Untangling temporal signal and the phylodynamic threshold of microbial sequence data sets

John H Tay[1], Sebastian Duchene[1,2*].

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

[2]Department of Computational Biology, Institut Pasteur, Paris, France.

*email: sduchene@pasteur.fr

Abstract ($\leq$300 words) Our understanding of the evolution of many microbes has been revolutionised by the molecular clock, a statistical tool to infer evolutionary rates and timescales. The fundamental assumption of molecular clock models is that the rate at which substitutions accumulate can be described by a statistical process. In all molecular clock models evolutionary rates and times are unidentifiable, and therefore 'calibration' information is essential to obtain estimates in calendar time.

For many microbes, the sequence sampling times themselves can be often used for calibration. Phylogenetic tests of temporal signal are often used to decide whether such calibrations are reliable. Critically, in addition to the calibration information, the full Bayesian phylogenetic model also includes the molecular clock model and a branching process (tree prior). As a result, there are multiple sources of information that are difficult to untangle.

We assessed temporal signal in three microbial data sets of human and animal diseases with a range of evolutionary characteristics and with ancient DNA sequences; *Powassan virus* (POWV), the cholera bacetrium (*Vibrio cholerae*), and the syphilis bacterium (*Treponema palladium.*). We found that the tree prior can have a substantial impact on whether temporal signal is detected. To investigate this problem we conducted extensive simulations and calculated the sensitivity and specificity of these tests under several tree priors.

We find that highly informative sequence data sets are generally robust to the tree prior. In contrast, in data sets with low information content, choosing a prior that is highly informative and inconsistent with the data can result in the false rejection of temporal signal.

We propose prior sensitivity analyses and prior predictive simulations to determine whether the prior is reasonable and to improve the detection of temporal signal and maximise the information that can be drawn from molecular sequence data sets.

Key words: Molecular clock, Ancient DNA, Bayesian phylogenetics, Microbial evolution.

1 Introduction

Molecular sequence data have been essential to unravel the evolutionary history of many organisms. The molecular clock is a statistical tool that posits that molecular evolution, in the form of substitutions, follows a statistical process. For example, under the simplest molecular clock model, known as the strict clock, substitutions accumulate linearly over time along a lineage, such that the evolutionary rate is constant over time (Zuckerkandl and Pauling, 1965). At the other end of the spectrum, relaxed molecular clocks allow every lineage in a phylogenetic tree to display a different evolutionary rate (Drummond et al. (2006) and reviewed in (Ho and Duchêne, 2014)).

All molecular clock models have a fundamental limitation, where evolutionary rates and times are unidentifiable. That is, there exist an infinite number of combinations of evolutionary rates and times that are compatible with an amount of evolutionary divergence (Dos Reis and Yang, 2013, Yang and Rannala, 2006). For this reason, external information, known as a molecular clock calibration is necessary to produce estimates in calendar units. As a case in point, consider two sequences whose genetic divergence from their most recent common ancestor is 10 subs/site. In the absence of calibrating information it is impossible to know *how rapidly* they evolve and *when* they diverged. The calibration can be a known evolutionary rate, such as 1 subs/site/year, or a divergence date, such as 1 year before present. The genetic distance can be divided by the evolutionary rate to infer the divergence time to infer a time to the most recent common ancestor of 10 years, or the genetic distance can be divided by the divergence date to infer the evolutionary rate, 10/subs/site/year in this case.

The finding that some organisms accumulate substitutions in a measurable timescale prompted the use of sequence sampling times for calibration (Drummond et al., 2003). The motivation behind this practice is that sequence data collected at different points in time would have accumulated a corresponding number of substitutions. With the example above, a sequence collected six months of the common ancestor would have accumulated 5 subs/site (10 subs/site/year × 0.5 years = 5 subs/site), whereas one collected after 1 year would have accrued 10 subs/site (10 subs/site/year× 1 year = 10 subs/site). As a result, sequence sampling times act as a time-calibration that is intuitively informative about the evolutionary rate.

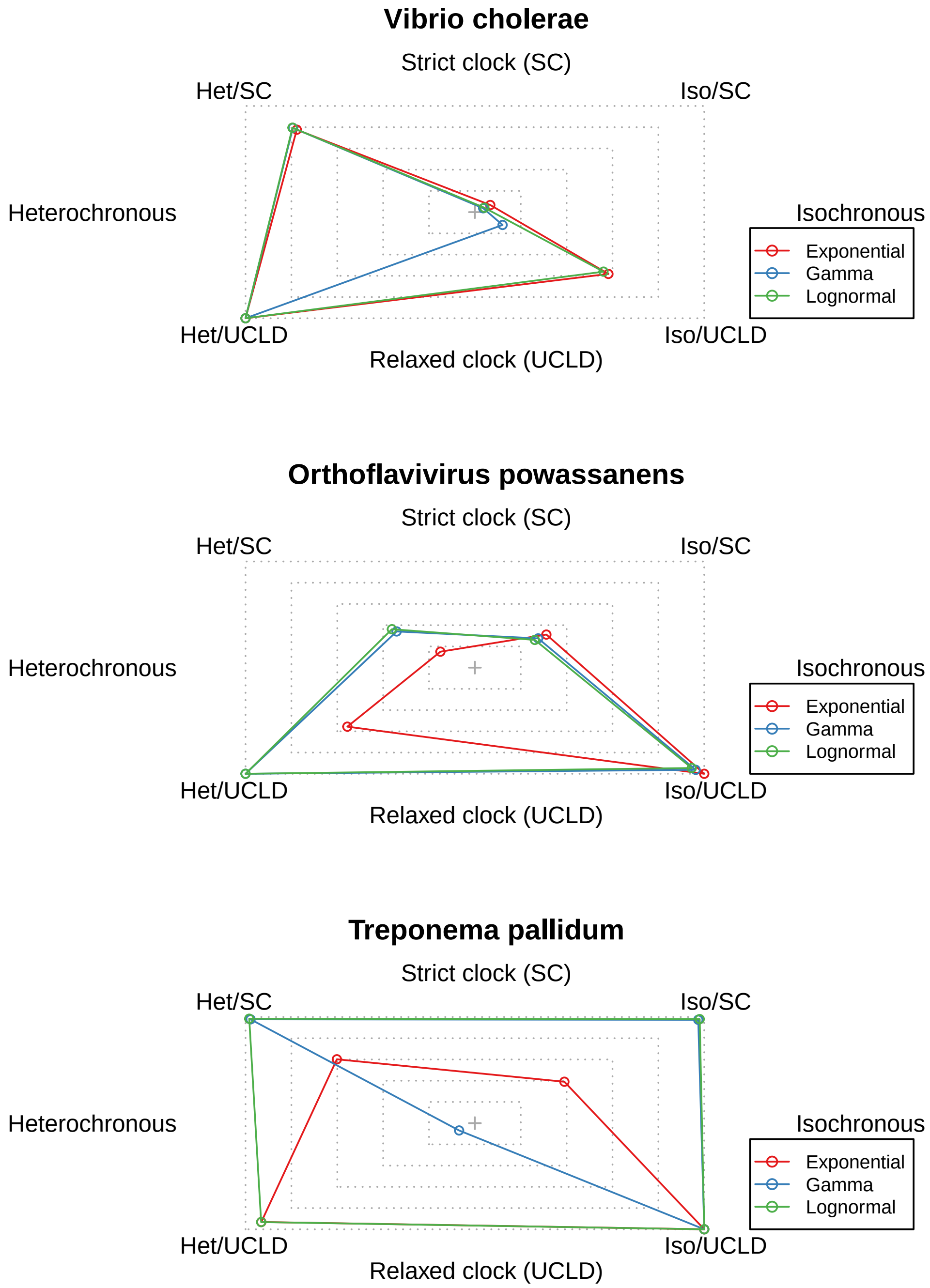
A fundamental question to warrant the use of sampling times for molecular calibration is the extent to which a data set behaves as a measurably evolving population.

Outline

Key points for Results section - 28 / 02 / 2024

- The impact of the tree prior and hyperpriors on model selection. Here show the polygons and tables. Pow is particularly interesting because the exponential prior on Φ does not support temporal signal. **Result: Priors obviously impact marginal likelihoods, but sometimes also model choice.**

  To do: Make figure with polygons for all data sets, add legends, etc... Also make latex table with marginal likelihoods, but only those that have sufficient ESS.



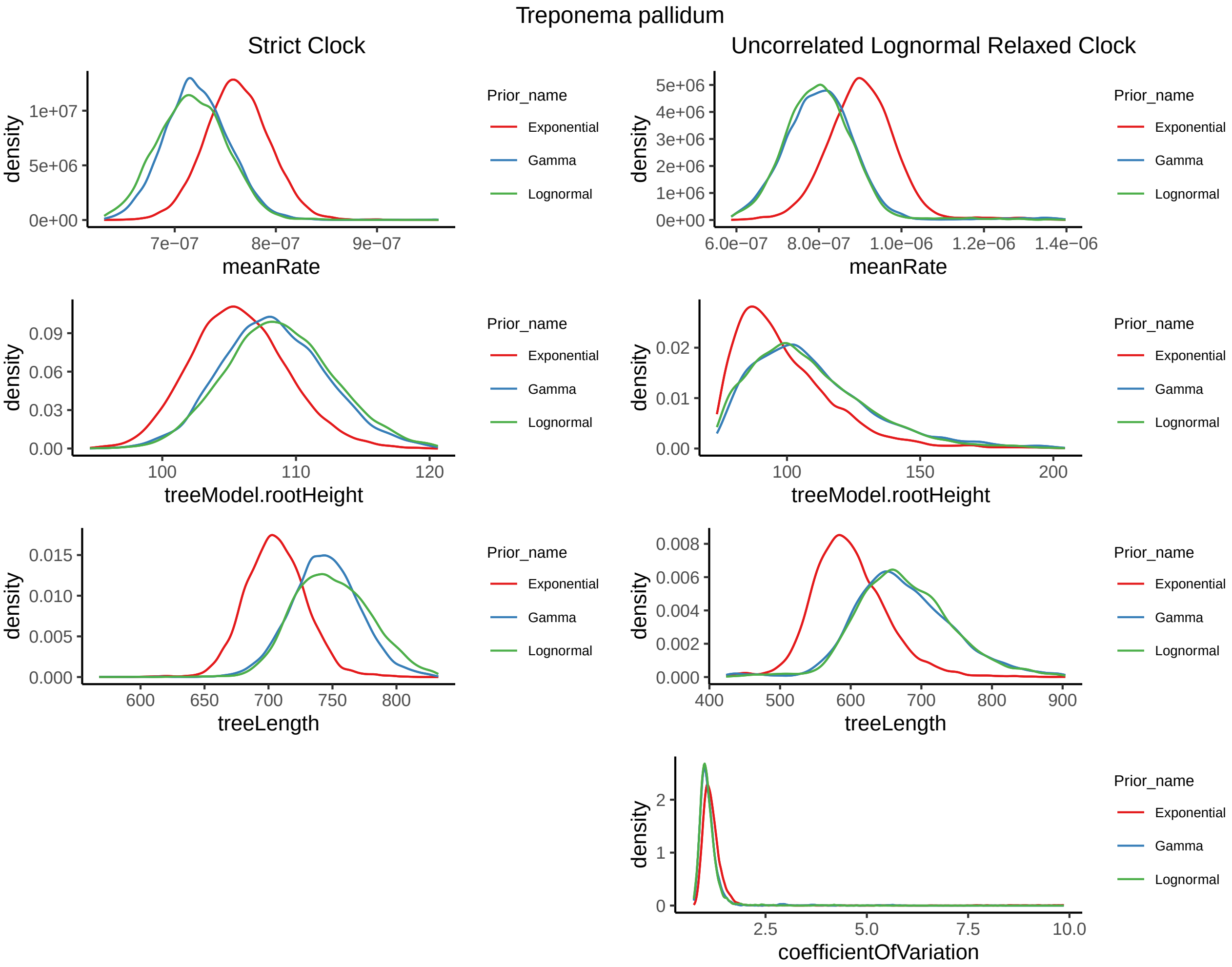
A polygon plot of the relative log marginal likelihoods of each microbe dataset under a different population size prior, analysed with four different configurations. Het (heterochronous) includes sampling, while iso (isochronous) does not include any sampling times. SC is strict clock and UCLD is the uncorrelated lognormal relaxed clock. Red represents an exponential hyperprior on the population size, blue is a gamma hyperprior, and green is a lognormal hyperprior.

Table 1: Log Bayes factors between isochronous and heterochronous models for each dataset, separated by prior on population size

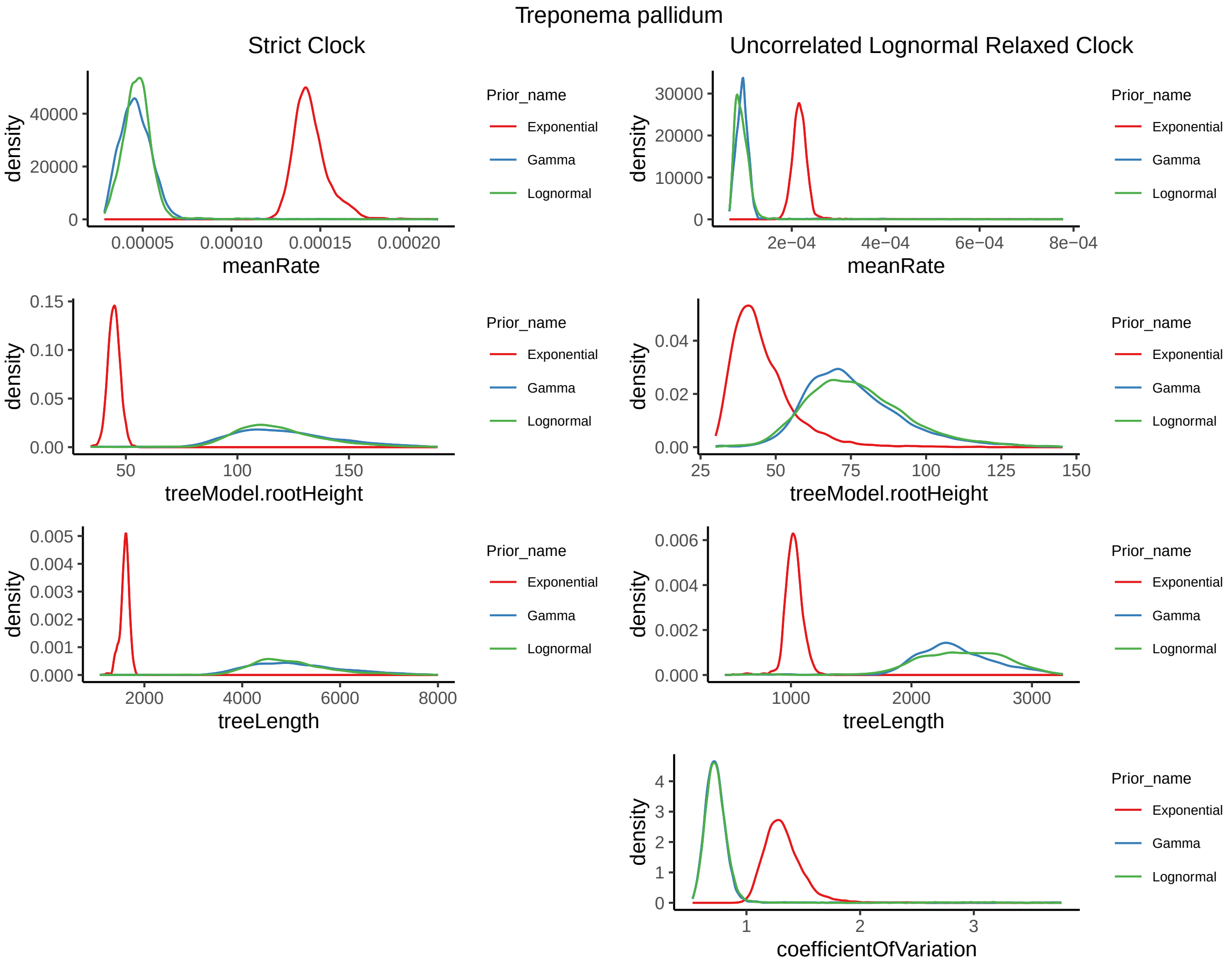
	Exponential	Gamma	Lognormal	
Vibrio cholerae; Strict Clock	355.18	379.63	382.10	
Vibrio cholerae; Relaxed Clock	208.97	439.63	219.60	
Orthoflavivirus powassanense; Strict Clock	-80.63	32.67	50.29	n
Orthoflavivirus powassanense; Relaxed Clock	-221.94	18.79	27.23	
Treponema pallidum; Strict Clock	105.80	2.17	1.85	
Treponema pallidum; Relaxed Clock	-34.37	-1474.14	-34.04	

- Show clock rate (ucld.mean and clock.rate), root height, tree length, and coefficient of rate variation (only for the relaxed clock) distributions under each clock model and prior. **Result: Estimates of evolutionary rates and times are generally robust to the tree prior even if the marginal likelihood is not**

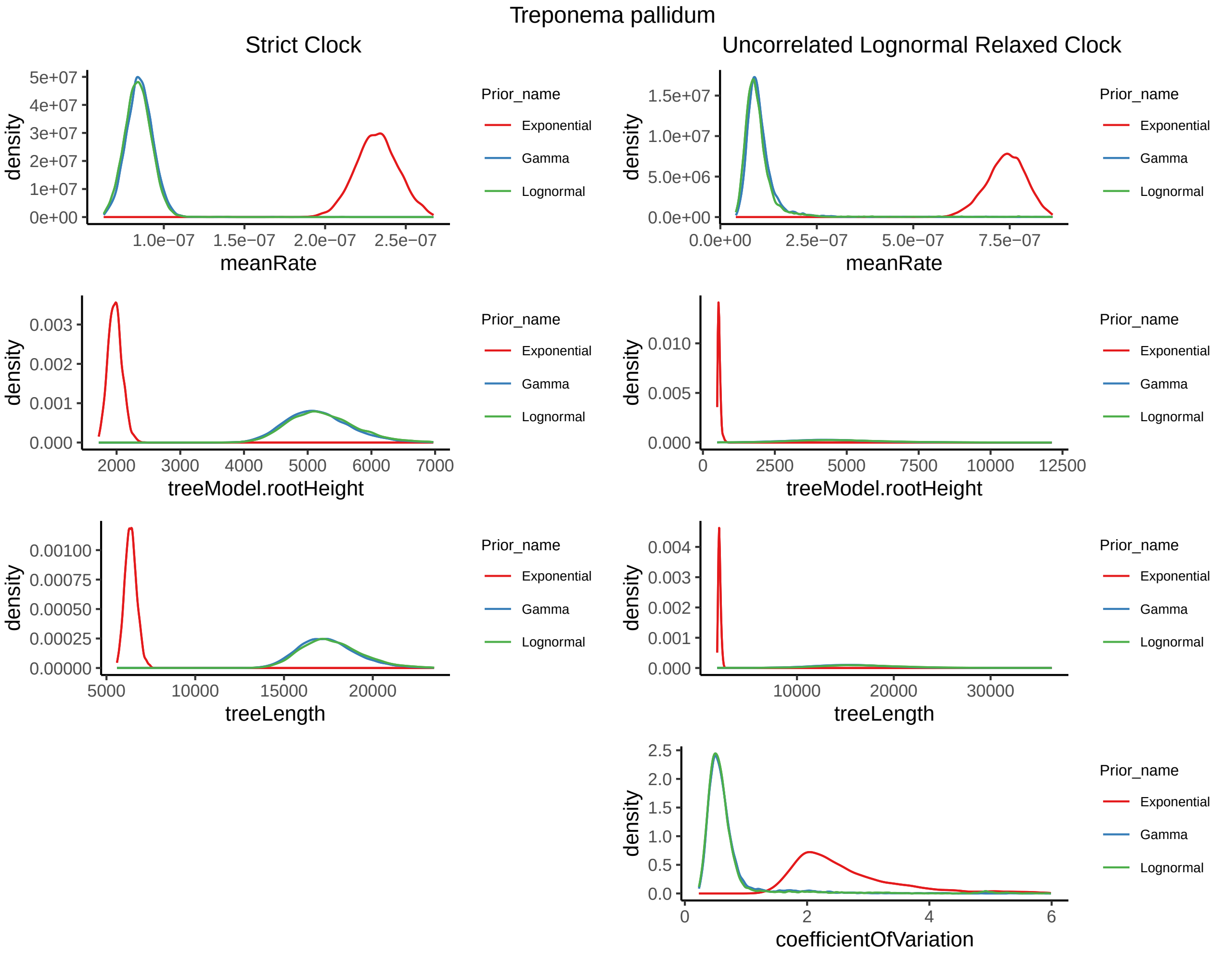
  To do: Plot densities of clock rates, root height, tree length, coefficient of rate variation. One panel per data set, show density for the three priors and the same colours as the fig above.



Density plots of the clock rate, root height, tree length, and coeffcient of rate variation of cholera.



Density plots of the clock rate, root height, tree length, and coeffcient of rate variation of powassan virus.



Density plots of the clock rate, root height, tree length, and coeffcient of rate variation of treponema palladium.

- Simulations **Pending for now - does the tree prior impact the accuracy to detect temporal signal?**

To do: This is where the simulations come in. Think of visuals here. Probably best displayed as tables? We can also just have plots of error bars per analysis to show coverage of key parameters...

Table 2: tree prior

	Exponential	Gamma	Lognormal
Vibrio cholerae; Strict Clock	1.0	1.0	1.0
Vibrio cholerae; Relaxed Clock			
Orthoflavivirus powassanense; Strict Clock			
Orthoflavivirus powassanense; Relaxed Clock			
Treponema pallidum; Strict Clock	0.0	0.0	0.0
Treponema pallidum; Relaxed Clock			

- Why does the prior on Φ impact temporal signal? Because it is linked to tree length and to the clock rate via the coalescent and the CTMC reference prior. Here show pairs plots for the root height, Φ, tree length, clock rate **Result: Some parameters in the Bayesian hierarchical model are naturally correlated, the marginal prior is not obvious, but can be easily inspected**

  To do: we should do pairwise plots to show the association of the parameters above (Φ, tree length, tree height, clock rate). This is just sampling from the prior, which I have done. Can we do this for the three priors on Φ, using the colours above?

- Inspecting the marginal prior is important and should be conducted prior to tests of temporal signal. For example, including monophyletic constraints, and other prior information can impact temporal signal.

- Prior sensitivity is also important to assess the robustness of the estimates.

General outline and key points

- Talk about tests of temporal signal

- The role of BETS

- pitfals of BETS: that we don't know the effect of the tree prior, all priors should be proper, impact of the clock model (because it can absorb rate variation among lineages).

- introduce data sets

- analyses of empirical data

  Use constant coalescent with the three population size priors and explain that they result in different priors on the clock rate and on the tree height. What are the implications here?

  Use the skyride and show the marginal prior on clock rate and tree height.

- Simulations. We should only do these for the best fitting model clock model per data set. But here, let's stick to the constant-size coalescent because it is more practical.

  How well do they classify data sets as having temporal signal? think about error rates.

  Does the classification accuracy depend on the average logBFs for those data sets that do have temporal signal? What I mean here is that the cholera ones would naturally have stronger temporal signal than the treponema ones, so we expect different error rates.

- Conclude: What is the best strategy here regarding tree prior? comment on the clock prior? where doe the regression fit in here? what about prior and posterior predictive checks?

- What does the root height, tree length, and clock rate prior look in each case (coalescent with all priors, skyride)?

  How do these priors impact the phylogenetic likelihood? The point here is that if the tree prior is very wrong (i.e. it is in conflict with the likelihood) then it can mislead tests of temporal signal.

with the relaxed clock it is useful to inspect whether the coefficient of rate variation is larger

- Ideally the test should assess only the sampling times

Prior sensitivity to run (21 Feb)

- Run all three priors on Φ and check

  tree height, clock rate, tree length, pop size - and correlations

- Run with a set of fixed Φ values and show how the mean and variance of clock rate, tree height and tree length changes

new results, as of 12 Feb

- The prior on the population size seems to matter when we don't have temporal signal. It tends to lead to false positives.

- A data set with temporal signal, will generally be robust to the prior.

- Redefine temporal signal. It is when the data are more informative than the prior in a Bayesian context.

- But this is not a problem in itself. One should choose a prior that is sensible. Some data sets, like cholera will be robust to the choice of prior. They are more informative and have stronger temporal signal.

Introduction

Pending

Table 3: Proportion of simulations with temporal signal according to log Bayes factors support (log BF)

Prior on Φ and clock model	$logBF \geq 0$	$logBF \geq 1$	$logBF \geq 3$	$logBF \geq 5$
$\Gamma(\kappa = 0.001, \theta = 1000)$; SC	0.25	0.23	0.21	0.21
$Lnorm(\mu = 1.0, \sigma = 5)$; SC	0.42	0.40	0.33	0.26
$Exp(\mu = 1.0)$; SC	NA	NA	NA	NA

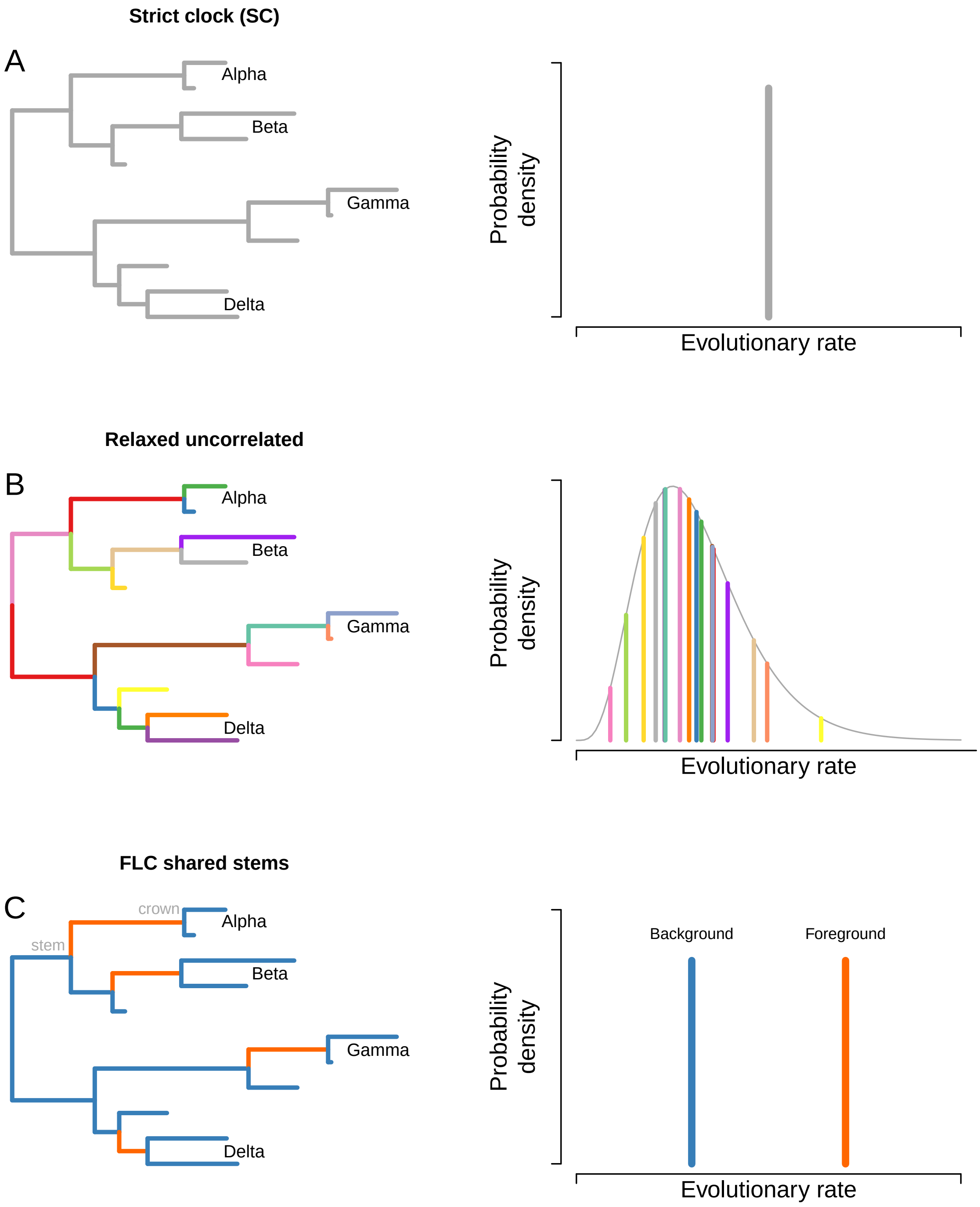


Figure 1: Examples of molecular clock models. A. is a strict clock model (SC), where all branches share a single evolutionary rate. B. is an uncorrelated relaxed clock model, in which the evolutionary rates across branches are independent draws from an underlying statistical distribution, such as a lognormal or a gamma (Γ) distribution (typically abbreviated as UCLD and UCGD, respectively). Note that branches are coloured according to those drawn from the distribution on the right. C. represents a fixed local clock model, where stem branches leading up to variants of concern (VOCs; those leading to clades labelled as Alpha, Beta, Gamma and Delta, following variant names) are designated as the 'foreground' and assigned a rate that differs from the 'background'. In C, we have labelled the stem and crown nodes of variant Alpha, where evolutionary rate changes would have occurred.

2 Supplementary Material

Supplementary data are available at Molecular Biology and Evolution online.

3 Acknowledgements

The Authors thank two anonymous reviewers and the Editor for helpful comments in earlier versions of this manuscript.

JHT and SD were supported by the Australian Research Council (FT220100629) and the Australian National Health and Medical Research Council (grant number 2017284). GB acknowledges support from the Internal Funds KU Leuven under grant agreement C14/18/094 and from the Research Foundation – Flanders ('Fonds voor Wetenschappelijk Onderzoek – Vlaanderen', G0E1420N and G098321N). The authors acknowledge efforts by originating and submitting laboratories for the sequence data in GISAID EpiCoV on which our empirical analyses are based. 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.

4 Data availability

The data underlying this article are available in GISAID at gisaid.org, and all accession numbers are provided in Supplementary Material online.

References

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

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

Z. Yang and B. Rannala. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Molecular biology and evolution*, 23(1):212–226, 2006.

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