August 6, 2021

The emergence of SARS-CoV-2 variants of concern is driven by episodic acceleration of the genomic rate of molecular evolution

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

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

*email: sduchene@unimelb.edu.au

Abstract

The ongoing SARS-CoV-2 pandemic has seen an unprecedented amount of rapidly generated genome data. These data have revealed the emergence of lineages with mutations associated to transmissibility and antigenicity, known as variants of concern (VOCs). A striking aspect of VOCs is that many of them involve a high number of defining mutations. Current phylogenetic estimates of the evolutionary rate of SARS-CoV-2 suggest that its genome accrues around 2 mutations per month. However, VOCs can have around 15 defining mutations and it is hypothesised that they emerged over the course of a few months, implying that the evolutionary rate would be several fold higher. A plausible scenario that is difficult to demonstrate empirically is that such rapid evolution has occurred within immunocompromised patients over a short period of time. We analysed genome sequence data from the GISAID database to assess whether the emergence of VOCs can be attributed to changes in the evolutionary rate of the virus and whether this pattern can be detected at a phylogenetic level using genome data. We fit a range of molecular clock models and assessed their statistical fit. Our analyses indicate that the emergence of VOCs is driven by an episodic increase in the evolutionary rate of around 6-fold the background phylogenetic rate estimate. Our results underscore the importance of monitoring the molecular evolution of the virus as a means of understanding the circumstances under which VOCs may emerge.

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

1 The molecular clock of SARS-CoV-2

Genome sequence data of viruses have been extensively used to track the evolution and spread of these pathogens. The ongoing SARS-CoV-2 pandemic has seen an unprecedented number of genomes generated,

which have revealed the emergence of lineages with a large number of mutations. Many of the mutations occur in the spike protein which increase transmissibility or disease severity, or which may produce a reduction in neutralising antibodies from previous infection or vaccination (Abdool Karim and de Oliveira, 2021). Such lineages are known as variants of concern (VOCs) and they are characterised at a genomic level by a number of fixed mutations in the S1 subunit of the spike protein, the most common of which are mutations N501Y and D614G (Team et al., 2021). For a lineage to be formally classified as a VOC there must be evidence of increased transmissibility, virulence, and/or immunity (Mascola et al., 2021). SARS-CoV-2 lineages are classified using a dynamic nomenclature system, known as PANGO (Rambaut et al., 2020). Recently the World Health Organisation assigned variants of concern letters of the greek alphabet (Konings et al., 2021). At present the United States CDC recognises four variants of concern; Alpha (PANGO lineage B.1.1.7) first identified in the UK, Beta (PANGO lineage B.1.351) first identified in South Africa, Gamma (PANGO lineage P.1) first identified in Brazil, and Delta (PANGO lineage B.1.617.2) first identified in India (of Disease Control, 2021). The mechanisms under which VOCs have emerged is not entirely clear, particularly because they have a large number of protein altering mutations (Harvey et al., 2021). Variant Alpha has 14 protein-altering mutations and three deletions, with eight of these being in the spike protein. One of the deletions ΔH69/ΔV70 enhances infectivity in vitro and has been detected in immunocompromised patients where immune escape occurred (Kemp et al., 2021, Plante et al., 2021). Variant Beta has nine protein-altering mutations with five altering the receptor binding domain. (Tegally et al., 2021). Variant Gamma has 17 mutations, with 10 found in the spike protein and including N501Y and E484K (Faria et al., 2021). Importantly, Alpha, Beta and Gamma share several important mutations, including N501Y and E404K, which likely enhance affinity to human the ACE2 receptor (Nelson et al., 2021). Variant Delta is characterised by 7 mutations in the spike protein, several of which have been associated with altered immune response and increased viral replication, viral load, and thus transmission (Lopez Bernal et al., 2021). The sheer number of genetic changes observed in these variants is much higher than what would be expected under phylogenetic estimates of the nucleotide evolutionary rate of SARS-CoV-2, which range from around 7×10^{-4} to 1.1×10^{-3} subs/site/year (Duchene et al., 2020), meaning that only about 2 mutations would accumulate per month along a lineage. In these circumstances, the 14 mutations in Alpha would require a period of at least six months, a time period that is inconsistent with its first detection in September 2020, because it would have had to evolve from around March 2020 with several mutations undetected for many months. There is compelling evidence to suggest that variants likely emerged in chronic infection within immunocompromised hosts during treatment with

convalescent plasma therapy (Kemp et al., 2021). In influenza models, prolonged infection has been shown to accelerate the evolutionary rate because selection has had more time and opportunities to shape viral evolution (Xue et al., 2017, 2018).

1.1 Bayesian molecular clock models

We investigated whether the emergence of variants of concern is associated with an increase in the evolutionary rate that can be detected using phylogenetic analyses of genome data and in the absence of intrahost sequences. To this end, we analysed publicly available nucleotide sequence data from GISAID (Elbe and Buckland-Merrett, 2017, Shu and McCauley, 2017) under a range of molecular clock models that describe the evolutionary rate along branches in phylogenetic trees, shown in Table 1 and illustrated in Figure 1. We consider each model as a hypothesis for which we can assess statistical support using Bayesian model selection techniques. Critically, our analyses do not intend to detect signatures of natural selection, nor to identify genomic regions with higher mutation rates. Instead, our framework serves to characterise the main patterns of evolutionary rate variation in the genome of the virus that underpin the emergence of VOCs. The simplest molecular clock model is known as strict molecular clock (SC; Zuckerkandl, 1962, Zuckerkandl and Pauling, 1965) that posits a single evolutionary rate for all branches, and thus serves as a null model here. A more complex model is the uncorrelated relaxed clock with an underlying lognormal distribution (UCLN; Drummond et al., 2006) that assumes that branch rates are independent and identically distributed draws from a lognormal distribution. We also considered a range of fixed local clock models (FLC; Yoder and Yang, 2000). These models require an *a priori* definition of a set of 'background' branches and set of branches with different rates, known as 'foreground' branches. For example, foreground branches can be defined based on some biological expectation (e.g. Worobey et al., 2014), and thus represent a formal evolutionary hypothesis. The evolutionary rate is constant for a given group of branches, although there exist approaches where branches can be assigned a range of relaxed molecular clocks (Fourment and Darling, 2018). These models differ in their number of parameters and biological assumptions (Table 1).

We specified six configurations of the FLC model, where the evolutionary rate could vary within VOC clades (FLC clades model in Fig. 1) or along the stem (FLC stems+clades in Fig. 1), only at stem branches (FLC stems in Fig. 1), or where these rates could be shared among all VOCs (FLC shared stems, FLC shared clades and FLC shared clades+stems in Fig. 1).

Models in which the rate only changes along the stem branches of VOCs represent a situation where

Table 1: Molecular clock model configurations and parameterisation. The term "clades" correspond to monophyletic groups of either of the four variants of concern (VOC) in the data set (Alpha, Beta, Gamma, Delta) and the stems are the branches leading up to them. Note that "clock rate" refers to the global clock rates of the strict clock and "b. rate" is for the background rate.

Model family	**Model abbreviation**	**Parameters**	**Number of parameters**
Strict clock	SC	clock rate	1
Relaxed	UCLN	Num. branches, mean, s.d.	358+2
Relaxed	UGM	Num. branches, mean, shape	358+2
Relaxed	UCE	Num. branches, mean	358+1
Random local clock	RLC*	b. rate, num. rate changes	variable
Fixed local clock	FLC clades	b. rate, num. clades	1+4
Fixed local clock	FLC stems+clades	b. rate, num. clades & stems	1+4
Fixed local clock	FLC stems only	b. rate, num. stems	1+4
Fixed local clock	FLC shared stem	b. rate, stem rate	1+1
Fixed local clock	FLC shared clade	b. rate, clade rate	1+1
Fixed local clock	FLC shared clades+stems	b. rate, clade & stem rate	1+1

the evolutionary rate may increase for a short period of time before returning to the background rate. For example, if nucleotide mutations in VOCs have appeared within patients with prolonged infections (Davies et al., 2021, Kemp et al., 2021). In contrast, models where the clade also undergoes a rate change would imply that VOCs have a rate that is statistically different from the background.

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

1.2 Bayesian hypothesis testing

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

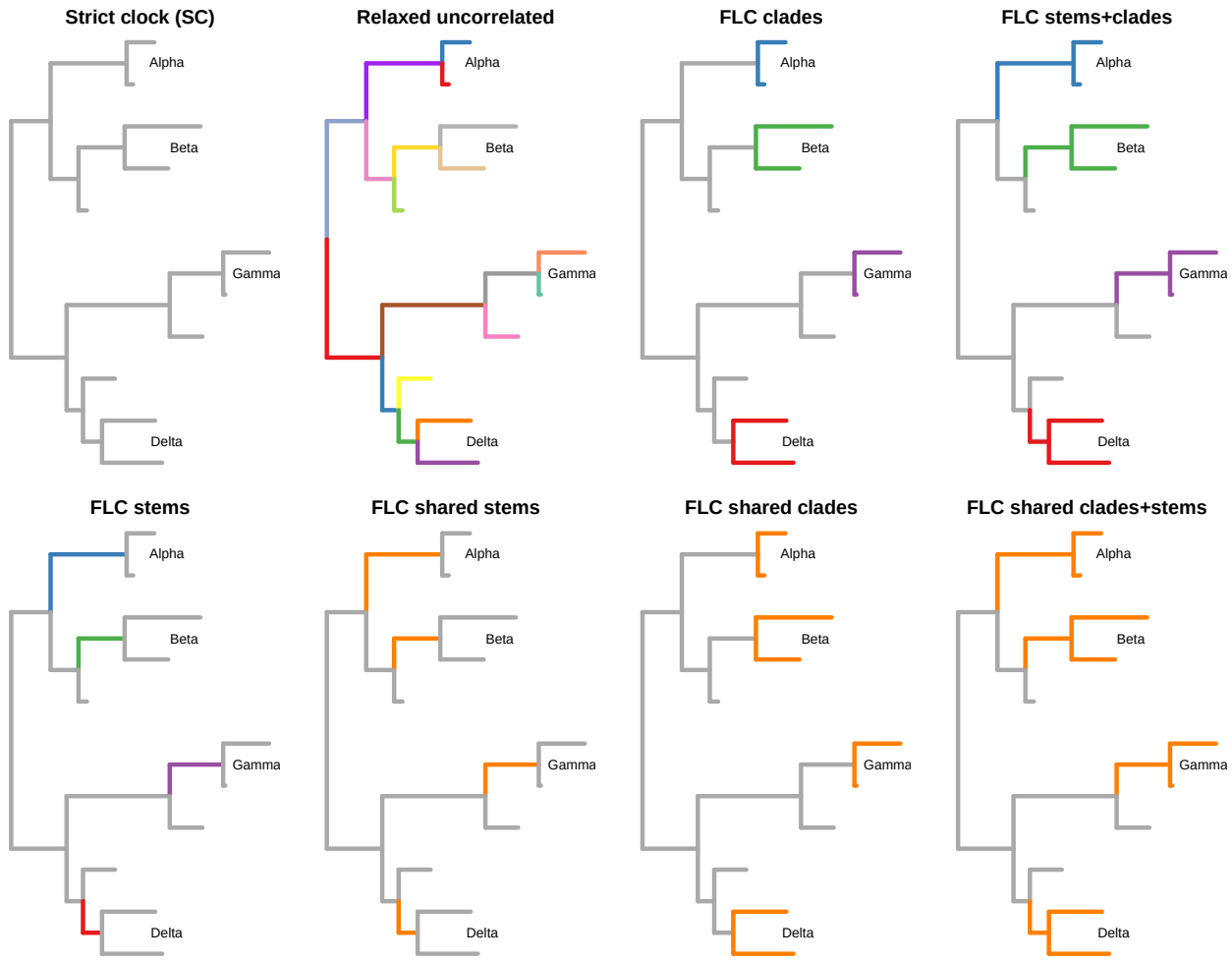

Figure 1: Illustration of molecular clock models considered in hypothetical trees with the four variants of concern (VOCs) and background genetic diversity. SC stands for strict clock and FLC for fixed local clock. Branches are coloured according to the rate assigned in each case, with grey corresponding to the "background" branches and those in colours being the "foreground" branches. Model names match those of Table 1.

2 Results

2.1 Model selection

The UCLN model had the highest statistical fit, with a log Bayes factor of over 27 compared to the next best-fitting model (27.297 with path-sampling and 27.484 with stepping-stone sampling; Table 2). The next model with highest log marginal likelihood was the FLC shared stems, followed by the FLC stems. These two models had very similar log marginal likelihoods that differed only by 1.1 and 1.4 log likelihood units, using path-sampling and stepping-stone, respectively. The next best-fitting model was the SC, with a decisively

lower log marginal likelihood. The log Bayes factor of the UCLN relative to the SC was 38.160 and 38.357 with path sampling and stepping stone, respectively.

Interestingly, FLC models where clades were defined as foreground had decisively lower statistical performance than those where only stem branches were labelled as foreground (Table 2). In fact, even the SC model, which is generally considered unrealistic for empirical data, had a log Bayes factor of at least 2 with respect to the FLC models with clades Table 2.

Table 2: Model selection results for complete genomes. Estimates of log marginal likelihoods using path sampling and stepping-stone (ps logML and ss logML, respectively). Bayes factors 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**	**ps BF**	**ss rank**	**ss BF**
UCLN	-54589.560	-54590.114	1	0.000	1	0.000
FLC shared stems	-54616.857	-54617.598	2	27.297	2	27.484
FLC stems	-54617.911	-54618.979	3	28.351	3	28.866
SC	-54627.720	-54628.471	4	38.160	4	38.357
FLC shared clades	-54630.514	-54631.420	5	40.954	5	41.306
FLC shared clades+stems	-54632.647	-54633.645	6	43.088	6	43.531
FLC	-54641.062	-54641.765	7	51.502	7	51.651
FLC clades+stems	-54646.807	-54647.603	8	57.247	8	57.489

2.2 Rates of evolution of variants of concern

The coefficient of rate variation of the UCLN model was indicative of departure from clocklike evolution in the data. To investigate whether VOC stem branch rates differed from the rest we extracted individual branch rates and compared the VOC stem branch rates to the mean of all other branches. We found evidence that VOC stem branch rates were higher than the mean of other branches, with higher means, but very high uncertainty and 95% credible intervals that overlapped with the mean of other branches (Fig 2). The mean evolutionary rate of branches other than the VOC stems was 0.74×10^{-3} subs/site/year (95% CI: 0.66 - 0.83×10^{-3}), slightly lower, but within the rage of early estimates of the global rate of the virus (Duchene et al., 2020). In contrast, the VOC stem mean evolutionary rates were: 1.21×10^{-3} subs/site/year (95% credible interval, CI: 0.52 - 2.2×10^{-3}) for Alpha , 1.42×10^{-3} (95% CI: 0.69 - 3.4×10^{-3}) for Beta, 2.05×10^{-3} (1.08 - 3.4×10^{-3}) for Gamma, and 1.05×10^{-3} (95% CI: 0.37 - 2.77×10^{-3}) for Delta. The percentile where VOC stems rates fell with respect to other branches also supported the finding that their rates were particularly high. For Alpha 0.73 of posterior density had the stem rate in the top 75% of fastest evolving branches, with the corresponding numbers for the other VOCs being 0.86, 0.94, and 0.67 Beta, Gamma, and Delta, respectively.

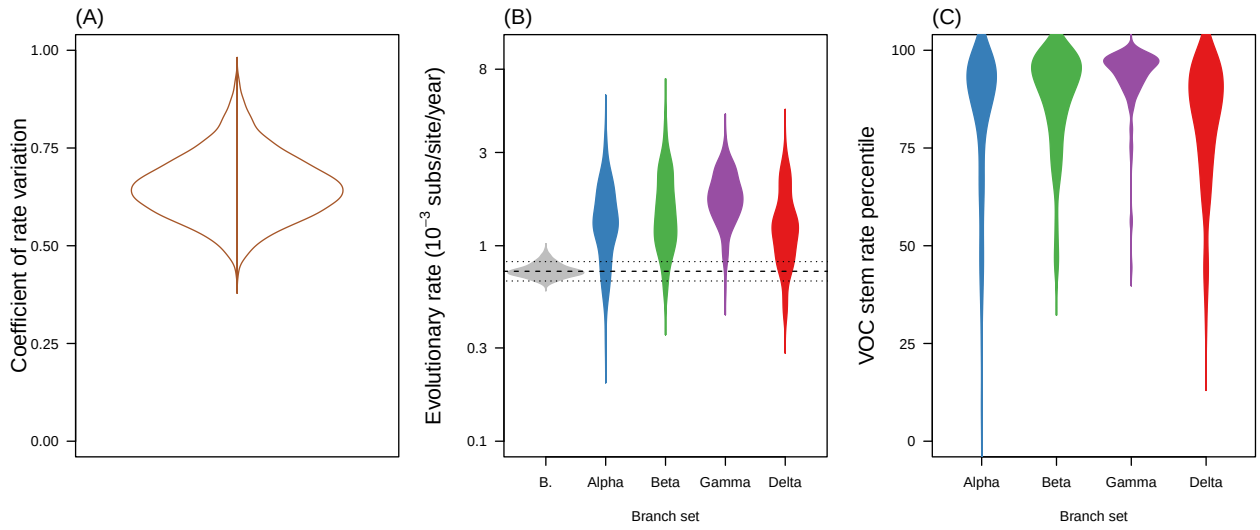

Figure 2: Violin plots of posterior statistics for the uncorrelated relaxed clock with underlying lognormal distribution (UCLN; see 1 and 1). (A) shows the coefficient of rate variation, which is the standard deviation of branch rates divided by the mean rate, and indicates clocklike behaviour when it is abutting zero (Drummond et al., 2006, Ho et al., 2015). In (B) 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) 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.

The best fitting FLC models that supported VOC stem rates differing from the background (FLC shared stems and FLC stems) suggested a marked increase in the rates for these branches. The FLC shared stems model had a mean background evolutionary rate of 0.66×10^{-3} subs/site/year (95% CI: 0.60 - 0.74×10^{-3}), while that for the VOC stems was 4.16×10^{-3} subs/site/year (95% CI: 2.10 - 11.36×10^{-3}). As such, the VOC stems rate was around 6 fold higher than the background (mean 6.24, 95% CI: 3.07 - 17.22) (Fig 3).

Although the FLC stems model that assigned each VOC stem branch a different rate had very high uncertainty, it also suggested much higher rates for these branches. The mean background rate under this model was 0.68×10^{-3} subs/site/year (95% CI: 0.60 - 0.0747×10^{-3}). The corresponding values for VOC were 10.40×10^{-3} subs/site/year (95% CI: 1.5 - 254.91×10^{-3}) for Alpha, 9.43×10^{-3} (95% CI: 1.57 - 246.33×10^{-3}) for Beta, 4.05×10^{-3} (95% CI: 1.55 - 54.58×10^{-3}) for Gamma, and 10.93×10^{-3} (95% CI: 1.61 - 167.66×10^{-3}) for Delta. Clearly, these estimates were several fold higher than that of the background branches and in spite of their high uncertainty least 0.98 of the posterior density was above the mean background rate (Fig 3).

The RLC model produced less clear results than the other molecular clock models. The maximum

a posteriori number of rate changes was 4, with the 95% CI ranging between 2 and 6. However, the posterior probability of rate changes in VOC stem branches or clades was 0.0. Instead, rate changes were not consistently found on particular branches. It is conceivable that this model poses a heavy penalty on rate changes, such that it lacked statistical power to assess support for our hypotheses here, especially noting that there is very high uncertainty in cases where a single VOC stem branch rate is estimated (e.g. FLC stems model; Fig 2). This model, however, had an evolutionary rate estimate that was comparable to that of other models (mean 0.74×10^{-3} subs/site/year; 95% CI: 0.76 - 0.83×10^{-3}).

3 Discussion

The molecular evolutionary rate of SARS-CoV-2 displays substantial variation among lineages, a pattern that has been apparent since early phylogenetic analyses of the virus (Duchene et al., 2020). Evolutionary rate variation is sometimes stochastic in nature and elucidating its causes is often difficult in empirical data. Our analyses explicitly test hypotheses by using our current knowledge of the impact and spread of VOCs. Such model testing framework has been previously used to understand viral evolution among host species in influenza (Worobey et al., 2014), and the host range and capabilities of SARS-CoV-2 (MacLean et al., 2021). We suggest that explicit model testing frameworks may be preferable to using highly parametric models, such as the UCLN, because they tend to have very high variance in parameters of interest, such as evolutionary rates of particular branches.

We find compelling evidence that episodic, instead of long-term, increases in the evolutionary rate underpin the emergence of VOCs. All models where VOC clades were assigned a different rate to the background had poor statistical fit, even when compared to the SC "null" model, providing further support for such rate increases to occur over a short period of time. The increase in evolutionary rate required to give rise to the four VOCs examined was estimated to be around 6-fold compared to the background, although such estimates may carry high uncertainty when estimated for individual stem branches. Under these circumstances the number of mutations required to give rise to a VOC, such as Alpha, would have accumulated in about one month, which appears plausible in chronic infections in SARS-CoV-2 (Harvey et al., 2021, Kemp et al., 2021).

Our genomic analyses demonstrate that these signatures of increased evolutionary rates are detectable using phylogenetic methods and genome surveillance data. A key requirement here is that multiple VOC genome samples are needed to define clades and their stems. In the absence of multiple genomes for a VOC

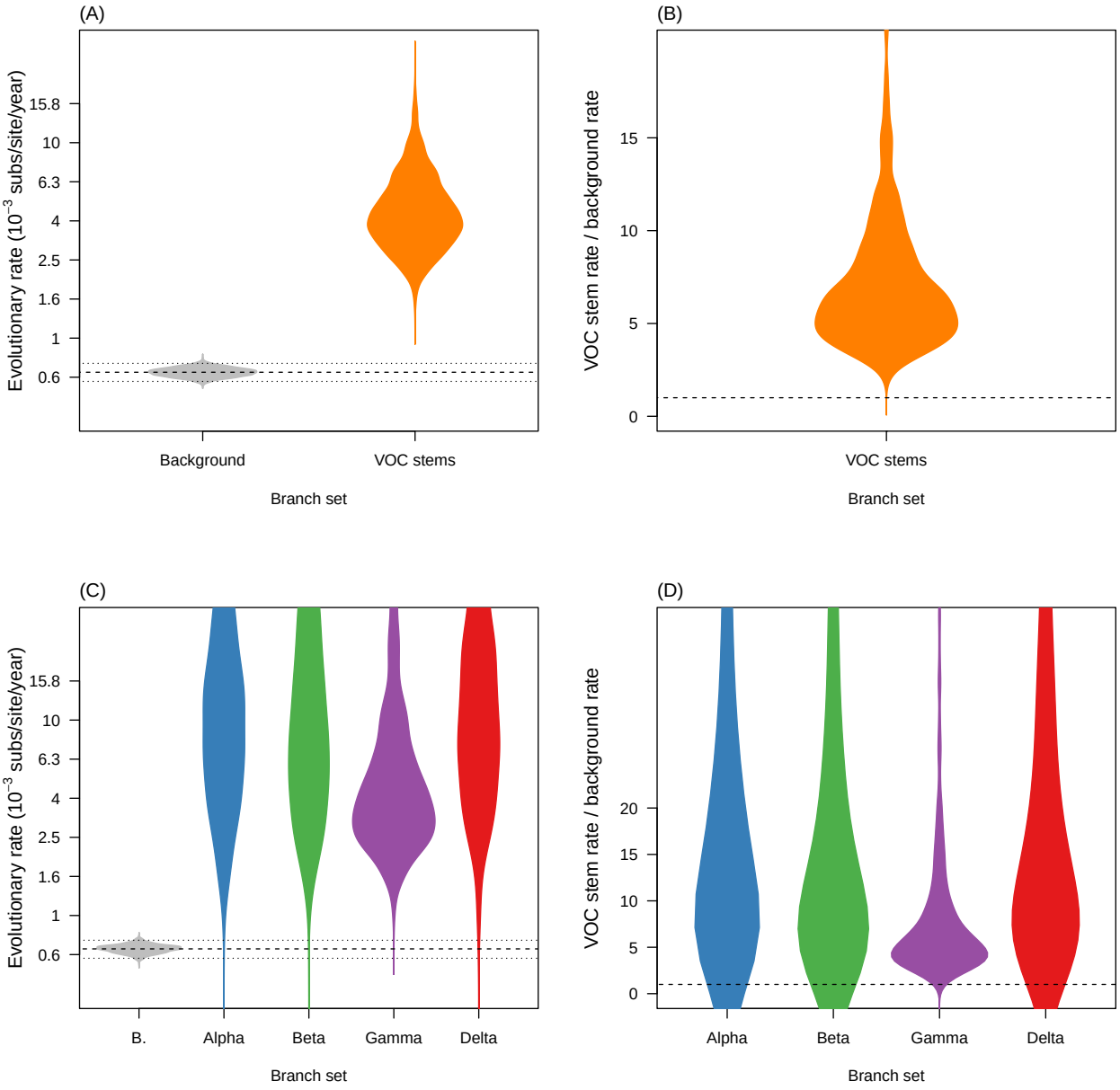

Figure 3: Violin plots for posterior statistics of fixed local clock models (FLC). (A) is for a FLC where the stem branches of VOCs share an evolutionary rate that is different to that of the background (model "FLC shared stems" in Table 1 and in Fig 1. The evolutionary rate for variants of concern (VOC) 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.

candidate only the terminal branches can be assessed, such that episodic evolution may not be detectable. Recent advances in molecular clock models may provide increased sensitivity (Fisher et al., 2021). An obvious area that should be explored is the action of natural selection in shaping VOCs. Such studies may benefit from using explicit codon evolution models, partitioning among genes to model mutational heterogeneity. Relaxing some of the fundamental assumptions of substitution models may also give more nuanced inferences given that the evolutionary process within hosts may be markedly different to that between hosts. We hope that our results will motivate such research to ultimately help the early detection and understanding of the circumstances under which viral lineages with epidemiological impacts emerge.

4 Materials and methods

4.1 Data set construction

We downloaded 100 randomly selected sequences from the latest global NextStrain SARS-CoV-2 build (Hadfield et al., 2018), from the GISAID database (Elbe and Buckland-Merrett, 2017, Shu and McCauley, 2017). This set of sequences did not include any of those belonging to the four VOCs (Alpha, Beta, Gamma, or Delta) and we also excluded samples drawn from non human hosts. We downloaded 20 randomly selected sequences from the four VOCs to generate a data set of 180 genomes, which we aligned using MAFFT (Katoh and Standley, 2013). To verify that samples classified as VOCs were correctly assigned as such, we estimated a phylogenetic tree using maximum likelihood as implemented in IQ-TREE2 (Minh et al., 2020), using the GTR+Γ substitution model and with approximate Bayes branch support (Anisimova et al., 2011). We ensured that all VOC samples that were monophyletic with other VOC samples with an approximate Bayes support <0.95.

4.2 Bayesian phylogenetic analyses

Our Bayesian analyses require specifying a substitution model, a tree, prior, priors for all parameters in BEAST 1.10 (Suchard et al., 2018). We chose the GTR+Γ_4 substitution model and a coalescent exponential tree prior. Although the tree prior is not necessarily realistic here, it is expected to have little impact in molecular clock estimates Ritchie et al. (2017). It can accommodate changes in population size via the exponential growth function and it is fully parametric, such that setting proper priors for all parameters is possible. To calibrate the molecular clock we specified the sequence sampling times. The FLC models require

constraining monophyly in VOCs, which we also did for other clock models to ensure that the prior on the tree topology was the same.

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

4.3 Marginal likelihood estimation

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

5 Acknowledgements

This work was supported by the Australian Research Council (DE190100805) and the Medical Research Future Fund (MRF9200006). This research was undertaken using the LIEF HPC-GPGPU Facility hosted at the University of Melbourne. This Facility was established with the assistance of LIEF Grant LE170100200. We also acknowledge efforts by originating and submitting laboratories for the sequence data in GISAID EpiCoV on which our analyses are based.

6 Supplementary material

List of accession numbers from GISAID: GISAID acknowledgements

References

S. S. Abdool Karim and T. de Oliveira. New sars-cov-2 variants—clinical, public health, and vaccine implications. *New England Journal of Medicine*, 384(19):1866–1868, 2021.

M. Anisimova, M. Gil, J.-F. Dufayard, C. Dessimoz, and O. Gascuel. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Systematic biology*, 60(5):685–699, 2011.

G. Baele and P. Lemey. Bayesian model selection in phylogenetics and genealogy-based population genetics. 2014.

G. Baele, P. Lemey, T. Bedford, A. Rambaut, M. A. Suchard, and A. V. Alekseyenko. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Molecular biology and evolution*, 29(9):2157–2167, 2012.

G. Baele, P. Lemey, and S. Vansteelandt. Make the most of your samples: Bayes factor estimators for high-dimensional models of sequence evolution. *BMC bioinformatics*, 14(1):1–18, 2013.

G. Baele, P. Lemey, and M. A. Suchard. Genealogical working distributions for bayesian model testing with phylogenetic uncertainty. *Systematic Biology*, 65(2):250–264, 2016.

V. Boskova, S. Bonhoeffer, and T. Stadler. Inference of epidemiological dynamics based on simulated phylogenies using birth-death and coalescent models. *PLoS Computational Biology*, 10(11):e1003913, 2014.

L. Bromham, S. Duchêne, X. Hua, A. M. Ritchie, D. A. Duchêne, and S. Y. Ho. Bayesian molecular dating: opening up the black box. *Biological Reviews*, 93(2):1165–1191, 2018.

N. G. Davies, S. Abbott, R. C. Barnard, C. I. Jarvis, A. J. Kucharski, J. D. Munday, C. A. Pearson, T. W. Russell, D. C. Tully, A. D. Washburne, et al. Estimated transmissibility and impact of sars-cov-2 lineage b. 1.1. 7 in england. *Science*, 372(6538), 2021.

A. J. Drummond and M. A. Suchard. Bayesian random local clocks, or one rate to rule them all. *BMC Biology*, 8(1):1–12, 2010.

A. J. Drummond, S. Y. W. Ho, M. J. Phillips, and A. Rambaut. Relaxed phylogenetics and dating with confidence. *PLoS Biology*, 4(5):e88, 2006.

S. Duchene, L. Featherstone, M. Haritopoulou-Sinanidou, A. Rambaut, P. Lemey, and G. Baele. Temporal signal and the phylodynamic threshold of sars-cov-2. *Virus evolution*, 6(2):veaa061, 2020.

S. Elbe and G. Buckland-Merrett. Data, disease and diplomacy: Gisaid's innovative contribution to global health. *Global challenges*, 1(1):33–46, 2017.

N. R. Faria, T. A. Mellan, C. Whittaker, I. M. Claro, D. d. S. Candido, S. Mishra, M. A. Crispim, F. C. Sales, I. Hawryluk, J. T. McCrone, et al. Genomics and epidemiology of the p. 1 sars-cov-2 lineage in manaus, brazil. *Science*, 372(6544):815–821, 2021.

M. A. Ferreira and M. A. Suchard. Bayesian analysis of elapsed times in continuous-time markov chains. *Canadian Journal of Statistics*, 36(3):355–368, 2008.

A. A. Fisher, X. Ji, A. Nishimura, P. Lemey, and M. A. Suchard. Shrinkage-based random local clocks with scalable inference. *arXiv preprint arXiv:2105.07119*, 2021.

M. Fourment and A. E. Darling. Local and relaxed clocks: the best of both worlds. *PeerJ*, 6:e5140, 2018.

M. Fourment, A. F. Magee, C. Whidden, A. Bilge, F. A. Matsen IV, and V. N. Minin. 19 dubious ways to compute the marginal likelihood of a phylogenetic tree topology. *Systematic Biology*, 69(2):209–220, 2020.

A. Gelman and X.-L. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, pages 163–185, 1998.

J. Hadfield, C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, and R. A. Neher. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, 34(23):4121–4123, 2018.

W. T. Harvey, A. M. Carabelli, B. Jackson, R. K. Gupta, E. C. Thomson, E. M. Harrison, C. Ludden, R. Reeve, A. Rambaut, S. J. Peacock, et al. Sars-cov-2 variants, spike mutations and immune escape. *Nature Reviews Microbiology*, 19(7):409–424, 2021.

S. Y. Ho and S. Duchêne. Molecular-clock methods for estimating evolutionary rates and timescales. *Molecular ecology*, 23(24):5947–5965, 2014.

S. Y. Ho, S. Duchêne, and D. Duchêne. Simulating and detecting autocorrelation of molecular evolutionary rates among lineages. *Molecular ecology resources*, 15(4):688–696, 2015.

R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430): 773–795, 1995.

K. Katoh and D. M. Standley. Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4):772–780, 2013.

S. A. Kemp, D. A. Collier, R. P. Datir, I. A. Ferreira, S. Gayed, A. Jahun, M. Hosmillo, C. Rees-Spear, P. Mlcochova, I. U. Lumb, et al. Sars-cov-2 evolution during treatment of chronic infection. *Nature*, 592 (7853):277–282, 2021.

F. Konings, M. D. Perkins, J. H. Kuhn, M. J. Pallen, E. J. Alm, B. N. Archer, A. Barakat, T. Bedford, J. N. Bhiman, L. Caly, et al. Sars-cov-2 variants of interest and concern naming scheme conducive for global discourse. *Nature Microbiology*, pages 1–3, 2021.

N. Lartillot and H. Philippe. Computing bayes factors using thermodynamic integration. *Systematic biology*, 55(2):195–207, 2006.

J. Lopez Bernal, N. Andrews, C. Gower, E. Gallagher, R. Simmons, S. Thelwall, J. Stowe, E. Tessier, N. Groves, G. Dabrera, et al. Effectiveness of covid-19 vaccines against the b. 1.617. 2 (delta) variant. *New England Journal of Medicine*, 2021.

O. A. MacLean, S. Lytras, S. Weaver, J. B. Singer, M. F. Boni, P. Lemey, S. L. Kosakovsky Pond, and D. L.

Robertson. Natural selection in the evolution of sars-cov-2 in bats created a generalist virus and highly capable human pathogen. *PLoS biology*, 19(3):e3001115, 2021.

J. R. Mascola, B. S. Graham, and A. S. Fauci. Sars-cov-2 viral variants—tackling a moving target. *Jama*, 325(13):1261–1262, 2021.

B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. Von Haeseler, and R. Lanfear. Iq-tree 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular biology and evolution*, 37(5):1530–1534, 2020.

G. Nelson, O. Buzko, P. R. Spilman, K. Niazi, S. Rabizadeh, and P. R. Soon-Shiong. Molecular dynamic simulation reveals e484k mutation enhances spike rbd-ace2 affinity and the combination of e484k, k417n and n501y mutations (501y. v2 variant) induces conformational change greater than n501y mutant alone, potentially resulting in an escape mutant. *BioRxiv*, 2021.

J. Oaks, K. F. Cobb, V. N. Minin, and A. D. Leaché. Marginal likelihoods in phylogenetics: a review of methods and applications. *Systematic Biology*, 68(5):681–697, 2019.

C. of Disease Control. Sars-cov-2 variant classifications and definitions, 2021. URL `https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html`.

J. A. Plante, Y. Liu, J. Liu, H. Xia, B. A. Johnson, K. G. Lokugamage, X. Zhang, A. E. Muruato, J. Zou, C. R. Fontes-Garfias, et al. Spike mutation d614g alters sars-cov-2 fitness. *Nature*, 592(7852):116–121, 2021.

A. Rambaut, E. C. Holmes, Á. O'Toole, V. Hill, J. T. McCrone, C. Ruis, L. du Plessis, and O. G. Pybus. A dynamic nomenclature proposal for sars-cov-2 lineages to assist genomic epidemiology. *Nature microbiology*, 5(11):1403–1407, 2020.

A. M. Ritchie, N. Lo, and S. Y. Ho. The impact of the tree prior on molecular dating of data sets containing a mixture of inter-and intraspecies sampling. *Systematic Biology*, 66(3):413–425, 2017.

Y. Shu and J. McCauley. Gisaid: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance*, 22(13):30494, 2017.

J. S. Sinsheimer, J. A. Lake, and R. J. Little. Bayesian hypothesis testing of four-taxon topologies using molecular sequence data. *Biometrics*, pages 193–210, 1996.

M. A. Suchard, P. Lemey, G. Baele, D. L. Ayres, A. J. Drummond, and A. Rambaut. Bayesian phylogenetic and phylodynamic data integration using beast 1.10. *Virus Evolution*, 4(1):vey016, 2018.

E. E. Team et al. Updated rapid risk assessment from ecdc on the risk related to the spread of new sars-cov-2 variants of concern in the eu/eea–first update. *Eurosurveillance*, 26(3):2101211, 2021.

H. Tegally, E. Wilkinson, M. Giovanetti, A. Iranzadeh, V. Fonseca, J. Giandhari, D. Doolabh, S. Pillay, E. J. San, N. Msomi, et al. Detection of a sars-cov-2 variant of concern in south africa. *Nature*, 592(7854): 438–443, 2021.

E. M. Volz, S. L. Kosakovsky Pond, M. J. Ward, A. J. Leigh Brown, and S. D. Frost. Phylodynamics of infectious disease epidemics. *Genetics*, 183(4):1421–1430, 2009.

M. Worobey, G.-Z. Han, and A. Rambaut. A synchronized global sweep of the internal genes of modern avian influenza virus. *Nature*, 508(7495):254–257, 2014.

W. Xie, P. O. Lewis, Y. Fan, L. Kuo, and M.-H. Chen. Improving marginal likelihood estimation for bayesian phylogenetic model selection. *Systematic biology*, 60(2):150–160, 2011.

K. S. Xue, T. Stevens-Ayers, A. P. Campbell, J. A. Englund, S. A. Pergam, M. Boeckh, and J. D. Bloom. Parallel evolution of influenza across multiple spatiotemporal scales. *Elife*, 6:e26875, 2017.

K. S. Xue, L. H. Moncla, T. Bedford, and J. D. Bloom. Within-host evolution of human influenza virus. *Trends in microbiology*, 26(9):781–793, 2018.

A. D. Yoder and Z. Yang. Estimation of primate speciation dates using local molecular clocks. *Molecular Biology and Evolution*, 17(7):1081–1090, 2000.

E. Zuckerkandl. Molecular disease, evolution, and genetic heterogeneity. *Horizons in biochemistry*, pages 189–225, 1962.

E. Zuckerkandl and L. Pauling. Evolutionary divergence and convergence in proteins. In *Evolving genes and proteins*, pages 97–166. Elsevier, 1965.