Untangling temporal signal for calibrating the molecular clock of microbes

John H Tay[1], Arthur Kocher[2,3], Sebastian Duchene[1,2,*],

1 Peter Doherty Institute for Infection and Immunity, Department of Microbiology and Immunology, University of Melbourne, Melbourne, Australia
2 Pending pending.
3 Department of Computational Biology, Institut Pasteur, Paris, France

* sduchene@pasteur.fr

Abstract

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. Highly informative sequence data sets are generally robust to the tree prior. However, 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 demonstrate that prior sensitivity analyses and prior predictive simulations are an effective means to determine whether the prior is reasonable and to improve the detection of temporal signal and leverage the information that can be drawn from molecular sequence data sets.

Author summary

Pending.

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 [1]. At the other end of the spectrum, relaxed molecular clocks allow every lineage in a phylogenetic tree to display a different evolutionary rate ([2] and reviewed in [3]).

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 [4,5]. 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 [6,7]. 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 about using sampling times for molecular clock calibration is the extent to which the data can be assumed to have been sampled from a measurably evolving population. A measurably evolving population is defined as to the situation where the interval of time over which the samples were taken captures an appreciable amount of evolutionary change [8]. For rapidly evolving pathogens, such as RNA viruses, a measurably evolving population may be obtained by drawing samples over weeks or months. More slowly evolving microbes, such as the tuberculosis bacterium (*Mycobacterium tuberculosis*) may require sampling over many years.

There exist several statistical tests to determine whether a population has measurably evolving behaviour, also known as tests of temporal signal. The root-to-tip regression takes a phylogenetic tree for which the branch lengths measure evolutionary distance (i.e. a phylogram) and fits a linear regression of the distance from the root to the tips as a function of their sampling time [7]. The regression slope is a crude estimate of the evolutionary rate, the x-intercept is the time to the most recent common ancestor, and the R^2 is a measure of clocklike evolution. In general, the root-to-tip regression is a powerful tool for visual inspection of the data, for example to detect outliers or identify lineages with particularly low or high evolutionary rates [9–11]. However, it the data points are not statistically independent, such that it is not a formal statistical test of temporal signal and statistics, such as p-values are invalid [12].

Date randomisation tests consist of fitting a molecular clock to the data after permuting the sampling times, a procedure that is repeated multiple times to obtain a 'null' distribution of the evolutionary rate [13]. The data are considered to have

temporal signal if the evolutionary rate estimated with the correct sampling times falls outside such 'null' distribution [13, 14].

BETS (Bayesian Evaluation of Temporal Signal) is a fully Bayesian approach for assessing temporal signal [15], by treating it as a model selection problem. The premise of the test is that the data in question should have higher statistical fit when the sampling times are included than when they are not. In practice, the data are analysed with their correct sampling times and with all samples assigned the same date (typically at present), while keeping the rest of the phylogenetic model the same, including the molecular clock, tree prior and substitution model. The marginal likelihood is calculated in each case to compute Bayes factors, which quantify the amount of evidence for one model over another, here that with sampling times vs that without. A major advantage of BETS is that can consider the full model and it naturally accommodates important sources of uncertainty, including that due to radio carbon dating of ancient DNA studies [16].

Most parameters of the phylogenetic model have individual prior probability distributions that can be chosen by the user, for example, the evolutionary rate, or the transition-to-transversion ratio of the HKY substitution model. The phylogenetic tree topology and branch lengths are usually assigned a branching process, such as a coalescent or birth-death process. These tree priors implicitly impose a prior probability distribution on the ages of nodes and the tree length, and therefore may inadvertently impose highly informative calibration priors. Moreover, model selection, as is the case with BETS, can be sensitive to the choice of prior, even if the posterior is not [17, 18]. Here we investigate whether such prior information can overwhelm the signal from the data and potentially mislead Bayesian tests of temporal signal.

Results

Empirical data analyses

We collected genome sequences from three microbial data sets that have been shown to have temporal signal using Bayesian phylogenetic methods. The data sets were: *Vibrio cholerae* [19], the bacterium responsible for cholera; *Powassan virus (POWV)* [20], a tick-borne virus; and *Treponema pallidum* [21], the bacterium that causes syphilis. The *V. cholerae* and *T. pallidum* data sets involve ancient samples. We analysed the data sets using BETS under a coalescent tree prior with constant population size and two possible clock models; a strict and an uncorrelated relaxed clock with an underlying log-normal distribution. Our choice of the constant coalescent tree prior is based on statistical convenience, as it is fully parametric, and its simplicity, rather than describing a biological process. We set up our analyses in BEAST1.10 [22] and calculated log marginal likelihoods with and without sampling times for each combination of molecular clock model and tree prior.

To investigate the impact of the tree prior we considered different (hyper) prior distributions for the effective population size, Φ, the only parameter in the constant-size coalescent. This parameter is referred to as a *scale parameter* for time because large values imply more dispersion (the molecular clock rate is also a scale parameter), and it is typically assigned a $1/\Phi$ prior distribution, which is the Jeffrey's prior that is uninformative and invariant to reparameterisation [23]. This prior has attractive attributes because it maximises the signal from the data, but it is an improper distribution (it does not integrate to one over its domain, because $\int_0^\infty \frac{1}{x}$ is undefined), a problem for model comparison using Bayes factors, because marginal likelihood calculations require that all priors be proper distributions [24, 25]. Instead, we selected three prior distributions, an exponential, Γ, and log-normal, that have been used in

recent literature as shown in Table 1.

Our rationale for using different prior distributions on Φ is its impact on parameters that pertain to the molecular clock. Under the constant coalescent process, Φ is the single parameter in most implementations. Because it is associated with the genetic diversity and population size, it is expected that large values of Φ will result in a time to the most recent common ancestor that is older than if Φ is small. Thus a constant coalescent tree prior with a prior for Φ that favours large values will elicit old node times. The prior on Φ will also have an impact on the evolutionary rate for two key reasons. First, by impacting the overall age of the tree, it will also impact the length of time over which the sequence data evolved. Second, the default prior for the evolutionary rate in BEAST1.10 is a Gamma (Γ) distribution with shape (α) of 0.5 and beta (β, also known as the 'rate') equal to the tree length (sum of all branch lengths) [26, 27]. In this software, this prior is known as the CTMC-rate reference prior and its mean value is 0.5 / tree length, meaning that it is indirectly impacted by Φ.

Table 1. Prior distributions for the effective population size, Φ parameter

Probability distribution function	Parameters
Exponential	mean, $\mu = 1.0$
Γ (Gamma)	shape, $\kappa = 0.001$; scale, $\theta = 1000$
Log-normal	mean, $\mu = 1.0$; standard deviation $\sigma = 5.0$

The *V. cholerae* data set displayed overwhelming support for temporal signal (Table 2 and Fig 1), regardless of the molecular clock model and prior on Φ, with log Bayes factors of over 200. Note that a log Bayes factor of 3.2 corresponds to a model posterior probability ≈ 0.95 [28], and is considered as 'very strong support', following Kass and Raftery [29]. Although in this data set the prior on Φ did not impact model selection for detecting temporal signal, it did impact the magnitude of the Bayes factors.

For our other two data sets the impact of the prior on model selection was evident. For *Poawassan virus* the Γ and log-normal priors on Φ suggested strong temporal signal, whereas the exponential prior strongly favoured the exclusion of sampling times, according to the strict and relaxed molecular clock models. In our analyses of the *T. pallidum* data set we found that the strict molecular clock suggested temporal signal, according to all priors on Φ, although with very strong evidence only for the exponential prior and only 'positive evidence' for the Γ and log-normal priors. Under the relaxed molecular clock model all priors had very strong support against temporal signal.

Table 2. Log Bayes factors between isochronous and heterochronous models for each dataset, separated by prior on population size

Species; Clock Model	Exponential	Gamma	Log-normal
Vibrio cholerae; *Strict Clock*	355.18	379.63	382.10
Vibrio cholerae; *Relaxed Clock*	208.97	439.63	219.60
Powassan virus (POWV); *Strict Clock*	-80.63	32.67	50.29
Powassan virus (POWV); *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

Close here by stating that, as expected, the choice of the prior has an impact on model selection. This effect, however, is less notable in data sets with high information content. In particular, here the cholera data set has more informative sites and wider

Simulations based on empirical data   136

Table 3. Proportion of simulations with temporal signal under heterochronous simulated data

Species; Clock Model	Exponential	Gamma	Lognormal
Vibrio cholerae; *Strict Clock*	10	10	10
Vibrio cholerae; *Relaxed Clock*	10	10	10
Vibrio cholerae; *Best clock model*	10	10	10
Orthoflavivirus powassanense; *Strict Clock*			
Orthoflavivirus powassanense; *Relaxed Clock*			
Treponema pallidum; *Strict Clock*	0	0	0
Treponema pallidum; *Relaxed Clock*	0	10	8
Treponema pallidum; *Relaxed Clock*			

Table 4. Proportion of simulations without temporal signal under isochronous simulated data

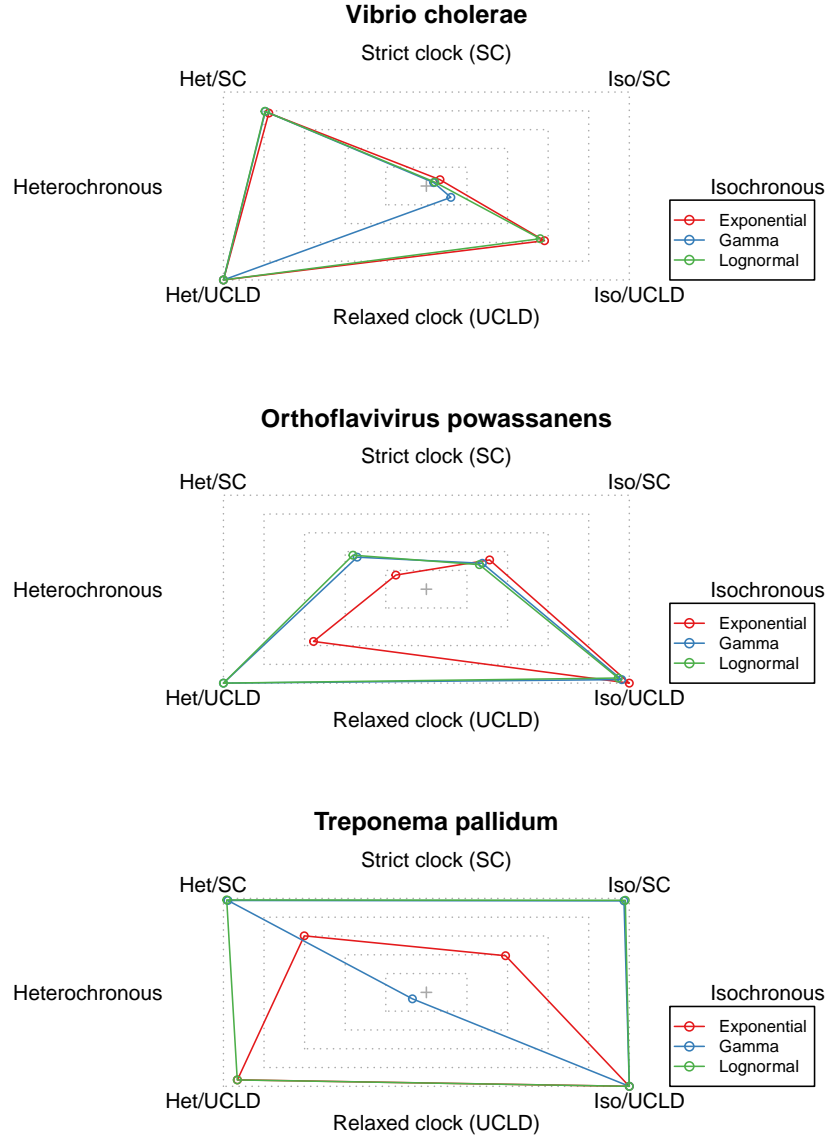| Species; Clock Model	Exponential	Gamma	Lognormal
Vibrio cholerae; *Strict Clock*	10	0	0
Vibrio cholerae; *Relaxed Clock*	0	0	0
Vibrio cholerae; *Best clock model*	0	0	0
Orthoflavivirus powassanense; *Strict Clock*			
Orthoflavivirus powassanense; *Relaxed Clock*			
Treponema pallidum; *Strict Clock*	10	10	10
Treponema pallidum; *Relaxed Clock*	10	0	0
Treponema pallidum; *Best clock model*			

Discussion   139

Fig 1. Relative log marginal likelihoods of empirical data sets. The polygons represent 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.

Materials and methods ₁₅₅

Empirical data ₁₅₆

John, please add information about the three sequence data sets here. I found some
notes we took on site patterns and others: Site patterns Empirical Cholera: 1392
Powassan: 3457 Treponema: 844 Simulations Cholera: 640-710 (ucld), 580-710 (sc)
Powassan: 14650 (het), 4800 (iso) Treponema: 110-150 (ucld), running (sc) (edited)
John, can you add details on how we ran these, e.g. MCMC length, marginal likelihood
settings, subst. model and parameter priors

Supporting information ₁₆₃

S1 Fig. **Bold the title sentence.** Add descriptive text after the title of the item
(optional).

S2 Fig. **Lorem ipsum.** Maecenas convallis mauris sit amet sem ultrices gravida.
Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula.
Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 File. **Lorem ipsum.** Maecenas convallis mauris sit amet sem ultrices gravida.
Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula.
Curabitur fringilla pulvinar lectus consectetur pellentesque.

Acknowledgments ₁₇₂

References

1. Zuckerkandl E, Pauling L. Evolutionary divergence and convergence in proteins.
   In: Evolving genes and proteins. Elsevier; 1965. p. 97–166.

2. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. Relaxed phylogenetics and
   dating with confidence. PLoS Biology. 2006;4(5):e88.

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

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

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

6. Rodrigo AG, Felsenstein J. Coalescent approaches to HIV population genetics.
   The evolution of HIV. 1999; p. 233–272.

7. Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, et al. Timing the
   ancestor of the HIV-1 pandemic strains. science. 2000;288(5472):1789–1796.

8. Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG. Measurably evolving populations. Trends in ecology & evolution. 2003;18(9):481–488.

9. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus evolution. 2016;2(1):vew007.

10. Featherstone LA, Rambaut A, Duchene S, Wirth W. Clockor2: Inferring global and local strict molecular clocks using root-to-tip regression. BioRxiv. 2023; p. 2023–07.

11. Volz E, Frost SD. Scalable relaxed clock phylogenetic dating. Virus evolution. 2017;3(2):vex025.

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

13. Ramsden C, Holmes EC, Charleston MA. Hantavirus evolution in relation to its rodent and insectivore hosts: no evidence for codivergence. Molecular biology and evolution. 2009;26(1):143–153.

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

15. Duchene S, Lemey P, Stadler T, Ho SY, Duchene DA, Dhanasekaran V, et al. Bayesian evaluation of temporal signal in measurably evolving populations. Molecular Biology and Evolution. 2020;37(11):3363–3379.

16. Molak M, Suchard MA, Ho SY, Beilman DW, Shapiro B. Empirical calibrated radiocarbon sampler: a tool for incorporating radiocarbon-date and calibration error into B ayesian phylogenetic analyses of ancient DNA. Molecular ecology resources. 2015;15(1):81–86.

17. Gelman A, Rubin DB. Avoiding model selection in Bayesian social research. Sociological methodology. 1995;25:165–173.

18. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis; 2014.

19. Devault AM, Golding GB, Waglechner N, Enk JM, Kuch M, Tien JH, et al. Second-pandemic strain of Vibrio cholerae from the Philadelphia cholera outbreak of 1849. New England Journal of Medicine. 2014;370(4):334–340.

20. Majander K, Pfrengle S, Kocher A, Neukamm J, Du Plessis L, Pla-Díaz M, et al. Ancient bacterial genomes reveal a high diversity of Treponema pallidum strains in early modern Europe. Current Biology. 2020;30(19):3788–3803.

21. Vogels CB, Brackney DE, Dupuis AP, Robich RM, Fauver JR, Brito AF, et al. Phylogeographic reconstruction of the emergence and spread of Powassan virus in the northeastern United States. Proceedings of the National Academy of Sciences. 2023;120(16):e2218012120.

22. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evolution. 2018;4(1):vey016.

23. Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics. 2002;161(3):1307–1320.

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

25. Baele G, Li WLS, Drummond AJ, Suchard MA, Lemey P. Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. Molecular Biology and Evolution. 2013;30(2):239–243. doi:10.1093/molbev/mss243.

26. Wang Y, Yang Z. Priors in Bayesian phylogenetics. Bayesian phylogenetics: methods, algorithms, and applications. 2014; p. 5–24.

27. Ferreira MA, Suchard MA. Bayesian analysis of elapsed times in continuous-time Markov chains. Canadian Journal of Statistics. 2008;36(3):355–368.

28. Tay JH, Baele G, Duchene S. Detecting episodic evolution through Bayesian inference of molecular clock models. Molecular Biology and Evolution. 2023;40(10):msad212.

29. Kass RE, Raftery AE. Bayes factors. Journal of the American Statistical Association. 1995;90(430):773–795.

30. Menardo F, Duchêne S, Brites D, Gagneux S. The molecular clock of Mycobacterium tuberculosis. PLoS pathogens. 2019;15(9):e1008067.

31. Biek R, Pybus OG, Lloyd-Smith JO, Didelot X. Measurably evolving pathogens in the genomic era. Trends in ecology & evolution. 2015;30(6):306–313.

32. Duchêne S, Ho SY, Carmichael AG, Holmes EC, Poinar H. The recovery, interpretation and use of ancient pathogen genomes. Current Biology. 2020;30(19):R1215–R1231.

33. Spyrou MA, Bos KI, Herbig A, Krause J. Ancient pathogen genomics as an emerging tool for infectious disease research. Nature Reviews Genetics. 2019;20(6):323–340.

34. Ho SY, Duchêne S. Dating the emergence of human pathogens. Science. 2020;368(6497):1310–1311.