The phylodynamic threshold of measurably evolving populations

Ariane Weber[1,*], Sanni Översti[1], Julia Kende[2], and Sebastian Duchene[3,4,*].

[1] Pending.

[2] Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, Paris, France.

[3] ED-ID unit, Dept of Computational Biology, Institut Pasteur, Paris, France.

[4] Peter Doherty Institute for Infection and Immunity, Dept of Microbiology and Immunology, University of Melbourne, Melbourne, Australia.

*email: weber@gea.mpg.de, sduchene@pasteur.fr

Abstract

Pending.

Keywords: Measurably evolving populations, phylodynamic threshold, molecular clock, Bayesian phylogenetics, microbial evolution.

1 Introduction

Molecular sequence data have become nearly ubiquitous for studying the evolution of modern and ancient organisms. A fundamental concept in molecular evolution is the 'molecular clock', which posits that substitutions accumulate roughly constantly over time (Zuckerkandl and Pauling, 1965). An underlying assumption of the classic molecular clock is that selective constraints are negligible for most sites and over time. The development of molecular clock models as a statistical processes relax this and other assumptions by allowing for rate variation among branches in phylogenetic trees (reviewed by Ho and Duchêne (2014)).

Molecular clock models necessarily involve two key quantities, the evolutionary timescale and the 'evolutionary rate', with the latter representing the combination of mutations and substitutions that accrue over time. However, evolutionary times and rates cannot be jointly identified using genetic sequence data alone, that is they are undentifiable (Dos Reis and Yang (2013) and reviewed by Bromham et al. (2018)). To overcome this problem, all molecular clock methods require prior assumption about evolutionary times or rates, known as a 'molecular clock calibration'. For example, one can constraint the age of the common ancestor between two lineages (i.e. an internal node in a phylogenetic tree) to a given time or fix the evolutionary rate

to a known value. The choice of calibration depends on the information available and its reliability (Duchêne et al., 2014, Warnock et al., 2012).

1.1 Measurably evolving populations

Rapidly evolving organisms, notably viruses and some bacteria, have been found to accrue an appreciable number of mutations over the sampling timescale. Influenza viruses, for example, have evolutionary rates of around 6×10^{-3} subs/site/year (Ghafari et al., 2021, Sanjuán, 2012). Assuming a genome size of 13,500 Kb, one would expect to observe one mutation every 4 to 5 days ($\frac{365\ days/year}{13500\ sites\ \times\ 6{\times}10^{-3} subs/site/year} \approx 4.5\ days/subs$). If genome samples are collected over the course of a few weeks, then the sampling times themselves can be used to calibrate the molecular clock, a practice known as 'tip calibration' (reviewed by Rieux and Balloux (2016)). Data sets for which the molecular clock can be calibrated using sequence sampling times are considered to have been sampled from a 'measurably evolving population' (Drummond et al., 2003b).

Advances in sequencing technologies have dramatically expanded the range of organisms from which data sets can be considered to have been sampled from a measurably evolving population. First, ancient DNA techniques have effectively expanded the genome sampling window for many organisms (Duchene et al., 2020b, Spyrou et al., 2019). One of many examples, is mitochrondrial DNA recovered from dogs from 36,000 years before present (Thalmann et al., 2013), which was found to be sufficient to calibrate its molecular clock. Second, whole genome sequencing has meant that data sets of 'slowly' evolving microbes often carry sufficient information for calibrating the molecular clock (Biek et al., 2015). The causative agent of tuberculosis, the bacterium *Mycobacterium tuberculosis*, was commonly considered to evolve too slowly for calibrating the molecular clock using samples collected over a few years Duchene et al. (2016). However, data sets involving the full genome, of about 4.4 Mb, have demonstrated that a genome sampling window of a few decades might be sufficient for reliable clock calibration (Menardo et al., 2019).

1.2 The phylodynamic threshold

The emergence of SARS-CoV-2 saw the rapid generation of genome data from the early stages of the outbreak, with phylodynamic analyses conducted in near to real-time (Attwood et al., 2022). The first attempts to estimate the evolutionary rate and time of origin were highly uncertain due to two factors; a narrow sampling window and low genetic diversity. In Duchene et al. (2020a) Bayesian phylodynamic analyses of the available genomes were conducted as the outbreak unfolded, such that the number of genomes and the width of the

sampling window increased over time and ranged from 22 genomes sampled over 31 days to 122 genomes sampled over 63 days. The early estimates of the evolutionary rate and time of origin had high uncertainty, but they rapidly converged to values that were robust to the addition of more data. The term 'phylodynamic threshold' refers to the time where an organism has accrued a sufficient amount of genetic change *since its emergence* for tip-calibrations to be informative (Duchene et al., 2020a).

The terms *phylodynamic threshold* and *measurably evolving population* are different, albeit related, concepts. A population is measurably evolving if the samples available are sufficiently informative as to warrant tip calibration. In contrast, the phylodynamic threshold is the amount of time over which we would need to draw samples for the data set to behave as a measurably evolving population. For a recently evolving pathogen it would simply correspond to the point in time at which it has accrued sufficient genetic diversity since its emergence, under the condition that the data have been collected constantly over time. In contrast, an ancient organism may have attained its phylodynamic threshold, with substantial genetic diversity, but drawing a samples from a very short time window may fail to capture a representative amount of such genetic diversity (fig. 1).

1.3 Temporal signal

Tests of temporal signal are designed to assess our ability to extract information from data for estimating evolutionary rates and timescales (reviewed by Rieux and Balloux (2016)). These tests do not differentiate between recently emerging organisms (i.e. they have not attained their phylodynamic threshold; fig. 1a) and those with narrow sampling windows (i.e. the data cannot be treated as being drawn from a measurably evolving population; fig. 1d), both of which may lack temporal signal. Additionally, these tests assume that the phylogenetic model adequately captures the evolutionary process. However, pervasive evolutionary rate variation (an 'overdispersed molecular clock') may lead to rejection of temporal signal under a strict clock model. Thus, temporal signal is not solely a property of the data but also depends on model performance. Recent research suggests that the choice of tree prior and molecular clock model significantly impacts the sensitivity and specificity of temporal signal tests, with relaxed molecular clocks being particularly effective (Tay et al., 2024).

Various methods exist for assessing temporal signal, including root-to-tip regression (Buonagurio et al., 1986, Drummond et al., 2003a, Gojobori et al., 1990), date-randomization tests (Duchêne et al., 2015, Duchene et al., 2018, Ramsden et al., 2009, Trovão et al., 2015), and the Bayesian Evaluation of Temporal Signal

(BETS; Duchene et al. (2020c)). Root-to-tip regression fits a regression of root-to-tip distance against sampling time, with the R^2 value indicating clocklike behavior. Date-randomization tests compare evolutionary rate estimates using correct sampling times against permutations. BETS evaluates whether including sampling times improves model performance using Bayes factors. Importantly, a lack of temporal signal does not necessarily preclude estimating evolutionary rates and timescales, as additional sources of information (e.g., prior rate estimates or known internal node ages) can provide valid calibration, albeit potentially less precise than sequence sampling times for microbial pathogens.

A lack of temporal signal is typically considered to be associated with unreliable molecular clock estimates. In practice, this situation is thought to be alleviated by broadening the sampling window or by including more ancient samples. However, the presence and direction of a possible bias in estimates of evolutionary rates and timescales is poorly understood.

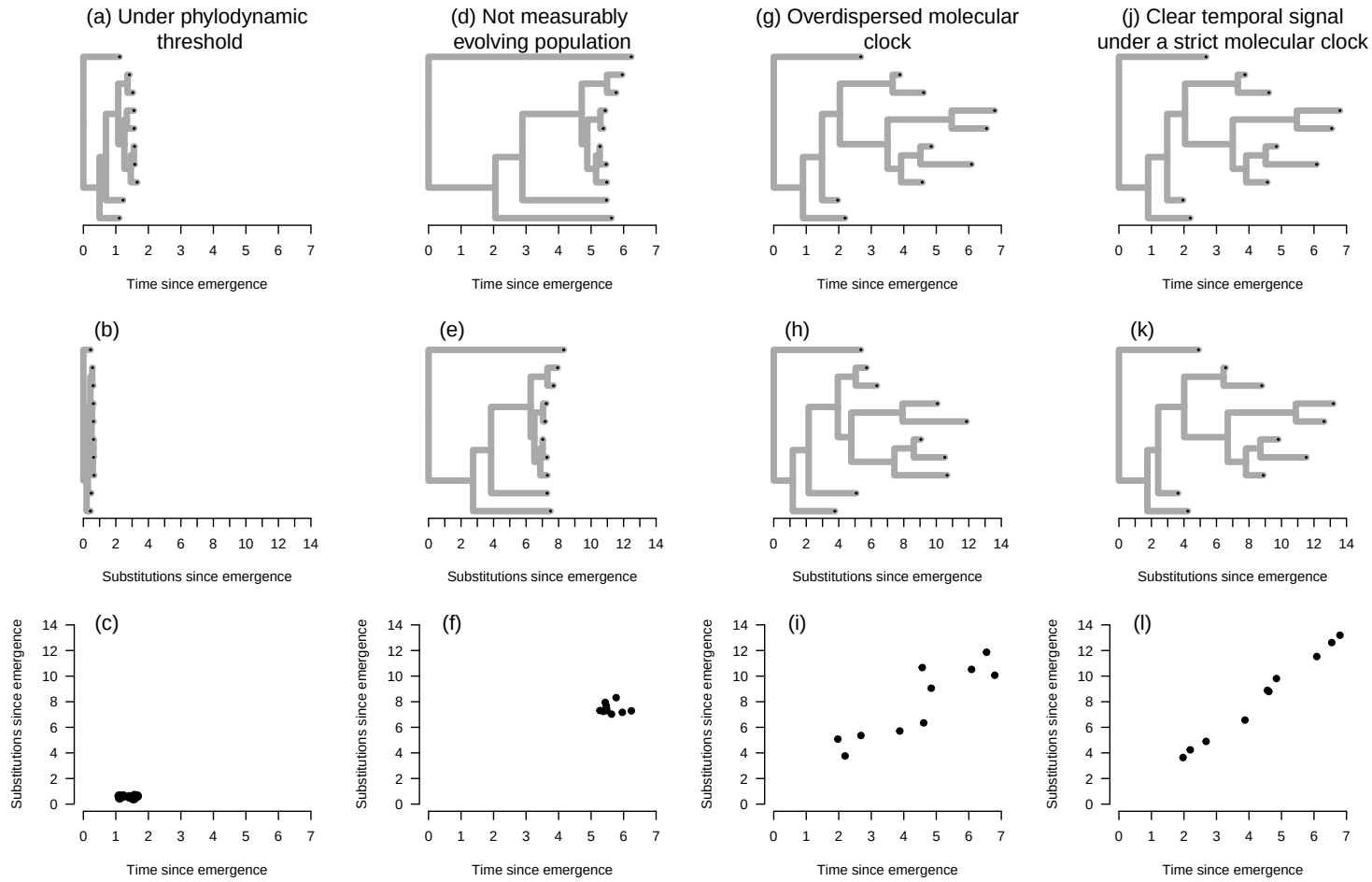

Figure 1: Examples of situations where temporal signal is typically not detected. An organism that has not attained its phylodynamic threshold has a recent time of emergence (with a phylogenetic time tree shown in (a)) because it has not had sufficient time to accrue an appreciable number of substitutions (phylogenetic tree with branch lengths in subs/site, i.e. a 'phylogram', shown in (b)), such that it is not possible to establish a statistical relationship between molecular evolution (substitutions) and time (shown in (c)). Sequence data from an organism that has evolved for a substantial amount of time may have been sampled over a very narrow window of time that is not sufficient to treat it as a measurably evolving population (time tree in (d) and phylogram in (e)), which results in no temporal signal (root-to-tip regression in (f)). A data set may involve a wide sampling window of time and from a population that has attained its phylodynamic threshold, but an overdispersed molecular clock (substantial rate variation among lineages; panels (g) - (i)) may result in a lack of temporal signal. In (j) through (l) we show the situation where an organism has attained its phylodynamic threshold and it has been sampled for sufficiently long, as to produce a clear relationship between molecular evolution and time, and unequivocal temporal signal.

2 Results

In this study we sought to pinpoint the impact of sampling strategies on these estimates. We focused our attention on two major problems for emerging microbes and studies involving ancient DNA. First, we varied the sampling window of a population that has attained its phylodynamic threshold. Second, we subsampled a population over time to vary the number of ancient samples, leading to a temporal sampling bias. Finally, we illustrate these results in an empirical data set of Hepatitis B virus (HBV) that includes a large number of ancient samples (Kocher et al., 2021). This virus has been the subject of intense research due to its close association with human populations and complex evolutionary dynamics (Kahila Bar-Gal et al., 2012, Paraskevis et al., 2013, Ross et al., 2018).

2.1 Sampling windows relative to the phylodynamic threshold

We simulated sequence data that resembled the evolution of HBV, a double-stranded DNA (dsDNA) virus that has evolved in humans for around 10 thousand years (Kocher et al., 2021). Our synthetic data had a genome length of 3,200 Kb and an evolutionary rate of 1.5^{-5} subs/site/year (Kocher et al., 2021, Mühlemann et al., 2018). Under these conditions we expect to observe one mutation every 20 years ($\frac{1}{3200\ sites\times\ 1.5\times10^{-5} subs/site/year} \approx 20\ years$). This number is important for the design of our simulation experiments: 20 years is the time we would expect for it to attain its phylodynamic threshold and it is would typically be the required sampling window to detect temporal signal.

2.2 Calibration windows for tips

We conceived a simulation process under which the evolutionary timescale had an expected time of 10 thousand years and with a sampling window of 0, 10, 20, 200 , or 2,000 years. A sampling window spanning 0 years results in ultrametric trees with the sampling times providing no calibration information. In contrast, a sampling window of 10 years is half of the phylodynamic threshold and is expected to have weak temporal signal (see fig. 1(d)-(f)). Sampling windows of 20 years (the phylodynamic threshold) or wider are expected to behave as measurably evolving populations and with increasingly strong temporal signal (see fig. 1(j)-(l)). All our simulations were analysed under Bayesian phylogenetic framework, as implemented in the BEASTv2.5 platform (Bouckaert et al., 2019).

All our simulations produced posterior estimated that included the correct value used to generate the data (i.e. high coverage). Increasingly wide sampling windows improved the precision of the estimates, while

still including the correct value. We do not observe a systematic bias associated with the sampling window width, contrary to the expectation that low temporal signal necessarily results in an underestimation of the evolutionary rate and an overestimation of the tree height (Duchêne et al., 2015).

This result is largely expected due to our configuration of the prior. Notably, the prior on the phylogenetic tree and the evolutionary rate are particularly influential for estimating evolutionary rates and timescales (for a detailed investigation see Tay et al. (2024)).Here tree prior is a constant-size coalescent for which the prior on the population size (known as Θ) is an exponential distribution with mean of 5,000, which matches the value used to generate the data. Similarly, the evolutionary rate had a prior in the form of a Γ distribution with shape=1.5 and rate=10^6, whose mean is shape$\times$rate= 1.5×10^{-6} and thus also matches the 'true' value. Although these priors are centred on the correct values, they are vague, and it is important to note than in all cases, the posterior distribution of the evolutionary rate and tree height was narrower than the prior, meaning that even in the absence of sampling times the sequence data provide some information about these two parameters.

Note to self: That precision generally improves, but there does not seem to be a systematic bias if the prior is reasonable. When the prior is unreasonable, then we do need to sample quite past the phylodynamic threshold

WRITING UP TO HERE. THE REST IS RUBBISH!

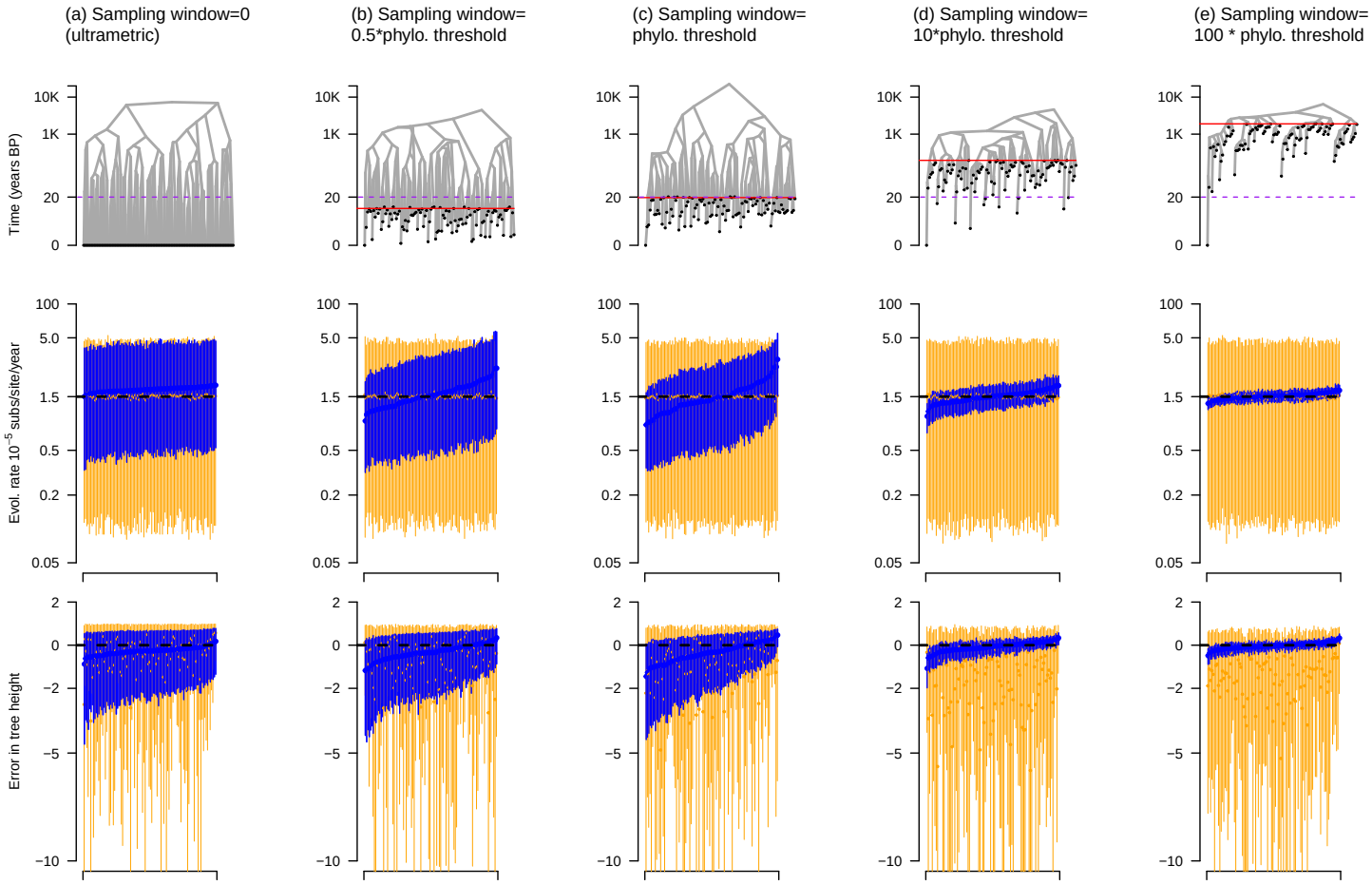

Figure 2: Simulations of varying sampling window widths. Each column corresponds to a simulation setting: (a) is for ultrametric trees where all samples are collected at the same point in time, (b) is for the situation where the sampling window is 10 years (half the phylodynamic threshold), (c) is where the sampling window is exactly the phylodynamic threshold of 20 years. Scenarios (d) and (e) denote sampling windows that are 10 and 100 times the phylodynamic threshold. The first row is an example of a simulated phylogenetic tree with branch lengths scaled in units of time. The black circles represent genomic samples. The purple dashed line is is the phylodynamic threshold and the solid red line is for the oldest sample, such that it represents the sampling window. Note that time here is shown in $\log_{10}$ scale. The second row is the estimated evolutionary rate over 100 simulations. The dashed black line is the value used to generate the data (i.e. the ground truth), the blue bars are the posterior, and those in orange are the prior. For the prior and the posterior we use solid circles to show the mean estimate. The third row is the estimate of the error in tree height (the age of the tree). The error in tree height is calculated as the $\frac{true-estimated}{true}$.

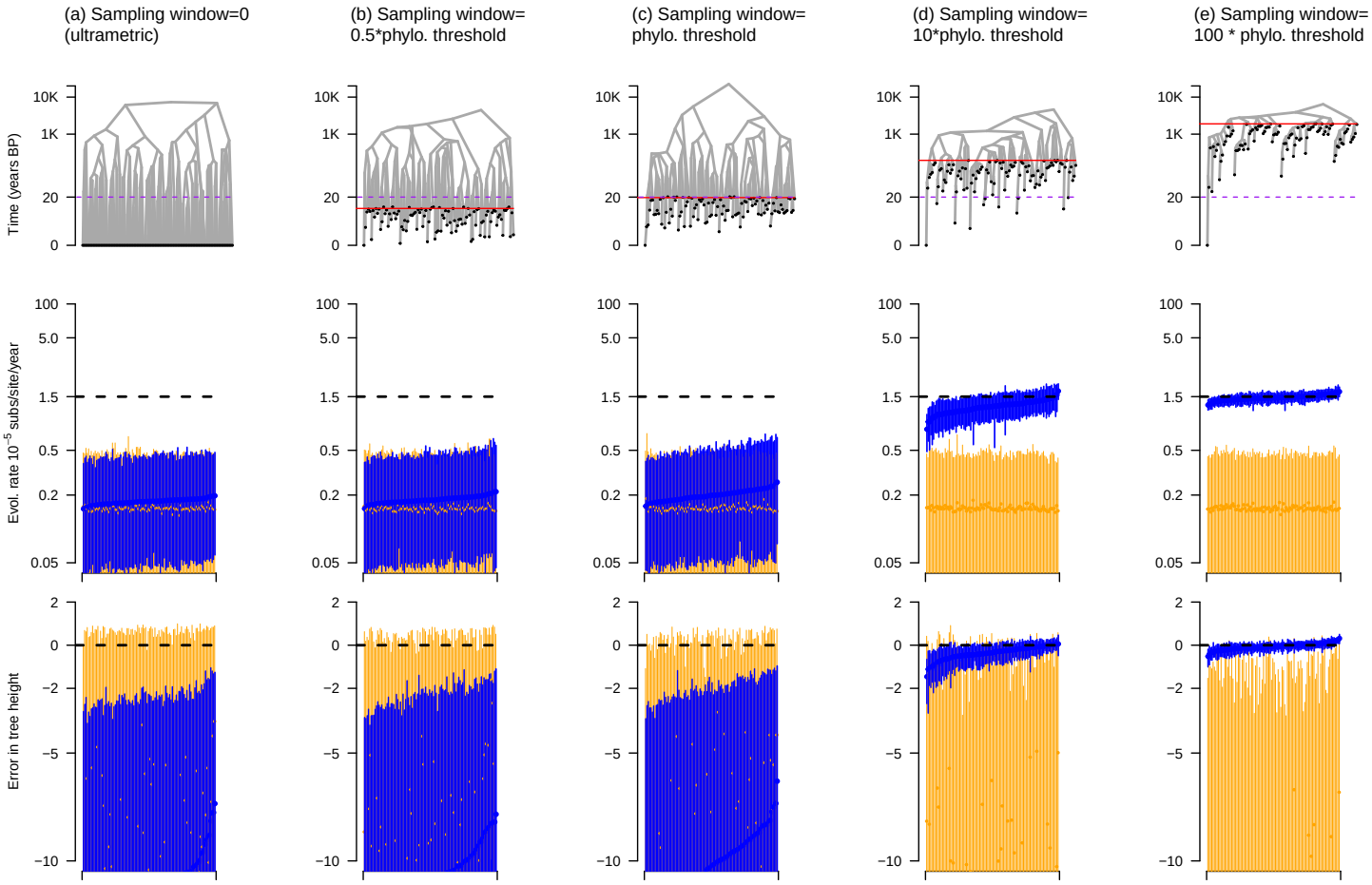

Figure 3: Simulations of varying sampling window widths. The colours, labels and legends match those from fig 2. However, in these analyses we deliberate use misleading priors on two key parameters, with a exponential distribution with mean 5,000 for the coalescent population size (true value=5,000), and a $\Gamma(shape = 1.5, rate = 10^6)$ with mean=1.5×10^{-6} (true value=1.5×10^{-5}).

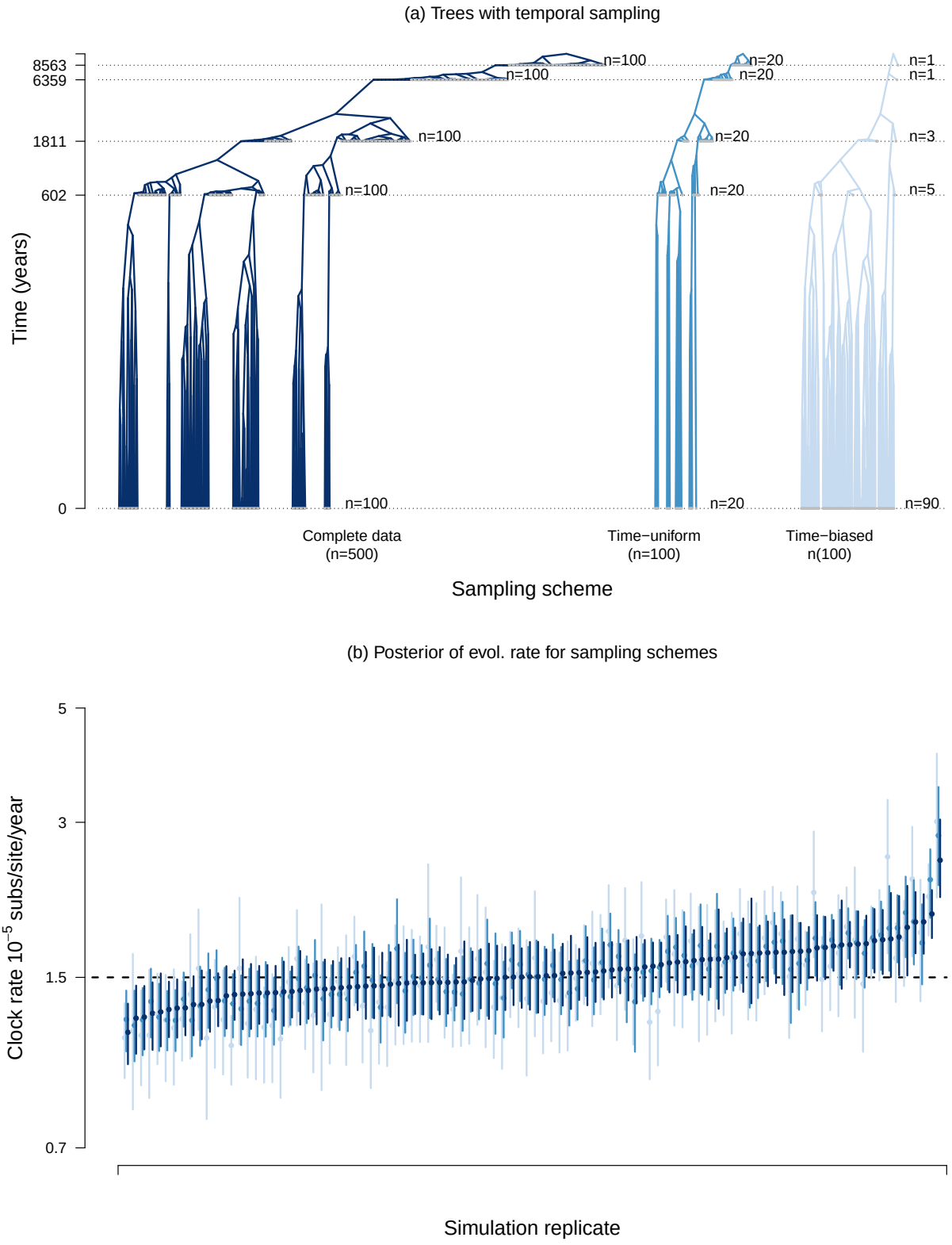

Figure 4: Trees plot thing pending

142 Blah

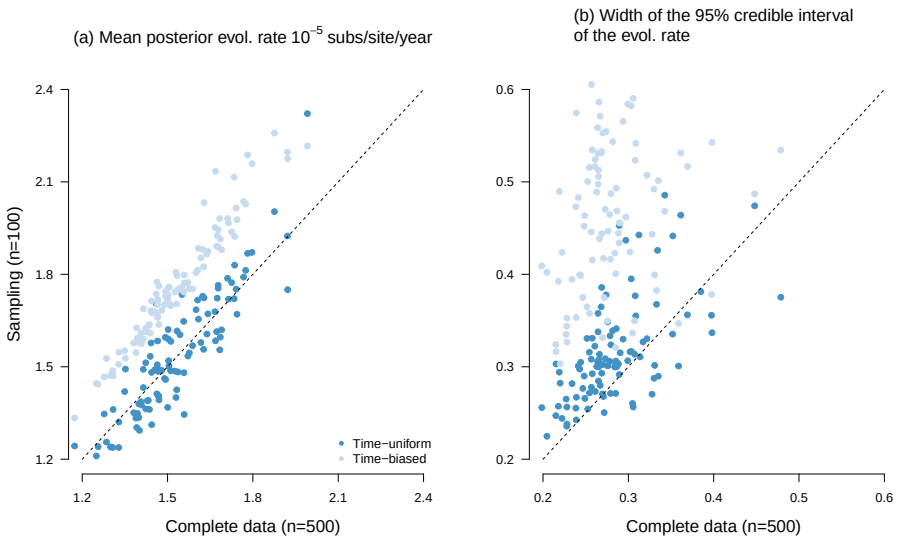

Figure 5: Trees plot thing pending

143 2.3 Bayesian molecular clock models

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

154 a

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). 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**	**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

3 Discussion

Our mean rate estimates over all lineages are somewhat lower than earlier estimates (Duchene et al., 2020a), which is consistent with the notion that the virus has had time to evolve and remove transient deleterious mutations since its emergence (Ghafari et al., 2021). However, 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., 2020a). Evolutionary rate variation is

4 Materials and methods

safdsf

4.1 Data set construction

verify convergence of independent chains and we ensured that the effective sample size of all parameters was at least 200.

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 acknowledge efforts by originating and submitting laboratories for the sequence data in GISAID EpiCoV on which our analyses are based. We are also grateful to Prof. Edward Holmes for useful suggestions and comments on ideas developed in this study.

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.

S. W. Attwood, S. C. Hill, D. M. Aanensen, T. R. Connor, and O. G. Pybus. Phylogenetic and phylodynamic approaches to understanding and combating the early sars-cov-2 pandemic. *Nature Reviews Genetics*, 23 (9):547–562, 2022.

R. Biek, O. G. Pybus, J. O. Lloyd-Smith, and X. Didelot. Measurably evolving pathogens in the genomic era. *Trends in ecology & evolution*, 30(6):306–313, 2015.

R. Bouckaert, T. G. Vaughan, J. Barido-Sottani, S. Duchêne, M. Fourment, A. Gavryushkina, J. Heled, G. Jones, D. Kühnert, N. De Maio, et al. Beast 2.5: An advanced software platform for bayesian evolutionary analysis. *PLoS computational biology*, 15(4):e1006650, 2019.

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.

D. A. Buonagurio, S. Nakada, J. D. Parvin, M. Krystal, P. Palese, and W. M. Fitch. Evolution of human influenza a viruses over 50 years: rapid, uniform rate of change in ns gene. *Science*, 232(4753):980–982, 1986.

M. Dos Reis and Z. Yang. The unbearable uncertainty of bayesian divergence time estimation. *Journal of Systematics and Evolution*, 51(1):30–43, 2013.

A. Drummond, O. G. Pybus, and A. Rambaut. Inference of viral evolutionary rates from molecular sequences. *Adv Parasitol*, 54:331–358, 2003a.

A. J. Drummond, O. G. Pybus, A. Rambaut, R. Forsberg, and A. G. Rodrigo. Measurably evolving populations. *Trends in ecology & evolution*, 18(9):481–488, 2003b.

S. Duchêne, R. Lanfear, and S. Y. Ho. The impact of calibration and clock-model choice on molecular estimates of divergence times. *Molecular phylogenetics and evolution*, 78:277–289, 2014.

S. Duchêne, D. Duchêne, E. C. Holmes, and S. Y. Ho. The performance of the date-randomization test in phylogenetic analyses of time-structured virus data. *Molecular Biology and Evolution*, 32(7):1895–1906, 2015.

S. Duchene, K. E. Holt, F.-X. Weill, S. Le Hello, J. Hawkey, D. J. Edwards, M. Fourment, and E. C. Holmes. Genome-scale rates of evolutionary change in bacteria. *Microbial genomics*, 2(11):e000094, 2016.

S. Duchene, D. A. Duchene, J. L. Geoghegan, Z. A. Dyson, J. Hawkey, and K. E. Holt. Inferring demographic parameters in bacterial genomic data using bayesian and hybrid phylogenetic methods. *BMC evolutionary biology*, 18:1–11, 2018.

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, 2020a.

S. Duchene, S. Y. Ho, A. G. Carmichael, E. C. Holmes, and H. Poinar. The recovery, interpretation and use of ancient pathogen genomes. *Current Biology*, 30(19):R1215–R1231, 2020b.

S. Duchene, P. Lemey, T. Stadler, S. Y. Ho, D. A. Duchene, V. Dhanasekaran, and G. Baele. Bayesian evaluation of temporal signal in measurably evolving populations. *Molecular Biology and Evolution*, 37 (11):3363–3379, 2020c.

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

M. Ghafari, L. du Plessis, J. Raghwani, S. Bhatt, B. Xu, O. Pybus, and A. Katzourakis. Purifying selection determines the short-term time dependency of evolutionary rates in sars-cov-2 and ph1n1 influenza. *medRxiv*, 2021.

T. Gojobori, E. N. Moriyama, and M. KIMuRA. Molecular clock of viral evolution, and the neutral theory. *Proceedings of the National Academy of Sciences*, 87(24):10015–10018, 1990.

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.

G. Kahila Bar-Gal, M. J. Kim, A. Klein, D. H. Shin, C. S. Oh, J. W. Kim, T.-H. Kim, S. B. Kim, P. R. Grant, O. Pappo, et al. Tracing hepatitis b virus to the 16th century in a korean mummy. *Hepatology*, 56 (5):1671–1680, 2012.

A. Kocher, L. Papac, R. Barquera, F. M. Key, M. A. Spyrou, R. Hübler, A. B. Rohrlach, F. Aron, R. Stahl, A. Wissgott, et al. Ten millennia of hepatitis b virus evolution. *Science*, 374(6564):182–188, 2021.

F. Menardo, S. Duchêne, D. Brites, and S. Gagneux. The molecular clock of mycobacterium tuberculosis. *PLoS pathogens*, 15(9):e1008067, 2019.

B. Mühlemann, T. C. Jones, P. d. B. Damgaard, M. E. Allentoft, I. Shevnina, A. Logvin, E. Usmanova, I. P. Panyushkina, B. Boldgiv, T. Bazartseren, et al. Ancient hepatitis b viruses from the bronze age to the medieval period. *Nature*, 557(7705):418–423, 2018.

D. Paraskevis, G. Magiorkinis, E. Magiorkinis, S. Y. Ho, R. Belshaw, J.-P. Allain, and A. Hatzakis. Dating the origin and dispersal of hepatitis b virus infection in humans and primates. *Hepatology*, 57(3):908–916, 2013.

C. Ramsden, E. C. Holmes, and M. A. Charleston. Hantavirus evolution in relation to its rodent and insectivore hosts: no evidence for codivergence. *Molecular biology and evolution*, 26(1):143–153, 2009.

A. Rieux and F. Balloux. Inferences from tip-calibrated phylogenies: a review and a practical guide. *Molecular ecology*, 25(9):1911–1924, 2016.

Z. P. Ross, J. Klunk, G. Fornaciari, V. Giuffra, S. Duchêne, A. T. Duggan, D. Poinar, M. W. Douglas, J.-S. Eden, E. C. Holmes, et al. The paradox of hbv evolution as revealed from a 16th century mummy. *PLoS pathogens*, 14(1):e1006750, 2018.

R. Sanjuán. From molecular genetics to phylodynamics: evolutionary relevance of mutation rates across viruses. *PLoS pathogens*, 8(5):e1002685, 2012.

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

M. A. Spyrou, K. I. Bos, A. Herbig, and J. Krause. Ancient pathogen genomics as an emerging tool for infectious disease research. *Nature Reviews Genetics*, 20(6):323–340, 2019.

J. H. Tay, A. Kocher, and S. Duchene. Assessing the effect of model specification and prior sensitivity on bayesian tests of temporal signal. *PLoS Computational Biology*, 20(11):e1012371, 2024.

O. Thalmann, B. Shapiro, P. Cui, V. J. Schuenemann, S. K. Sawyer, D. L. Greenfield, M. B. Germonpré, M. V. Sablin, F. López-Giráldez, X. Domingo-Roura, et al. Complete mitochondrial genomes of ancient canids suggest a european origin of domestic dogs. *Science*, 342(6160):871–874, 2013.

N. S. Trovão, G. Baele, B. Vrancken, F. Bielejec, M. A. Suchard, D. Fargette, and P. Lemey. Host ecology determines the dispersal patterns of a plant virus. *Virus evolution*, 1(1):vev016, 2015.

R. C. Warnock, Z. Yang, and P. C. Donoghue. Exploring uncertainty in the calibration of the molecular clock. *Biology letters*, 8(1):156–159, 2012.

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