

1

# 2 The emergence of SARS-CoV-2 variants of concern is driven 3 by acceleration of the evolutionary rate

4 John H. Tay, Ashleigh F. Porter, Wytamma Wirth, and Sebastian Duchene\*.

5 Peter Doherty Institute for Infection and Immunity, University of Melbourne, Melbourne, Australia.

6 \*email: [sduchene@unimelb.edu.au](mailto:sduchene@unimelb.edu.au)

## 7 Abstract

8 The ongoing SARS-CoV-2 pandemic has seen an unprecedented amount of rapidly generated genome  
9 data. These data have revealed the emergence of lineages with mutations associated to transmissibility  
10 and antigenicity, known as variants of concern (VOCs). A striking aspect of VOCs is that many of them  
11 involve an unusually large number of defining mutations. Current phylogenetic estimates of the evolutionary  
12 rate of SARS-CoV-2 suggest that its genome accrues around 2 mutations per month. However, VOCs can  
13 have around 15 defining mutations and it is hypothesised that they emerged over the course of a few months,  
14 implying that they must have evolved faster for a period of time. We analysed genome sequence data from the  
15 GISAID database to assess whether the emergence of VOCs can be attributed to changes in the evolutionary  
16 rate of the virus and whether this pattern can be detected at a phylogenetic level using genome data. We fit a  
17 range of molecular clock models and assessed their statistical fit. Our analyses indicate that the emergence of  
18 VOCs is driven by an episodic increase in the evolutionary rate of around 4-fold the background phylogenetic  
19 rate estimate that may have lasted several weeks or months. These results underscore the importance of  
20 monitoring the molecular evolution of the virus as a means of understanding the circumstances under which  
21 VOCs may emerge.

22 **Keywords:** SARS-CoV-2 molecular evolution, variants of concern, molecular clock, Bayesian model  
23 selection.

## 24 1 The molecular clock of SARS-CoV-2

25 Genome sequence data of viruses have been extensively used to track the evolution and spread of these  
26 pathogens. The ongoing SARS-CoV-2 pandemic has seen an unprecedented number of genomes generated  
27 that have been used to gain rapid insight into epidemiological spread (Dellicour et al., 2021), identify the time  
28 of origin (Pekar et al., 2021), and to track mutations of functional importance. Most concerning mutations

29 occur in the spike protein and may increase transmissibility (Kraemer et al., 2021), or disease severity (Harvey  
30 et al., 2021), although vaccines are likely still effective against them (Dearlove et al., 2020). Such lineages  
31 are known as variants of concern (VOCs) and they are characterised at a genomic level by a number of fixed  
32 mutations in the S1 subunit of the spike protein, the most common of which are mutations N501Y and D614G  
33 (Eurosurveillance, 2021), with the the latter presenting evidence increased transmissibility and favoured by  
34 selection (Volz et al., 2021). For a lineage to be formally classified as a VOC there must be evidence of an  
35 impact in transmissibility, virulence, and/or immunity (Mascola et al., 2021).

36 SARS-CoV-2 lineages are classified using a dynamic nomenclature system, known as PANGO (Rambaut  
37 et al., 2020). Recently the World Health Organisation assigned variants of concern letters of the greek  
38 alphabet (Konings et al., 2021). At present the United States CDC recognises four variants of concern;  
39 Alpha (PANGO lineage B.1.1.7) first identified in the UK, Beta (PANGO lineage B.1.351) first identified in  
40 South Africa, Gamma (PANGO lineage P.1) first identified in Brazil, and Delta (PANGO lineage B.1.617.2)  
41 first identified in India (CDC, 2021).

42 The mechanisms under which VOCs have emerged is not entirely clear, but their defining mutations are  
43 well characterised. Variant Alpha has 14 protein-altering mutations and three deletions, with eight of these  
44 being in the spike protein. One of the deletions  $\Delta$ H69/ $\Delta$ V70 enhances infectivity in vitro and has been  
45 detected in immunocompromised patients where immune escape occurred (Kemp et al., 2021, Plante et al.,  
46 2021). Variant Beta has nine protein-altering mutations with five altering the receptor binding domain.  
47 (Tegally et al., 2021). Variant Gamma has 17 mutations, with 10 found in the spike protein and including  
48 N501Y and E484K (Faria et al., 2021). Importantly, Alpha, Beta and Gamma share several important  
49 mutations, including N501Y and E404K, which likely enhance affinity to human the ACE2 receptor (Nelson  
50 et al., 2021). Variant Delta is characterised by 7 mutations in the spike protein, several of which have been  
51 associated with altered immune response and increased viral replication, viral load, likely leading to increased  
52 viral fitness (CDC, 2021).

53 The sheer number of mutations observed in these four VOCs is much higher than what would be expected  
54 under phylogenetic estimates of the nucleotide evolutionary rate of SARS-CoV-2, which range from around  
55  $7 \times 10^{-4}$  to  $1.1 \times 10^{-3}$  subs/site/year (Duchene et al., 2020, Ghafari et al., 2021), meaning that only about 2  
56 mutations would accumulate per month along a lineage. In these circumstances, the 14 mutations in Alpha  
57 would require a period of at least six months, a time that is inconsistent with its first detection in September  
58 2020, because it would have had to evolve from around March 2020 with most defining mutations undetected

59 for many months.

## 60 1.1 Bayesian molecular clock models

61 We investigated whether the emergence of variants of concern is associated with an increase in the evolutionary  
62 rate that can be detected using phylogenetic analyses of genome data and in the absence of dense intrahost  
63 or transmission chain sampling. To this end, we analysed publicly available nucleotide sequence data from  
64 GISAID (Elbe and Buckland-Merrett, 2017, Shu and McCauley, 2017) under a range of molecular clock  
65 models that describe the evolutionary rate along branches in phylogenetic trees, shown in the Supplementary  
66 material. We consider each model as a hypothesis for which we can assess statistical support using Bayesian  
67 model selection techniques. Critically, our analyses do not intend to detect signatures of natural selection, nor  
68 to identify genomic regions with higher mutation rates, which have been described elsewhere (Abdool Karim  
69 and de Oliveira, 2021, Harvey et al., 2021). Instead, our framework serves to characterise the main patterns  
70 of evolutionary rate variation in the genome of the virus that underpin the emergence of VOCs.

71 The simplest molecular clock model is known as strict molecular clock (SC; Zuckerkandl, 1962, Zuck-  
72 erkandl and Pauling, 1965) that posits a single evolutionary rate for all branches in a phylogenetic tree, and  
73 thus serves as a 'null' model. A more complex model is the uncorrelated relaxed clock that assumes that  
74 branch rates are independent and identically distributed draws from a statistical distribution (Drummond  
75 et al., 2006), for which we considered either a lognormal or a  $\Gamma$  distribution (UCLN and UCG, respectively).

76 We also considered a range of fixed local clock models (FLC; Yoder and Yang, 2000). These models require  
77 an *a priori* definition of a set of 'background' branches and a set of branches with different rates, known as  
78 'foreground'. For example, foreground branches can be defined based on some biological expectation (e.g.  
79 Worobey et al., 2014), and thus represent a formal evolutionary hypothesis. The evolutionary rate is constant  
80 for a given group of branches, although there exist approaches where branches can be assigned a range of  
81 relaxed molecular clocks (Fourment and Darling, 2018). These models differ in their number of parameters  
82 and biological assumptions (Supplementary material Table S1; reviewed in Bromham et al., 2018 and Ho and  
83 Duchêne, 2014).

84 We specified six configurations of the FLC model, where the evolutionary rate could vary within VOC  
85 clades (FLC clades model in Supplementary material Figure S1) or along the stem (FLC stems+clades), only  
86 at stem branches (FLC stems), or where these rates could be shared among all VOCs (FLC shared stems,  
87 FLC shared clades and FLC shared clades+stems in Supplementary material Figure S1).

88 Models in which the rate only changes along the stem branches of VOCs represent a situation where  
89 the evolutionary rate may increase for a short period of time before returning to the background rate. In  
90 contrast, models where the clade also undergoes a rate change would imply that VOCs have a rate that is  
91 statistically different from the background.

92 An alternative approach to the FLC is the random local clock (RLC; Drummond and Suchard, 2010).  
93 The evolutionary rate can change at particular nodes in the tree and the location of such changes and actual  
94 rates are inferred. The RLC is a general form of all local clock models, where the simplest form is the strict  
95 clock, as a case of no rate changes (Bromham et al., 2018, Ho and Duchêne, 2014).

## 96 **1.2 Bayesian hypothesis testing**

97 We conducted model testing by calculating the log marginal likelihood, a measure of statistical fit, and  
98 ranking the models accordingly. The difference in log marginal likelihoods between two models is known as  
99 the log Bayes factor (Sinsheimer et al., 1996) and measures the relative support for two models given the  
100 data. In general, a log Bayes factor of at least 1.1 is considered as 'substantial evidence' in favour of a model,  
101 with 2.3 being 'strong' and 4.6 'decisive' (Kass and Raftery, 1995). We considered two marginal likelihood  
102 estimators, path sampling and stepping-stone sampling Gelman and Meng (1998), Lartillot and Philippe  
103 (2006), Xie et al. (2011).

## 104 **2 Results**

### 105 **2.1 Model selection**

106 The FLC shared stems model had the highest statistical fit, with a log Bayes factor of over 3 compared to  
107 the next best-fitting model (3.85 with path-sampling and 3.97 with stepping-stone sampling; Table 1). The  
108 next model with highest log marginal likelihood was the FLC stems. These two models assume that the stem  
109 branches of VOC have a rate that differs from the background and they only differ in that the FLC stems  
110 model allows each VOC stem branch to have its own rate.

111 The uncorrelated relaxed clocks had very similar performance, but ranked well below the best model,  
112 with a log Bayes factor of at least -6 relative to the FLC shared stems model 1. The log Bayes factors for  
113 the remaining models were at least -15, implying 'strong' evidence against them, relative to the FLC shared  
114 stems.

115 Interestingly, FLC models where VOC clades were defined as foreground had decisively lower statistical  
 116 performance than those where only stem branches were labelled as foreground (Table 1). In fact, even the  
 117 SC model, which is generally considered unrealistic for empirical data, had a log Bayes factor of at least 4  
 118 with respect to FLC shared clades and the FLC clades+stems Table 1.

Table 1: Model selection results for complete genomes. Estimates of log marginal likelihoods using path sampling and stepping-stone (ps logML and ss logML, respectively). log Bayes factors (BF) are shown for the best-fitting model, relative to all others (larger numbers mean lower statistical fit), and thus they are 0.0 for the top model.

Model	ps logML	ss logML	ps rank	ss rank	ps BF	ss BF
FLC shared stems	-55427.65	-55428.17	1	1	0	0
FLC stems	-55431.50	-55432.14	2	2	-3.85	-3.97
UCG	-55433.64	-55434.26	3	3	-6.00	-6.01
UCLN	-55434.32	-55434.69	4	4	-6.67	-6.52
FLC shared clades+stems	-55443.30	-55443.50	5	5	-15.64	-15.34
SC	-55443.53	-55444.21	6	6	-15.88	-16.04
FLC shared clades	-55449.89	-55450.52	7	7	-22.23	-22.35
FLC clades+stems	-55453.91	-55454.58	8	8	-26.25	-26.41
FLC	-55461.87	-55462.54	9	9	-34.21	-34.38

## 119 2.2 Rates of evolution of variants of concern

120 The FLC shared stems model had a mean background evolutionary rate of  $0.58 \times 10^{-3}$  subs/site/year (95%  
 121 CI:  $0.51 - 0.65 \times 10^{-3}$ ), while that for the VOC stems was  $2.45 \times 10^{-3}$  subs/site/year (95% CI:  $1.15 - 4.72 \times 10^{-3}$ ).  
 122 As such, the VOC stems rate was around 4 fold higher than the background (mean 4.25, 95% CI: 2.61 - 8.19)  
 123 (Fig 1).

124 Although the FLC stems model that assigned each VOC stem branch a different rate had very high  
 125 uncertainty, it also suggested much higher rates for these branches. The mean background rate under this  
 126 model was  $0.55 \times 10^{-3}$  subs/site/year (95% CI:  $0.49 - 0.62 \times 10^{-3}$ ). The corresponding values for VOC were  
 127  $8.47 \times 10^{-3}$  subs/site/year (95% CI:  $1.93 - 82.37 \times 10^{-3}$ ) for Alpha,  $1.71 \times 10^{-3}$  (95% CI:  $0.34 - 33.20 \times 10^{-3}$ ) for  
 128 Beta,  $2.76 \times 10^{-3}$  (95% CI:  $1.21 - 13.23 \times 10^{-3}$ ) for Gamma, and  $1.54 \times 10^{-3}$  (95% CI:  $0.62 - 7.35 \times 10^{-3}$ ) for Delta.  
 129 Clearly, these estimates were several fold higher than that of the background branches, and in spite of their  
 130 high uncertainty least 0.90 of the posterior density was above the mean background rate (Fig 1).

131 The coefficient of rate variation for both relaxed clock models, UCG and UCLN, was indicative of depart-  
 132 ure from clocklike evolution in the data. To investigate whether VOC stem branch rates differed from the  
 133 rest, we extracted individual branch rates and compared the VOC stem branch rates to the mean of all other

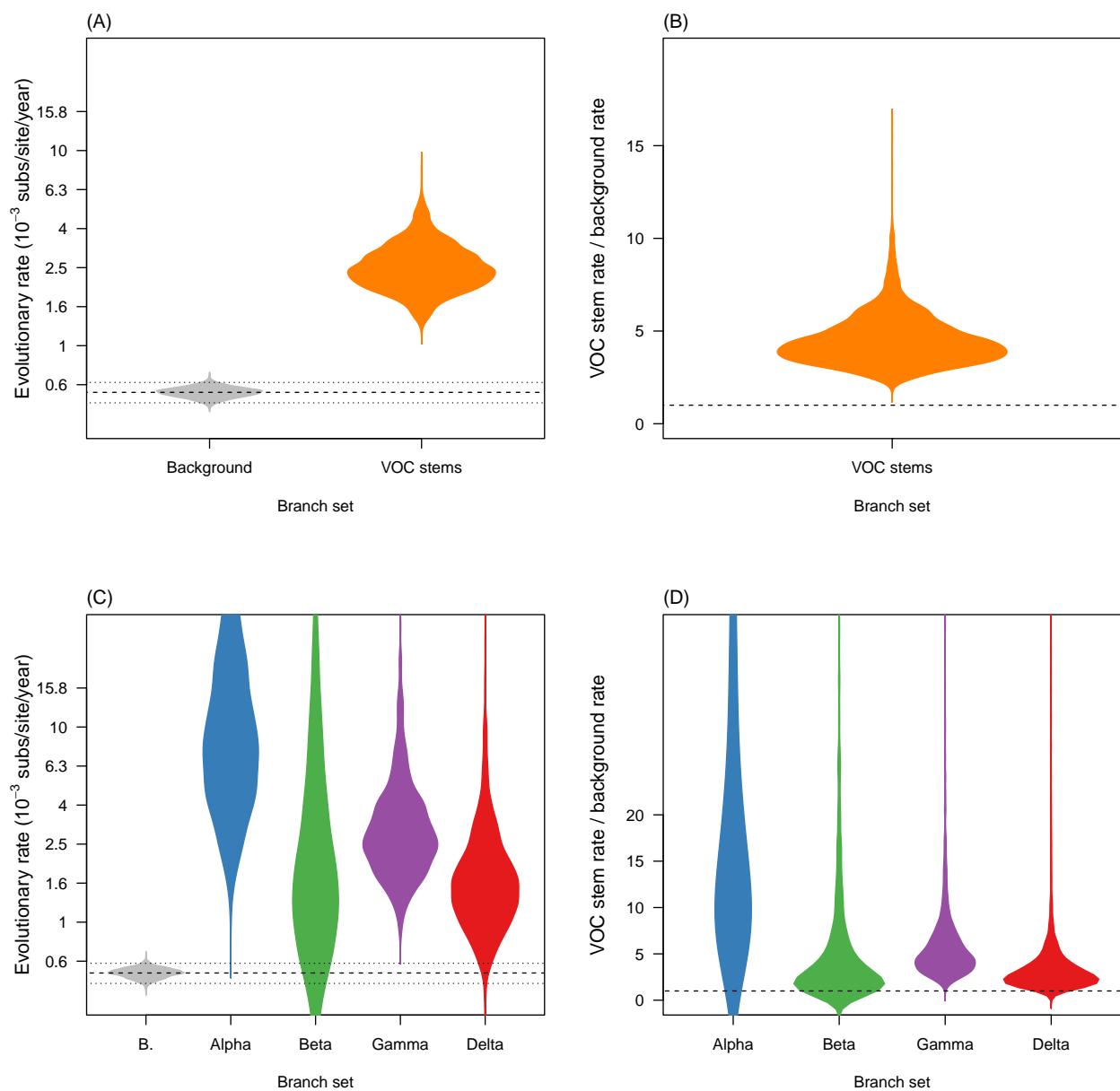


Figure 1: Violin plots for posterior statistics of fixed local clock models (FLC). (A) is for a FLC where the stem branches of VOCs share an evolutionary rate that is different to that of the background (model 'FLC shared stems' in Supplementary material Table S1 and Fig S1). The evolutionary rate for variants of concern (VOC) stem branches is shown in orange and the background in grey. The dashed line represents the mean background rate and the dotted lines are the 95% credible interval. (B) is the ratio of the evolutionary rate for VOC stem branches and the background under the same model and the dashed line represents a value of 1.0 where the background and VOC stem rate would be the same. (C) and (D) show the corresponding statistics for the FLC stems model, where the stem branch of every VOC has a different rate. Abbreviation 'B.' stands for background.

134 branches. We found evidence that VOC stem branch rates were higher than the mean of other branches,  
135 with higher means values, but very high uncertainty and 95% credible intervals that overlapped with the  
136 mean of other branches (Fig 2).

137 The mean evolutionary rate of branches other than the VOC stems was  $0.65 \times 10^{-3}$  subs/site/year (95%  
138 CI:  $0.58 - 0.77 \times 10^{-3}$ ) in the UCLN and  $0.69 \times 10^{-3}$  subs/site/year (95% CI:  $0.60 - 0.80 \times 10^{-3}$ ) for the UCG. In  
139 contrast, the VOC stem mean evolutionary rates for the UCLN were:  $1.29 \times 10^{-3}$  subs/site/year (95% credible  
140 interval, CI:  $0.76 - 2.56 \times 10^{-3}$ ) for Alpha,  $0.64 \times 10^{-3}$  (95% CI:  $0.32 - 1.57 \times 10^{-3}$ ) for Beta,  $1.29 \times 10^{-3}$  ( $0.82$   
141  $- 2.40 \times 10^{-3}$ ) for Gamma, and  $1.06 \times 10^{-3}$  (95% CI:  $0.50 - 2.38 \times 10^{-3}$ ) for Delta, and with comparable values  
142 for the UCG. The percentile where VOC stems rates fell with respect to other branches also supported the  
143 finding that their rates were particularly high in most cases. In the UCLN, for Alpha 0.96 of posterior density  
144 had the stem rate in the top 75% of fastest evolving branches, with the corresponding numbers for the other  
145 VOCs being 0.25, 0.98, and 0.81 Beta, Gamma, and Delta, respectively, and with comparable values in the  
146 UCG (0.92, 0.45, 0.96, and 0.91).

147 The RLC model produced less clear results than the other molecular clock models. The maximum  
148 *a posteriori* number of rate changes was 4, with the 95% CI ranging between 2 and 5. However, the  
149 posterior probability of rate changes in VOC stem branches or clades was 0.0. Instead, rate changes were  
150 not consistently found on particular branches. It is conceivable that this model poses a heavy penalty on  
151 rate changes. In particular, there is a very large number of local clock configurations in these data, which  
152 may be impossible to visit under this model and may result in low statistical power to assess support for our  
153 hypotheses here. This model, however, had an evolutionary rate estimate that was comparable to that of  
154 other models (mean  $0.60 \times 10^{-3}$  subs/site/year; 95% CI:  $0.49 - 0.72 \times 10^{-3}$ ).

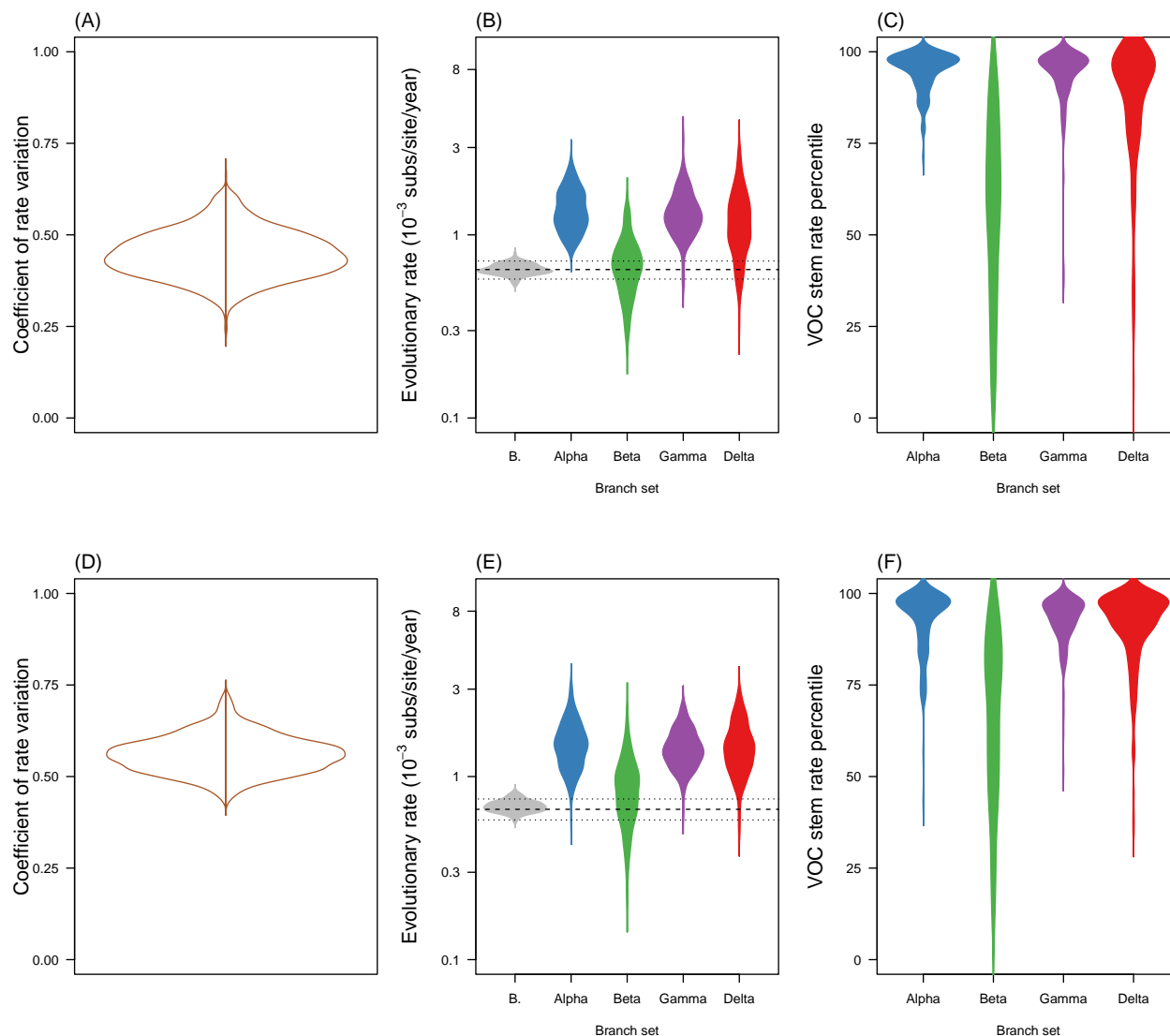


Figure 2: Violin plots of posterior statistics for the uncorrelated relaxed clocks with lognormal (UCLN) and gamma (UCG) distributions (see Supplementary material). The top row, (A) through (C), is for the UCLN and the bottom row, (D) through (F), is for the UCG. (A) and (D) show the coefficient of rate variation, which is the standard deviation of branch rates divided by the mean rate, and indicates clocklike behaviour when it is abutting zero (Drummond et al., 2006, Ho et al., 2015). In (B) and (E) the evolutionary rate is shown for the stem branches of variants of concern (VOC) and for the mean of background branches (i.e. those that are not the stems of VOCs), abbreviated as 'B.'. The dashed line denotes the mean background rate, while the dotted lines represent the upper and lower 95% credible interval. (C) and (F) shows the percentile in which stem branches for VOCs fall with respect to other branches. Note that the densities have been smoothed, but the maximum values are 100.



## 155 **2.3 Emergence time and expected genome substitutions**

156 We estimated the duration of time along VOC stem branches and the inferred total number of nucleotide  
157 substitutions along the complete genome. We focus on the best fitting model (FLC shared stems), with  
158 similar results for the second best model (FLC stems). The duration of time along these branches represents  
159 the time required before VOCs started to diversify, but it is important to note that they are contingent on  
160 sampling bias, and could therefore be shorter than estimated here. Under the FLC shared stems model, the  
161 stem branch leading up to VOs were; 14 weeks (95% CI:6 - 24) for Alpha, 4 (95% CI: 2 - 8) for Beta, 17  
162 (95% CI: 8 - 28) for Gamma, and 6 (3 - 11) for Delta (Supplementary material Fig S2).

163 The expected number of substitutions along the complete genome were; 21 (95% CI: 14 - 32) for Alpha,  
164 6 (95% CI:3 - 11) for Beta, 26 (95% CI: 18 - 35) for Gamma, and 9 (95% CI: 6 - 16) for Delta. Although,  
165 these numbers are loosely associated with the defining mutations, they are not directly comparable because  
166 they involve substitutions along the entire genome and they correspond to the inference from a standard  
167 phylogenetic substitution model (the GTR+ $\Gamma$  in this case).

## 168 **3 Discussion**

169 Our mean rate estimates over all lineages are somewhat lower than earlier estimates (Duchene et al., 2020),  
170 which is consistent with the notion that the virus has had time to evolve and remove transient deleterious  
171 mutations since its emergence (Ghafari et al., 2021). However, the molecular evolutionary rate of SARS-CoV-  
172 2 displays substantial variation among lineages, a pattern that has been apparent since early phylogenetic  
173 analyses of the virus (Duchene et al., 2020). Evolutionary rate variation is sometimes stochastic in nature  
174 and pinpointing its causes is often difficult in empirical data.

175 Our explicit hypothesis testing framework suggest that the emergence of VOCs explains much of the  
176 evolutionary rate variation in the virus. This model testing framework has been previously used to understand  
177 viral evolution among host species in influenza (Worobey et al., 2014), and the host range SARS-CoV-2 and  
178 closely related viruses (MacLean et al., 2021). We suggest that model testing may be preferable to using  
179 highly parametric models, such as relaxed molecular clock models for this purpose, because they tend to  
180 have very high variance in parameters of interest, such as evolutionary rates of particular branches. Recent  
181 advances in random local clock models may provide increased sensitivity (Fisher et al., 2021) of this family  
182 of models.

183 We find compelling evidence that episodic, instead of long-term, increases in the evolutionary rate un-

184 derpin the emergence of VOCs, a process that is probably driven the action of natural selection. All models  
185 where VOC clades were assigned a different rate to the background had poor statistical fit, even when com-  
186 pared to the SC 'null' model, providing further support for such rate increases to occur over a short period  
187 of time. The increase in evolutionary rate required to give rise to the four VOCs examined was estimated  
188 to be around 4-fold compared to the background, although such estimates may carry high uncertainty when  
189 estimated for individual stem branches. Under these circumstances the number of mutations required to  
190 give rise to a VOC, such as Alpha, would have accumulated in about three months, with some variants  
191 requiring less a few weeks, such as Beta and Delta. These timescales appear plausible in chronic infections of  
192 SARS-CoV-2 (Harvey et al., 2021, Kemp et al., 2021), but other circumstances are also likely, such as when  
193 transmission low and selection favours mutations that increase transmissibility between hosts.

194 Our genomic analyses demonstrate that signatures of increased evolutionary rates are detectable using  
195 phylogenetic methods and genome surveillance data, but the precise mechanism (ecological or intrahost) of  
196 how VOCs have emerged is still unclear. To elucidate these processes will require dense sampling between  
197 transmission chains, specifically in settings where transmission is unlikely and intra-host sequence data is  
198 available. An other important area that is currently under intense investigation is how natural selection  
199 shapes the emergence and persistence of VOCs (Tegally et al., 2021). Such studies may benefit from using  
200 explicit codon evolution models and careful partitioning among genes to model mutational heterogeneity. We  
201 recommend that further research focuses on early detection and understanding of the circumstances under  
202 which viral lineages with epidemiological impacts, such as VOCs, emerge.

## 203 4 Materials and methods

### 204 4.1 Data set construction

205 We downloaded 100 randomly selected sequences from the latest global NextStrain SARS-CoV-2 build (Had-  
206 field et al., 2018), from the GISAID database (Elbe and Buckland-Merrett, 2017, Shu and McCauley, 2017).  
207 This set of sequences did not include any of those belonging to the four VOCs (Alpha, Beta, Gamma, or  
208 Delta) and we also excluded samples drawn from non human hosts. We downloaded 20 randomly selected  
209 sequences from the four VOCs to generate a data set of 180 genomes, which we aligned using MAFFT (Ka-  
210 toh and Standley, 2013). Importantly, we ensured that the sequences consisted of complete genomes, with  
211 no stretches of more than 10 Ns and excluding those with low coverage (see Supplementary material). To

212 verify that samples classified as VOCs were correctly assigned as such, we estimated a phylogenetic tree using  
213 maximum likelihood as implemented in IQ-TREE2 (Minh et al., 2020), using the GTR+ $\Gamma$  substitution model  
214 and with approximate Bayes branch support (Anisimova et al., 2011). We ensured that all VOC samples  
215 that were monophyletic with other VOC samples with an approximate Bayes support  $<0.95$ .

## 216 4.2 Bayesian phylogenetic analyses

217 Our Bayesian analyses require specifying a substitution model, a tree, prior, priors for all parameters in  
218 BEAST 1.10 (Suchard et al., 2018). We chose the GTR+ $\Gamma_4$  substitution model and a coalescent exponential  
219 tree prior. Although the tree prior is not necessarily realistic here, it is expected to have little impact in  
220 molecular clock estimates Ritchie et al. (2017). It can accommodate changes in population size via the  
221 exponential growth function and it is fully parametric, such that setting proper priors for all parameters is  
222 possible. To calibrate the molecular clock we specified the sequence sampling times. The FLC models require  
223 constraining monophyly in VOCs, which we also did for other clock models to ensure that the prior on tree  
224 topology was the same.

225 We used the default priors for the substitution model. The coalescent exponential tree prior has two  
226 parameters, the scaled population size,  $\Phi$ , and the growth rate  $r$ . The scaled population size is proportional  
227 to the number of infected individuals at present divided by the twice the coalescent rate,  $\lambda$ , (i.e.  $\Phi = \frac{I(0)}{2\lambda}$ )  
228 and the growth rate is inversely proportional to the doubling time by a factor of  $\log(2)$  (*doubling time* =  
229  $\frac{\log(2)}{r}$ ) (Boskova et al., 2014, Volz et al., 2009). We used priors with relatively low information content for  
230 these two parameters. For  $\Phi$  we used an exponential distribution with mean  $10^5$ , while for  $r$  we used a  
231 Laplace distribution with location 0 and scale 100. For all molecular clock rates we used a continuous-time  
232 Markov chain reference prior (Ferreira and Suchard, 2008). The UCLN and UCG models have an additional  
233 parameter; the standard deviation of the lognormal distribution, and the shape of the  $\Gamma$  distribution. For  
234 these parameters we specified an exponential prior with mean 0.33. We ran our analyses for using a Markov  
235 chain Monte Carlo of length  $5 \times 10^7$ , sampling every  $5 \times 10^3$  and discarding 10% of the chain as burn-in. We  
236 repeated the analyses once to verify convergence of independent chains and we ensured that the effective  
237 sample size of all parameters was at least 200.

### 238 4.3 Marginal likelihood estimation

239 We used two techniques to infer the log marginal likelihood; path-sampling and stepping-stone (Gelman and  
240 Meng, 1998, Lartillot and Philippe, 2006, Xie et al., 2011), which have been found to have high performance  
241 in differentiating models in phylogenetics (Baele et al., 2012, 2013, Fourment et al., 2020), reviewed by  
242 Baele and Lemey (2014), Oaks et al. (2019). We chose these estimators over the more recently developed  
243 and highly accurate generalised stepping-stone because it requires a working genealogical distribution (Baele  
244 et al., 2016), which is not trivial here due to the monophyletic constraints. Our estimation setup had 200  
245 path steps distributed according to quantiles from a  $\beta$  distribution with parameter  $\alpha=0.3$ , with each of the  
246 resulting 201 power posterior inferences running for  $10^6$  iterations. We repeated these analyses three times  
247 to assess their variance. Our model testing approach considered the UCLN, SC, and all FLC models in Table  
248 1 and Supplementary material. We did not calculate log marginal likelihoods for the RLC because this is a  
249 model averaging method, where the number of parameters is less tractable than in other models. As a result  
250 it is difficult to conceive proper priors for all parameters, which is a fundamental aspect of Bayesian model  
251 selection.

## 252 5 Acknowledgements

253 This work was supported by the Australian Research Council (DE190100805) and the Medical Research  
254 Future Fund (MRF9200006). This research was undertaken using the LIEF HPC-GPGPU Facility hosted at  
255 the University of Melbourne. This Facility was established with the assistance of LIEF Grant LE170100200.  
256 We acknowledge efforts by originating and submitting laboratories for the sequence data in GISAID EpiCoV  
257 on which our analyses are based. We are also grateful to Prof. Edward Holmes for useful suggestions and  
258 comments on ideas developed in this study.

## 259 References

- 260 S. S. Abdool Karim and T. de Oliveira. New sars-cov-2 variants—clinical, public health, and vaccine impli-  
261 cations. *New England Journal of Medicine*, 384(19):1866–1868, 2021.
- 262 M. Anisimova, M. Gil, J.-F. Dufayard, C. Dessimoz, and O. Gascuel. Survey of branch support methods  
263 demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Systematic*  
264 *Biology*, 60(5):685–699, 2011.

- 265 G. Baele and P. Lemey. Bayesian model selection in phylogenetics and genealogy-based population genetics.  
266 In M. Chen, K. L, and L. PO, editors, *Bayesian phylogenetics, methods, algorithms, and applications*,  
267 chapter 4, pages 59–93. CPC Press, Boca Raton (Florida), 2014.
- 268 G. Baele, P. Lemey, T. Bedford, A. Rambaut, M. A. Suchard, and A. V. Alekseyenko. Improving the accuracy  
269 of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty.  
270 *Molecular Biology and Evolution*, 29(9):2157–2167, 2012.
- 271 G. Baele, P. Lemey, and S. Vansteelandt. Make the most of your samples: Bayes factor estimators for  
272 high-dimensional models of sequence evolution. *BMC Bioinformatics*, 14(1):1–18, 2013.
- 273 G. Baele, P. Lemey, and M. A. Suchard. Genealogical working distributions for bayesian model testing with  
274 phylogenetic uncertainty. *Systematic Biology*, 65(2):250–264, 2016.
- 275 V. Boskova, S. Bonhoeffer, and T. Stadler. Inference of epidemiological dynamics based on simulated phylo-  
276 genies using birth-death and coalescent models. *PLoS Computational Biology*, 10(11):e1003913, 2014.
- 277 L. Bromham, S. Duchêne, X. Hua, A. M. Ritchie, D. A. Duchêne, and S. Y. Ho. Bayesian molecular dating:  
278 opening up the black box. *Biological Reviews*, 93(2):1165–1191, 2018.
- 279 CDC. Sars-cov-2 variant classifications and definitions, 2021. URL [https://www.cdc.gov/coronavirus/](https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html)  
280 [2019-ncov/variants/variant-info.html](https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html).
- 281 B. Dearlove, E. Lewitus, H. Bai, Y. Li, D. B. Reeves, M. G. Joyce, P. T. Scott, M. F. Amare, S. Vasani,  
282 N. L. Michael, et al. A sars-cov-2 vaccine candidate would likely match all currently circulating variants.  
283 *Proceedings of the National Academy of Sciences*, 117(38):23652–23662, 2020.
- 284 S. Dellicour, K. Durkin, S. L. Hong, B. Vanmechelen, J. Martí-Carreras, M. S. Gill, C. Meex, S. Bontems,  
285 E. André, M. Gilbert, et al. A phylodynamic workflow to rapidly gain insights into the dispersal history  
286 and dynamics of sars-cov-2 lineages. *Molecular biology and evolution*, 38(4):1608–1613, 2021.
- 287 A. J. Drummond and M. A. Suchard. Bayesian random local clocks, or one rate to rule them all. *BMC*  
288 *Biology*, 8(1):1–12, 2010.
- 289 A. J. Drummond, S. Y. W. Ho, M. J. Phillips, and A. Rambaut. Relaxed phylogenetics and dating with  
290 confidence. *PLoS Biology*, 4(5):e88, 2006.

- 291 S. Duchene, L. Featherstone, M. Haritopoulou-Sinanidou, A. Rambaut, P. Lemey, and G. Baele. Temporal  
292 signal and the phylodynamic threshold of sars-cov-2. *Virus evolution*, 6(2):veaa061, 2020.
- 293 S. Elbe and G. Buckland-Merrett. Data, disease and diplomacy: Gisaïd’s innovative contribution to global  
294 health. *Global Challenges*, 1(1):33–46, 2017.
- 295 Eurosurveillance. Updated rapid risk assessment from ecdc on the risk related to the spread of new sars-cov-2  
296 variants of concern in the eu/eea–first update. *Eurosurveillance*, 26(3):2101211, 2021.
- 297 N. R. Faria, T. A. Mellan, C. Whittaker, I. M. Claro, D. d. S. Candido, S. Mishra, M. A. Crispim, F. C.  
298 Sales, I. Hawryluk, J. T. McCrone, et al. Genomics and epidemiology of the p. 1 sars-cov-2 lineage in  
299 manaus, brazil. *Science*, 372(6544):815–821, 2021.
- 300 M. A. Ferreira and M. A. Suchard. Bayesian analysis of elapsed times in continuous-time markov chains.  
301 *Canadian Journal of Statistics*, 36(3):355–368, 2008.
- 302 A. A. Fisher, X. Ji, A. Nishimura, P. Lemey, and M. A. Suchard. Shrinkage-based random local clocks with  
303 scalable inference. *arXiv preprint arXiv:2105.07119*, 2021.
- 304 M. Fourment and A. E. Darling. Local and relaxed clocks: the best of both worlds. *PeerJ*, 6:e5140, 2018.
- 305 M. Fourment, A. F. Magee, C. Whidden, A. Bilge, F. A. Matsen IV, and V. N. Minin. 19 dubious ways to  
306 compute the marginal likelihood of a phylogenetic tree topology. *Systematic Biology*, 69(2):209–220, 2020.
- 307 A. Gelman and X.-L. Meng. Simulating normalizing constants: From importance sampling to bridge sampling  
308 to path sampling. *Statistical Science*, pages 163–185, 1998.
- 309 M. Ghafari, L. du Plessis, J. Raghvani, S. Bhatt, B. Xu, O. Pybus, and A. Katzourakis. Purifying selec-  
310 tion determines the short-term time dependency of evolutionary rates in sars-cov-2 and ph1n1 influenza.  
311 *medRxiv*, 2021.
- 312 J. Hadfield, C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, and  
313 R. A. Neher. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, 34(23):4121–4123,  
314 2018.
- 315 W. T. Harvey, A. M. Carabelli, B. Jackson, R. K. Gupta, E. C. Thomson, E. M. Harrison, C. Ludden,  
316 R. Reeve, A. Rambaut, S. J. Peacock, et al. Sars-cov-2 variants, spike mutations and immune escape.  
317 *Nature Reviews Microbiology*, 19(7):409–424, 2021.

- 318 S. Y. Ho and S. Duchêne. Molecular-clock methods for estimating evolutionary rates and timescales. *Molecular*  
319 *Ecology*, 23(24):5947–5965, 2014.
- 320 S. Y. Ho, S. Duchêne, and D. Duchêne. Simulating and detecting autocorrelation of molecular evolutionary  
321 rates among lineages. *Molecular Ecology Resources*, 15(4):688–696, 2015.
- 322 R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):  
323 773–795, 1995.
- 324 K. Katoh and D. M. Standley. Mafft multiple sequence alignment software version 7: improvements in  
325 performance and usability. *Molecular Biology and Evolution*, 30(4):772–780, 2013.
- 326 S. A. Kemp, D. A. Collier, R. P. Datir, I. A. Ferreira, S. Gayed, A. Jahun, M. Hosmillo, C. Rees-Spear,  
327 P. Mlcochova, I. U. Lumb, et al. Sars-cov-2 evolution during treatment of chronic infection. *Nature*, 592  
328 (7853):277–282, 2021.
- 329 F. Konings, M. D. Perkins, J. H. Kuhn, M. J. Pallen, E. J. Alm, B. N. Archer, A. Barakat, T. Bedford, J. N.  
330 Bhiman, L. Caly, et al. Sars-cov-2 variants of interest and concern naming scheme conducive for global  
331 discourse. *Nature Microbiology*, pages 1–3, 2021.
- 332 M. U. Kraemer, V. Hill, C. Ruis, S. Dellicour, S. Bajaj, J. T. McCrone, G. Baele, K. V. Parag, A. L. Battle,  
333 B. Gutierrez, et al. Spatiotemporal invasion dynamics of sars-cov-2 lineage b. 1.1. 7 emergence. *Science*,  
334 373(6557):889–895, 2021.
- 335 N. Lartillot and H. Philippe. Computing bayes factors using thermodynamic integration. *Systematic Biology*,  
336 55(2):195–207, 2006.
- 337 O. A. MacLean, S. Lytras, S. Weaver, J. B. Singer, M. F. Boni, P. Lemey, S. L. Kosakovsky Pond, and D. L.  
338 Robertson. Natural selection in the evolution of sars-cov-2 in bats created a generalist virus and highly  
339 capable human pathogen. *PLoS Biology*, 19(3):e3001115, 2021.
- 340 J. R. Mascola, B. S. Graham, and A. S. Fauci. Sars-cov-2 viral variants—tackling a moving target. *Jama*,  
341 325(13):1261–1262, 2021.
- 342 B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. Von Haeseler, and R. Lanfear.  
343 Iq-tree 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology*  
344 *and Evolution*, 37(5):1530–1534, 2020.

- 345 G. Nelson, O. Buzko, P. R. Spilman, K. Niazi, S. Rabizadeh, and P. R. Soon-Shiong. Molecular dynamic  
346 simulation reveals e484k mutation enhances spike rbd-ace2 affinity and the combination of e484k, k417n  
347 and n501y mutations (501y. v2 variant) induces conformational change greater than n501y mutant alone,  
348 potentially resulting in an escape mutant. *BioRxiv*, 2021.
- 349 J. Oaks, K. F. Cobb, V. N. Minin, and A. D. Leaché. Marginal likelihoods in phylogenetics: a review of  
350 methods and applications. *Systematic Biology*, 68(5):681–697, 2019.
- 351 J. Pekar, M. Worobey, N. Moshiri, K. Scheffler, and J. O. Wertheim. Timing the sars-cov-2 index case in  
352 hubei province. *Science*, 372(6540):412–417, 2021.
- 353 J. A. Plante, Y. Liu, J. Liu, H. Xia, B. A. Johnson, K. G. Lokugamage, X. Zhang, A. E. Muruato, J. Zou,  
354 C. R. Fontes-Garfias, et al. Spike mutation d614g alters sars-cov-2 fitness. *Nature*, 592(7852):116–121,  
355 2021.
- 356 A. Rambaut, E. C. Holmes, Á. O’Toole, V. Hill, J. T. McCrone, C. Ruis, L. du Plessis, and O. G. Pybus. A  
357 dynamic nomenclature proposal for sars-cov-2 lineages to assist genomic epidemiology. *Nature Microbiology*,  
358 5(11):1403–1407, 2020.
- 359 A. M. Ritchie, N. Lo, and S. Y. Ho. The impact of the tree prior on molecular dating of data sets containing  
360 a mixture of inter-and intraspecies sampling. *Systematic Biology*, 66(3):413–425, 2017.
- 361 Y. Shu and J. McCauley. Gisaid: Global initiative on sharing all influenza data—from vision to reality.  
362 *Eurosurveillance*, 22(13):30494, 2017.
- 363 J. S. Sinsheimer, J. A. Lake, and R. J. Little. Bayesian hypothesis testing of four-taxon topologies using  
364 molecular sequence data. *Biometrics*, pages 193–210, 1996.
- 365 M. A. Suchard, P. Lemey, G. Baele, D. L. Ayres, A. J. Drummond, and A. Rambaut. Bayesian phylogenetic  
366 and phylodynamic data integration using beast 1.10. *Virus Evolution*, 4(1):vey016, 2018.
- 367 H. Tegally, E. Wilkinson, M. Giovanetti, A. Iranzadeh, V. Fonseca, J. Giandhari, D. Doolabh, S. Pillay, E. J.  
368 San, N. Msomi, et al. Detection of a sars-cov-2 variant of concern in south africa. *Nature*, 592(7854):  
369 438–443, 2021.
- 370 E. Volz, V. Hill, J. T. McCrone, A. Price, D. Jorgensen, Á. O’Toole, J. Southgate, R. Johnson, B. Jackson,



- 371 F. F. Nascimento, et al. Evaluating the effects of sars-cov-2 spike mutation d614g on transmissibility and  
372 pathogenicity. *Cell*, 184(1):64–75, 2021.
- 373 E. M. Volz, S. L. Kosakovsky Pond, M. J. Ward, A. J. Leigh Brown, and S. D. Frost. Phylodynamics of  
374 infectious disease epidemics. *Genetics*, 183(4):1421–1430, 2009.
- 375 M. Worobey, G.-Z. Han, and A. Rambaut. A synchronized global sweep of the internal genes of modern  
376 avian influenza virus. *Nature*, 508(7495):254–257, 2014.
- 377 W. Xie, P. O. Lewis, Y. Fan, L. Kuo, and M.-H. Chen. Improving marginal likelihood estimation for bayesian  
378 phylogenetic model selection. *Systematic Biology*, 60(2):150–160, 2011.
- 379 A. D. Yoder and Z. Yang. Estimation of primate speciation dates using local molecular clocks. *Molecular*  
380 *Biology and Evolution*, 17(7):1081–1090, 2000.
- 381 E. Zuckerkandl. Molecular disease, evolution, and genetic heterogeneity. *Horizons in biochemistry*, pages  
382 189–225, 1962.
- 383 E. Zuckerkandl and L. Pauling. Evolutionary divergence and convergence in proteins. In *Evolving genes and*  
384 *proteins*, pages 97–166. Elsevier, 1965.