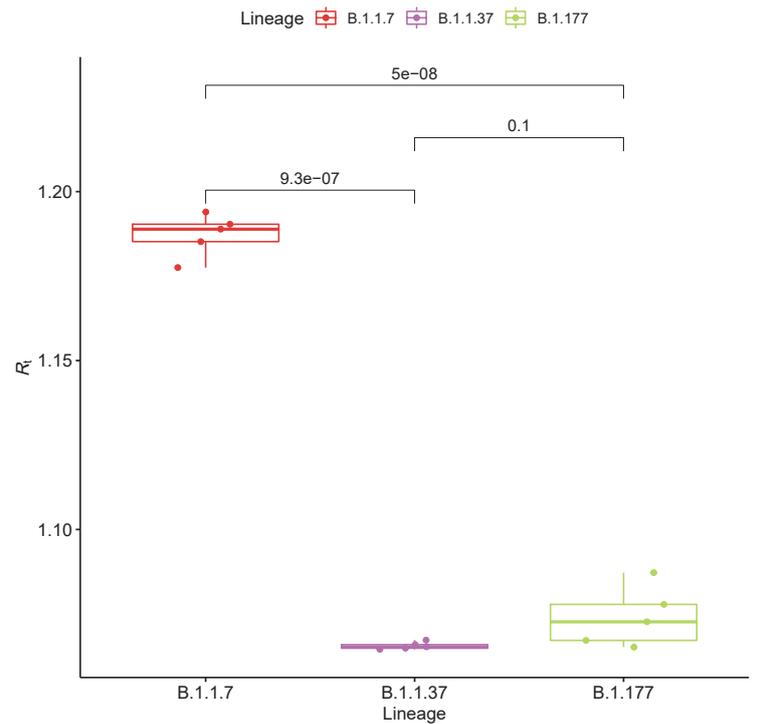
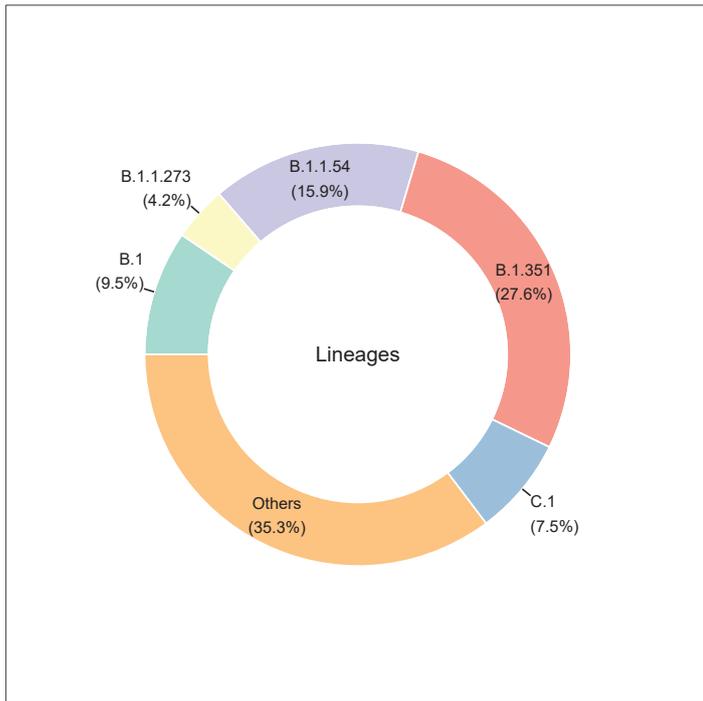


Figure 3

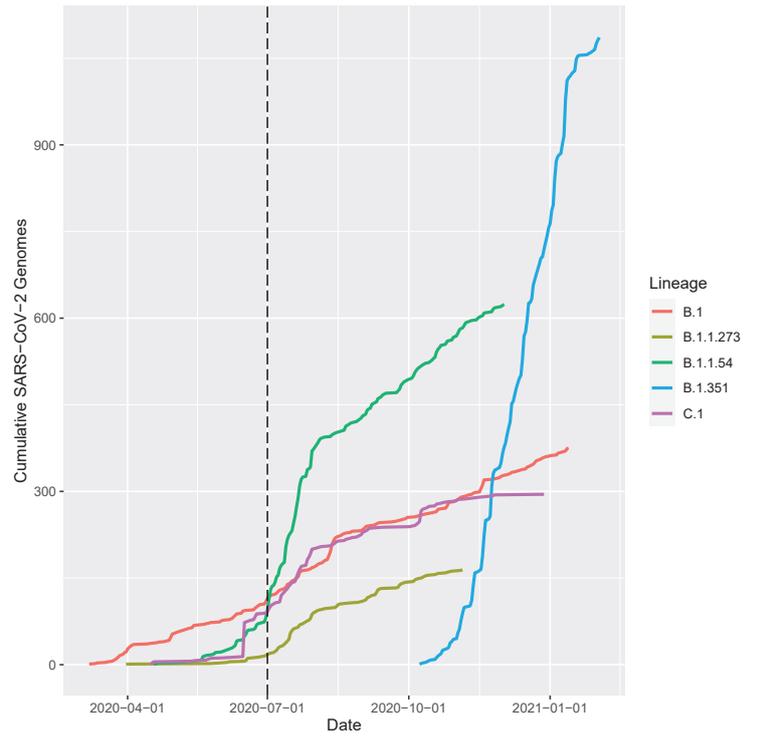
A

medRxiv preprint doi: <https://doi.org/10.1101/2021.06.25.21259565>; this version posted June 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY-NC-ND 4.0 International license](https://creativecommons.org/licenses/by-nc-nd/4.0/).

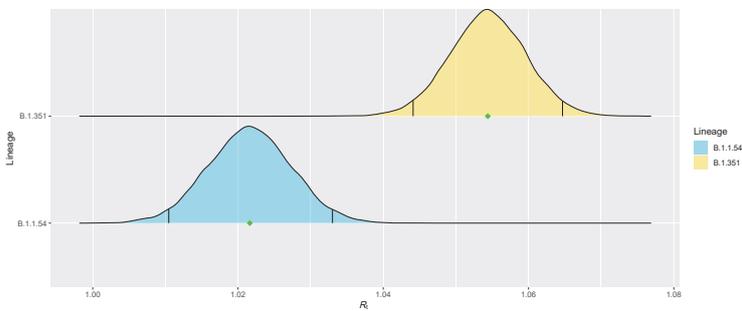
Distribution of lineages for SARS-CoV-2 isolates in South Africa



B



C



D

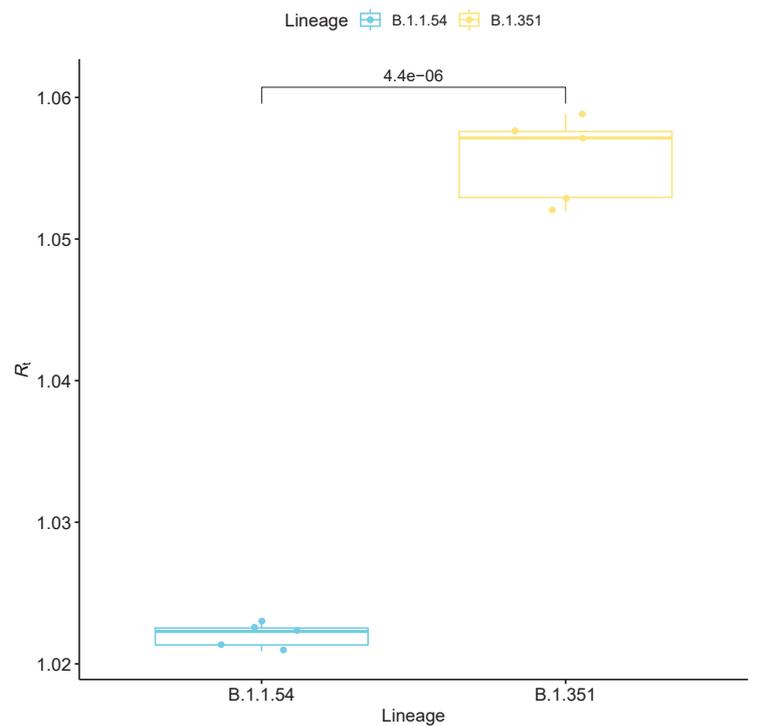
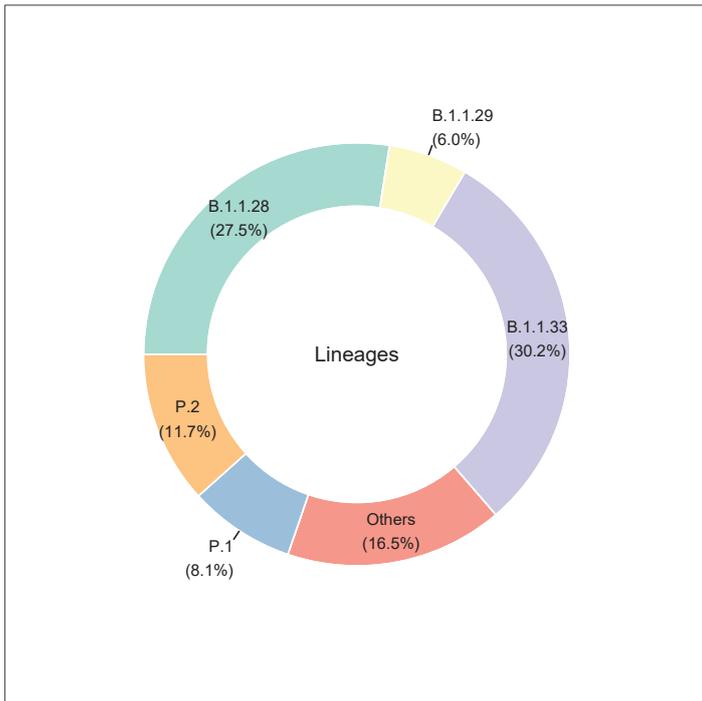


Figure4

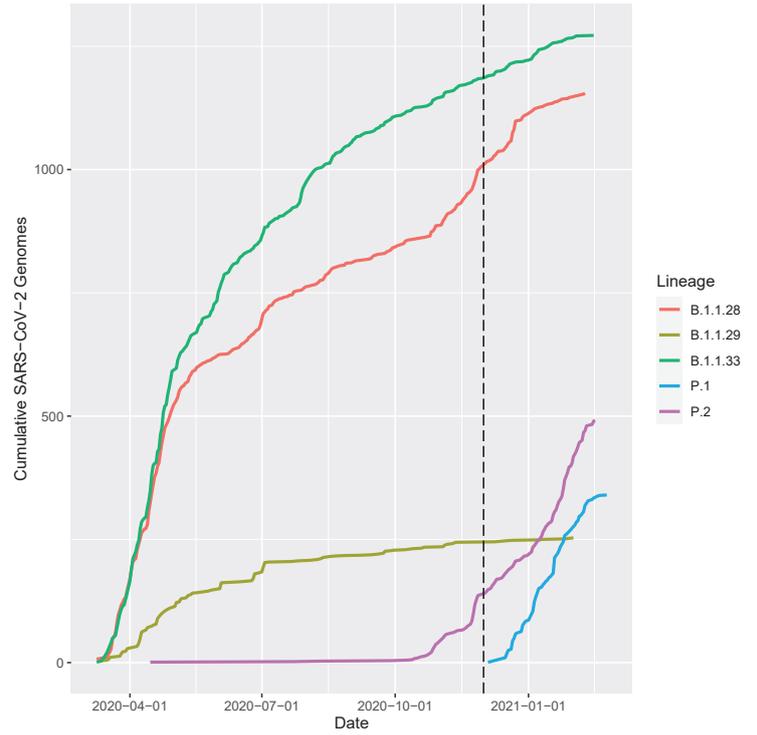
A

medRxiv preprint doi: <https://doi.org/10.1101/2021.06.25.21259565>; this version posted June 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY-NC-ND 4.0 International license](https://creativecommons.org/licenses/by-nc-nd/4.0/).

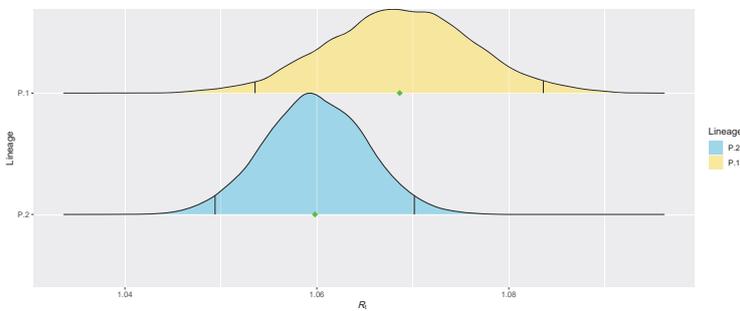
Distribution of lineages for SARS-CoV-2 isolates in South Africa



B



C



D

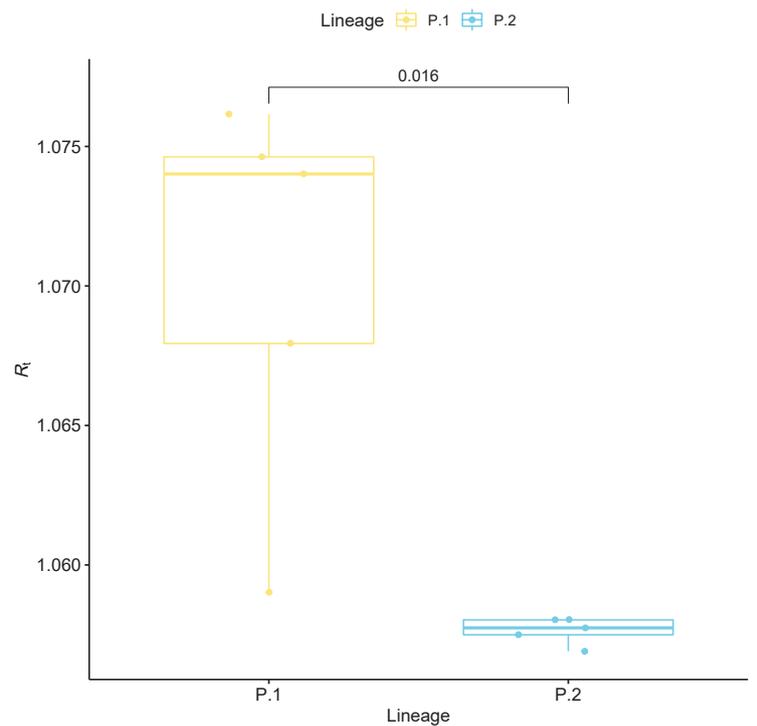
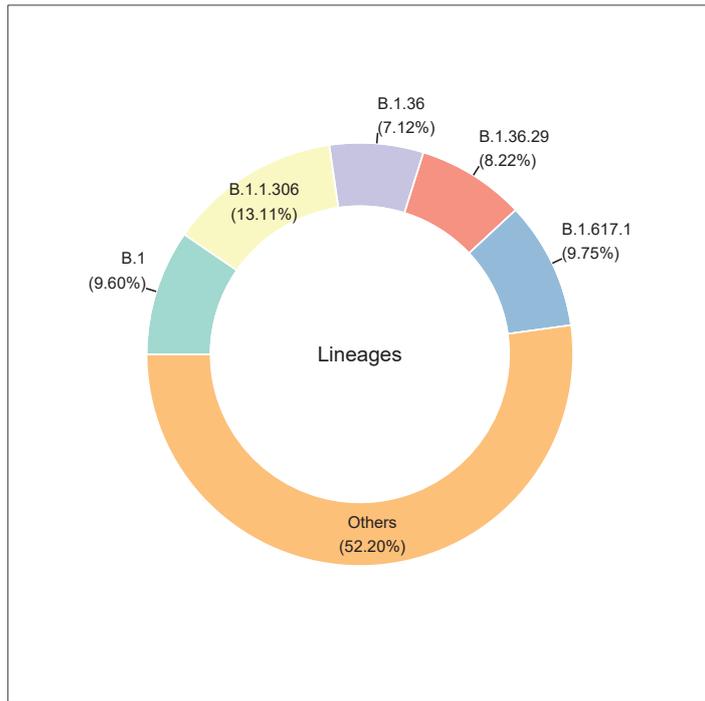


Figure 5

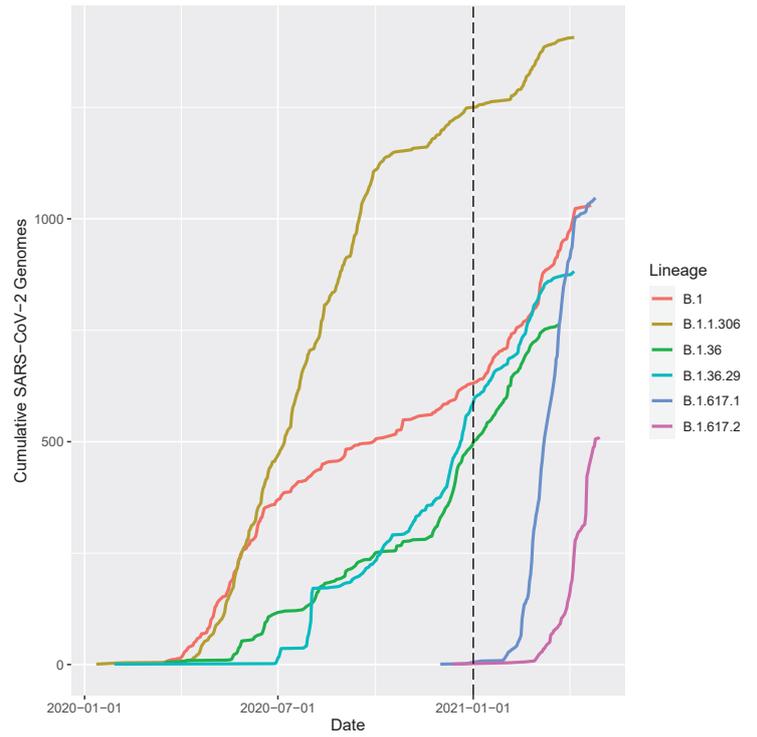
A

medRxiv preprint doi: <https://doi.org/10.1101/2021.06.25.21259565>; this version posted June 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY-NC-ND 4.0 International license](https://creativecommons.org/licenses/by-nc-nd/4.0/).

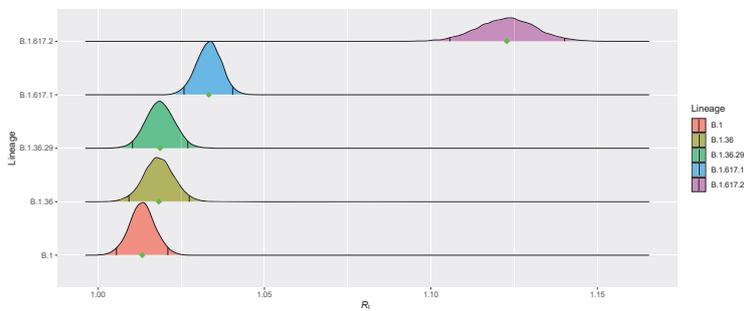
Distribution of lineages for SARS-CoV-2 isolates in India



B



C



D

