The phylodynamic threshold of measurably evolving populations

Ariane Weber[1,*], Julia Kende[2,3], Sanni Översti[1,‡] and Sebastian Duchene[3,4,5,‡,*].

[1] Max Planck Institute of Geoanthropology, Jena, Germany.

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

[3] Université Paris Cité, Paris, France.

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

[5] 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

‡ Equal contribution to the supervision of this work.

Abstract The molecular clock is a fundamental tool for understanding the time and pace of evolution, requiring calibration information alongside molecular data. Sampling times are often used for calibration since some organisms accumulate enough mutations over the course of their sampling period. This practice ties two key concepts: measurably evolving populations and the phylodynamic threshold. Current dogma suggests that populations meeting these criteria are suitable for molecular clock calibration via sampling times. However, the definitions and implications of these concepts remain unclear. Using Hepatitis B virus-like simulations and analyses of empirical data, this study shows that determining whether a population is measurably evolving or has reached the phylodynamic threshold does not only depend on the data, but also on model assumptions and sampling strategies. In Bayesian applications, a lack of temporal signal due to a narrow sampling window results in a prior that is overly informative relative to the data, such that a prior that is potentially misleading typically requires a wider sampling window than one that is reasonable. In our analyses we demonstrate that assessing prior sensitivity is more important than the outcome of tests of temporal signal. Our results offer guidelines to improve molecular clock inferences and highlight limitations in molecular sequence sampling procedures.

Keywords: Measurably evolving population, 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 statistical processes relaxes this and other assumptions by allowing for rate variation among branches in phylogenetic trees (reviewed by Guindon (2020), 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 are unidentifiable (Dos Reis and Yang (2013), as reviewed by Bromham et al. (2018)), and therefore cannot be jointly estimated using genetic sequence data alone. To make inferences about them from genetic sequences, all molecular clock methods require prior assumption about evolutionary times or rates, known as a 'molecular clock calibration'. Three main calibrations exist: First, the age of the most recent common ancestor between two samples can be constrained to a given time point or interval ('internal node calibration'). Second, a known estimate of the evolutionary rate can be incorporated (e.g. as a prior in Bayesian frameworks). Third, in cases where molecular sequences are sampled at different points in time (heterochronous sampling), the tips of the phylogeny can be anchored to these time points ('tip calibration'; reviewed by Rieux and Balloux (2016)). The choice of calibration depends on the information available and its reliability (Duchêne et al., 2014, Warnock et al., 2012). For instance, it would be remiss to ignore evidence about when two lineages shared a common ancestor if the fossil record is compelling (Gavryushkina et al., 2017, Ronquist et al., 2016). Similarly, multiple sources of calibration information can be provided for the molecular clock.

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 (substitutions per genomic site per year) (Ghafari et al., 2022, Sanjuán, 2012). Assuming a genome size of 13,500 Kb, one would expect to observe one mutation every 4 to 5 days ($\frac{365\ days/year}{13{,}500\ sites\ \times\ 6{\times}10^{-3} subs/site/year} \approx 4.5\ days/subs$). If genome samples are collected over the course of a few weeks, the sampling times themselves can be used to calibrate the molecular clock and tip calibration is therefore warranted. Data sets for which tip calibration is feasible are considered to have been sampled from a 'measurably evolving population' (Drummond et al., 2003b) and to have 'temporal signal'.

Measurably evolving populations are typically characterised either by a sampling period that is long relative to the evolutionary rate, a sufficiently big data set (long molecular sequences or many samples), or both. Traditionally, such characteristics were mainly found in rapidly evolving organisms. Nowadays, 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. Namely, ancient DNA techniques have effectively expanded the genome sampling window for many organisms (Duchene et al., 2020b, Spyrou et al., 2019a), and 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) even when the sampling period covers only a few decades (Menardo et al., 2019).

1.2 The phylodynamic threshold

Genomic data sets collected during the early stages of an outbreak, for example, often pose two problems: low genetic diversity and a narrow sampling window. Both can lead to highly uncertain estimates of evolutionary rate and time of origin. The point at which an organism has accumulated sufficient genetic changes since its emergence to allow for informative tip calibration is referred to as the 'phylodynamic threshold' (Duchene et al., 2020a). At a minimum, tip calibration requires that one mutation has occurred over the sampling period for the method to be informative. For a given organism, the minimum sampling period can be calculated as the inverse of the product of genome size and the evolutionary rate (i.e. $\frac{1}{genome\ size\ (sites) \times evol.\ rate(subs/site/year)} =$ years to observe one mutation). We refer to this amount of time as the expected phylodynamic threshold.

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 allow for tip calibration. In contrast, the phylodynamic threshold is the amount of time over which we would need to draw samples after their emergence for them to behave as from a measurably evolving population. For a recently evolving pathogen the phylodynamic threshold would simply correspond to the time until it can be considered a measurably evolving population, under the condition that the data have been collected constantly over time. In contrast, an organism that emerged further in the past may have accumulated considerable genetic diversity over time, effectively reaching its phylodynamic threshold. However, if samples are drawn from a very short time window they may fail to capture a representative amount of such genetic diversity.

1.3 Tests of temporal signal

Our ability to extract information from a tip calibration framework can be assessed through tests of temporal signal. The importance of performing such tests arises from the observation that a lack of temporal signal is associated with unreliable evolutionary rate estimates (Duchêne et al., 2015, Rieux and Balloux, 2016), although the presence and direction of a potential bias remain poorly understood. However, it is important to note that a lack of temporal signal does not necessarily preclude estimating evolutionary rates and timescales because alternative sources of calibration, such as prior estimates of evolutionary rates or constraints on internal node ages, can still be used to inform analyses.

In principle, frameworks developed to test for temporal signal do not differentiate between recently emerging organisms (fig. 1a) and those with narrow sampling windows (fig. 1d), both of which may lack temporal signal. As most of these tests involve fitting a phylogenetic model to the data, they implicitly assume that the model adequately captures the evolutionary process and thus their performance also highly depends on model fit. Recent research, for example, suggests that the choice of tree prior and molecular clock model significantly impacts the sensitivity and specificity of temporal signal tests (Tay et al., 2024). Thus, temporal signal is not solely a property of the data but also depends on model performance.

Various methods exist for assessing temporal signal. The root-to-tip regression (Buonagurio et al., 1986, Drummond et al., 2003a, Gojobori et al., 1990) fits a regression to the distance from the root to the tips in a phylogenetic tree against sampling time. High R^2 values of the regression suggest that phylogenetic distance can be sensibly modelled as a linear function of time and can thus be used as an indication of informative tip-calibration. Date-randomisation tests (Duchêne et al., 2015, Duchene et al., 2018, Ramsden et al., 2009, Trovão et al., 2015) compare evolutionary rate estimates using correct sampling times against those from permutations. Bayesian Evaluation of Temporal Signal (BETS; Duchene et al. (2020c)) evaluates whether a model with sampling times performs better than a model that assumes isochronous sampling using Bayes factors. Each method comes with a set of limitations and strengths, such that tests of temporal signal should rather be used in combination than being mutually exclusive (Duchene et al., 2020c, Rieux and Balloux, 2016).

1.4 Concepts of measurably evolving populations, the phylodynamic threshold, and temporal signal in practice

In fig. 1, we present four simple example cases to illustrate the relationships among the concepts of measurably evolving populations, the phylodynamic threshold, and temporal signal. The first example depicts an organism that has emerged recently and therefore has not yet reached its phylodynamic threshold (with a phylogenetic time tree shown in panel (a)). Due to its recent origin, there has not been enough time for the accumulation of a sufficient number of substitutions (represented in the phylogram in panel (b)), such that it is not possible to establish a statistical relationship between molecular evolution (i.e., substitutions) and time (as shown in panel (c)). A real-world example of such a case comes from the early phase of the SARS-CoV-2 outbreak: initial efforts to estimate the evolutionary rate and time of origin had substantial uncertainty due to a narrow sampling window and low genetic diversity (Boni et al., 2020). In Duchene et al. (2020a), Bayesian phylodynamic analyses were conducted on genome data as the outbreak unfolded. The number of available 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. Although early estimates of the evolutionary rate and time of origin were highly uncertain, they quickly converged to stable values as more data became available (Ghafari et al., 2022).

The second example in fig. 1 illustrates a case in which an organism has evolved over a long period, but the available sequence data have been collected within a very narrow timeframe, insufficient to treat the dataset as a measurably evolving population (time tree in panel (d) and phylogram in panel (e)). This results in no temporal signal, as demonstrated by the lack of correlation in the root-to-tip regression in panel (f). 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). Later studies have shown, however, that for *M. tuberculosis* a genome sampling window of a few decades might be sufficient for reliable clock calibration (Kühnert et al., 2018, Menardo et al., 2019, Merker et al., 2022).

The third example in fig. 1 describes a data set that may involve a wide sampling window of time and for which samples have been drawn from a population that has attained its phylodynamic threshold, but with substantial rate variation among lineages – i.e. overdispersed molecular clock -, resulting in a lack of temporal signal (panels (g) – (i)). This pattern appears to be the case in *Yersinia pestis*, the bacterium that causes the plague, for which some localised outbreaks display obvious temporal signal, but its long-term evolution has pervasive evolutionary rate variation (Andrades Valtueña et al., 2022, Eaton et al., 2023).

In the final example in fig. 1, a hypothetical organism has attained its phylodynamic threshold, has been sampled for sufficiently long time, and evolutionary rate variation among lineages is low. These conditions together produce a clear relationship between molecular evolution and time, thus providing unequivocal temporal signal (panels (j) – (l)). The long term evolution of *Vibrio cholerae*, the causative agent of cholera, and H3N2 influenza virus are exemplar microbes whose molecular evolution has been fairly constant across long periods of time (Devault et al., 2014, Rambaut et al., 2016).

In summary, the concepts of measurably evolving population, phylodynamic threshold and temporal signal describe the information that can be drawn from a sampled population about its evolutionary timescale. Because populations that are not measurably evolving have been observed to yield biased estimates, they remain important to consider (Gharbi et al., 2024). In Bayesian inference, such biases can be the result of complex interactions between prior distributions or model settings that do not align with the true data generating model, as these drive the inference in the absence of informative data. Traditionally, potential biases due to prior interactions (Tay et al., 2024) or model misspecification (Möller et al., 2018) have been found through simulations studies, while data analyses often involve little validations of the results (Mendes et al., 2025). Here, we illustrate through a range of examples the degree to which differing levels of temporal signal in a data set can interact with prior settings and model assumptions, both on simulated and empirical data.

2 Results

We sought to pinpoint the impact of sampling strategies on molecular clock estimates. We focused our attention on two major problems for emerging microbes and studies involving ancient DNA. First, we conducted simulations varying the sampling window of a population that had attained its phylodynamic threshold. In the second simulation scenario, 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 at least for around ten thousand years (Kocher et al., 2021). Our synthetic data had a genome length of 3,200 nucleotides and an evolutionary rate of 1.5×10^{-5} subs/site/year (Kocher et al., 2021, Mühlemann et al., 2018) with a moderate amount of rate variation among lineages (see Materials and methods). Under these conditions we expect to observe one mutation every 20 years $(\frac{1}{3{,}200\ sites \times\ 1.5 \times 10^{-5} subs/site/year} \approx 20\ years/subs)$. This number is important for the design of our simulation experiments: 20 years is the expected phylodynamic threshold, as introduced above, and typically serves as a good reference point from when on to expect temporal signal. We analysed the data under a Bayesian phylogenetic framework and considered whether the posterior contained the true value used to generate the data, known as coverage, and the width of the posterior, known as precision (a precise estimate has a narrow posterior distribution).

We conceived a simulation process under which the evolutionary timescale had an expectation of ten 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 expected phylodynamic threshold and is likely to have weak temporal signal (see fig. 1(d)-(f)). Sampling windows of 20 years (the expected phylodynamic threshold) or wider are more likely to behave as measurably evolving populations with increasingly strong temporal signal (see fig. 1(j)-(l)). Our synthetic data sets were analysed under Bayesian phylogenetic framework, as implemented in the BEAST 2 platform (Bouckaert et al., 2019).

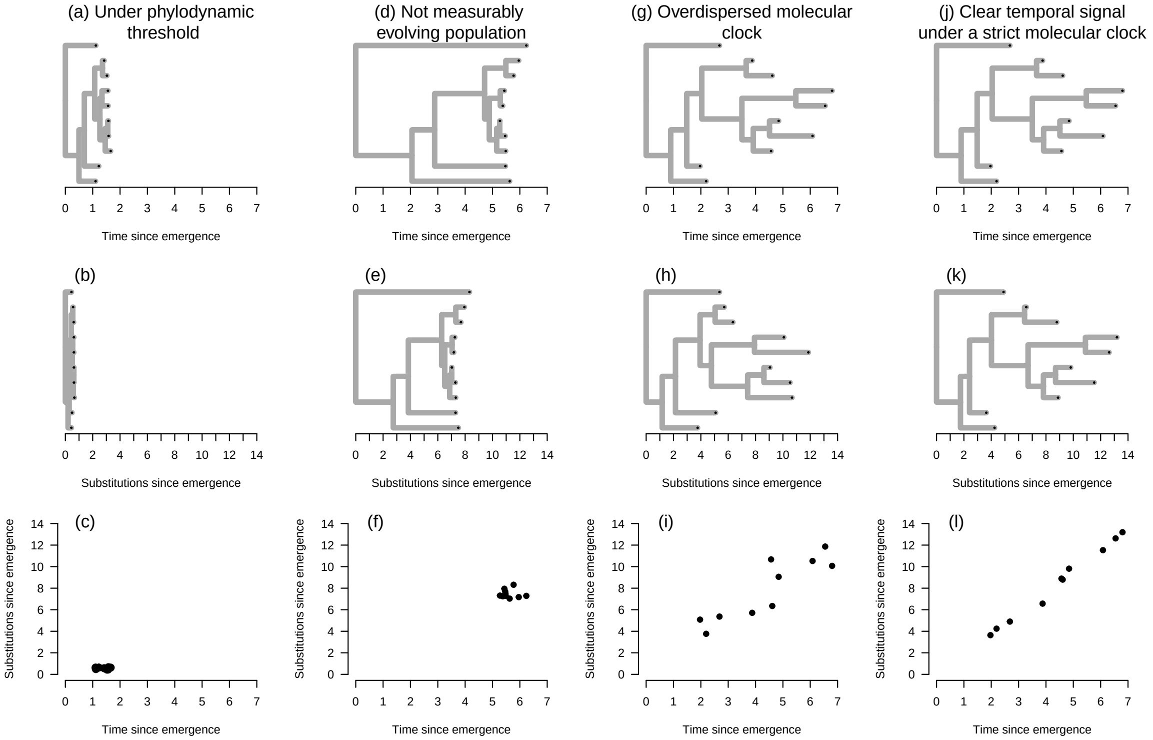


Figure 1: Examples of situations where temporal signal may or may not be detected. An organism that has not attained its phylodynamic threshold has a recent time of emergence (with a phylogenetic time tree shown in (a)) because it has not had sufficient time to accrue an appreciable number of substitutions (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, it has been sampled for sufficiently long, and where evolutionary rate variation among lineages is negligible, as to produce a clear relationship between molecular evolution and time, and thus unequivocal temporal signal.

All our simulations produced posterior estimates that included the correct value used to generate the data (i.e. 100% coverage; fig. 2). Increasingly wide sampling windows improved the precision of the estimates, while still including the correct value. This result is unsurprising, given our configuration of the prior. Here, the 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 $rate = 10^5$, whose mean is $shape/rate = 1.5 \times 10^{-5}$ 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 that in all cases, the posterior distributions of the evolutionary rate and tree height were narrower than their priors, meaning that even in the absence of sampling times the sequence data themselves provide some information about these two parameters.

We reanalysed these data with deliberately 'misleading' priors on the population size and the evolutionary rate. The prior on the population size was an exponential distribution with mean of 50,000, whereas the prior on the evolutionary rate was $\Gamma(shape = 1.5, rate = 10^6)$ (mean=1.5×10^{-6}). Under this configuration the mean prior mass corresponds to trees that are one order of magnitude older than the truth and evolutionary rates that are an order of magnitude slower. The objective of this experiment is to determine whether the sampling window is sufficiently informative to overcome such misleading prior information.

The posterior distribution was largely contained within the prior, resulting in low coverage for the evolutionary rate for sampling windows of 0, 10, and 20 years (0% coverage; fig. 3). A sampling window of 200 years was necessary to obtain 92% coverage, while a sampling window of 2,000 years had 100% coverage and even higher precision (fig. 3). These results demonstrate that a misleading prior that places a very low probability on the true value, requires a sampling window that is potentially much wider than the expected phylodynamic threshold.

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), we find that a lack of temporal signal due to narrow sampling windows may simply lend more influence to the prior. To confirm this point we conducted and additional set of simulations where the mean evolutionary rate prior was $\Gamma(shape = 1.5, rate = 10^4)$ (mean=1.5×10^{-4}, and 95% range from 1.1×10^{-5} to 4.7×10^{-4}), and thus should lead to an overestimation of this parameter. As expected, we also found that increasing the width of the sampling window resulted in a prior that was less influential on the posterior and with the latter converging to the value used to generate the data (see Supplementary material). Compared to the previous setting with incorrectly lower prior, a narrower sampling window already resulted in good coverage. This is unsurprising, as higher rate values are less likely when only few mutations are observed in a relatively long sampling window, while lower rate values cannot be excluded.

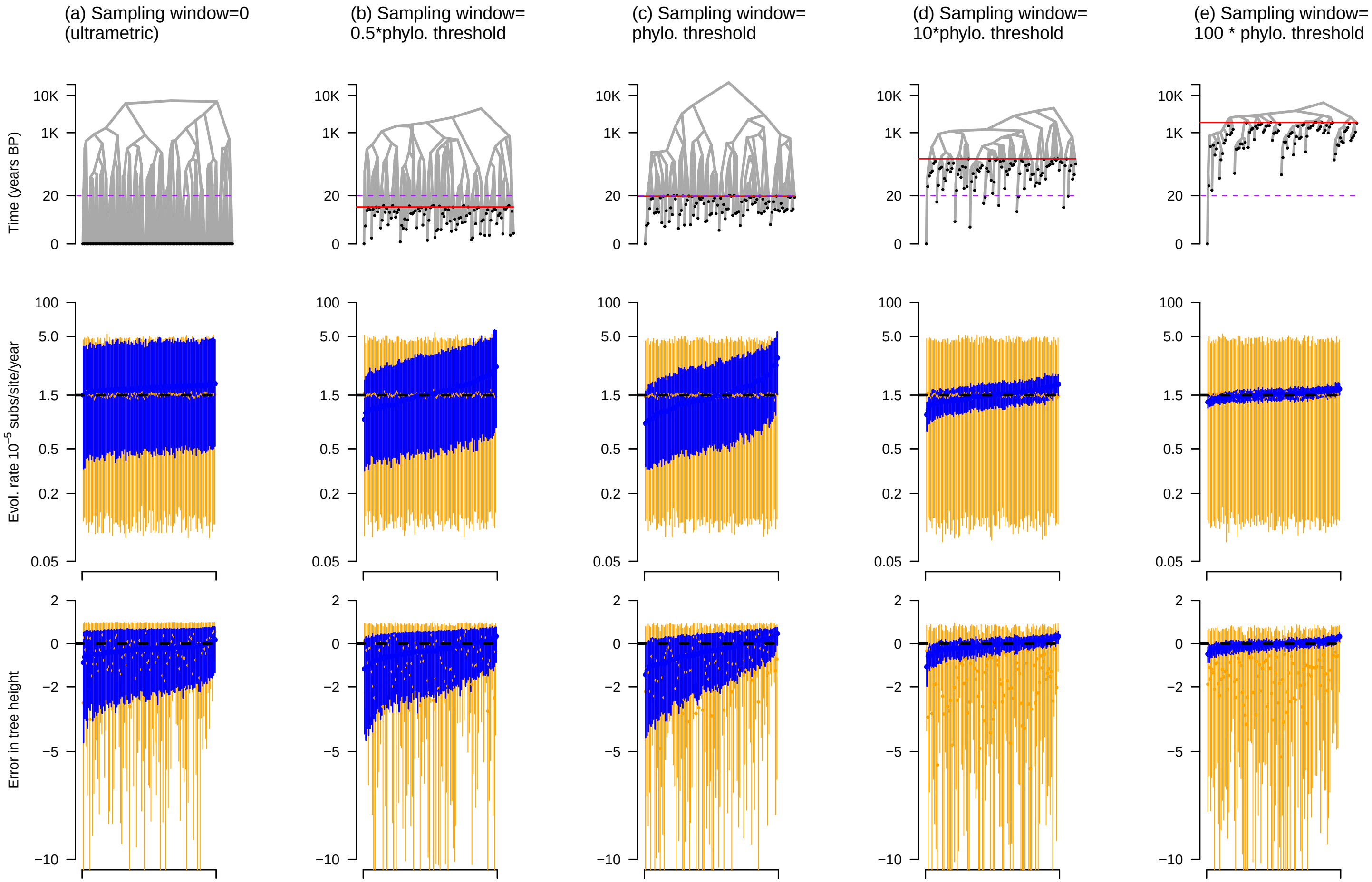


Figure 2: Simulations of varying sampling window widths. Each column corresponds to a simulation setting: (a) is for ultrametric trees where all samples are collected at the same point in time, (b) is for the situation where the sampling window is 10 years (half the expected phylodynamic threshold), (c) is where the sampling window is exactly the expected phylodynamic threshold of 20 years. Scenarios (d) and (e) denote sampling windows that are 10 and 100 times the expected 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 the expected 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 dark 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 and the width of the error bars denotes the 95% quantile range. 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 $\frac{true-estimated}{true}$.

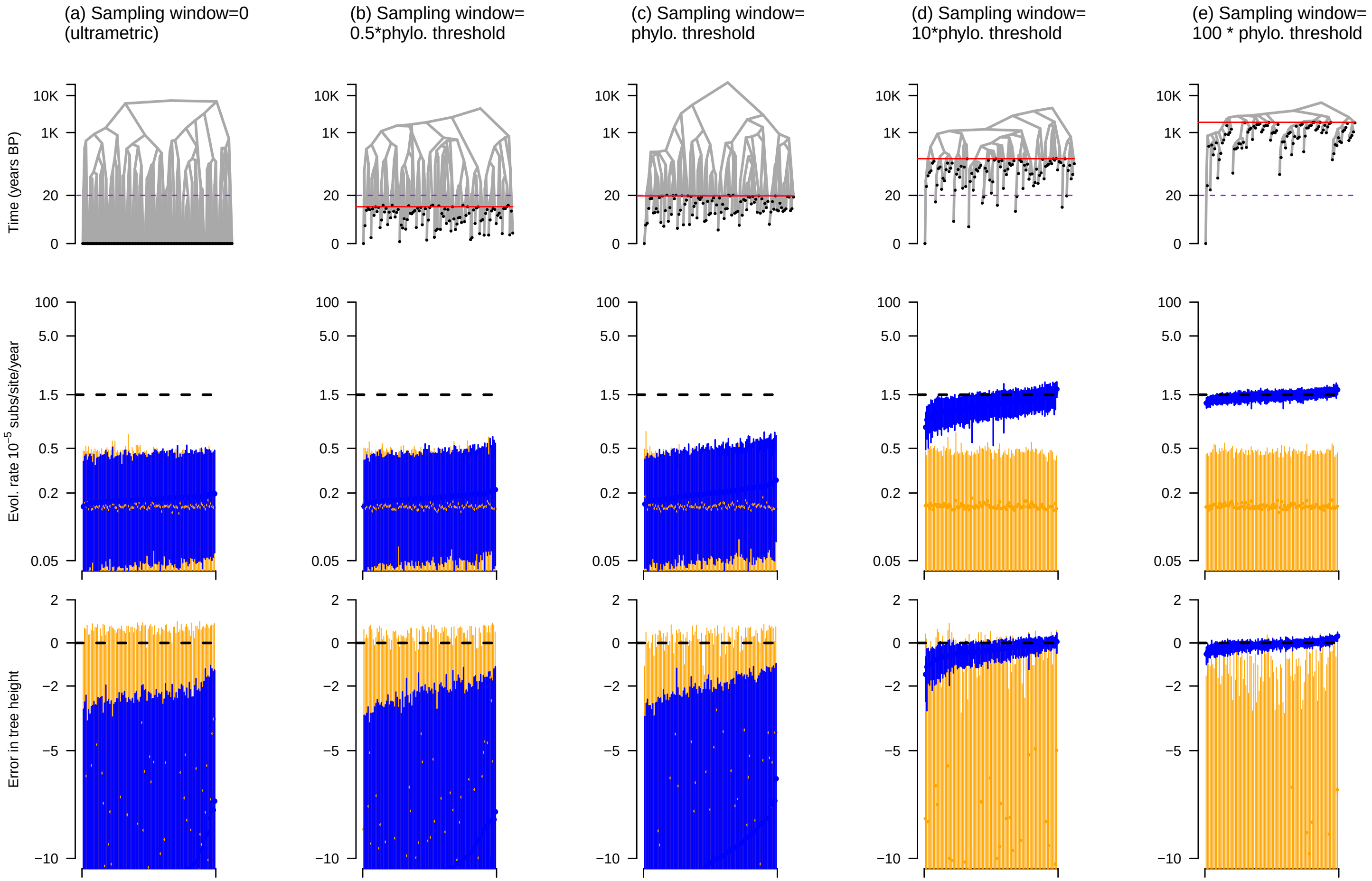


Figure 3: Simulations of varying sampling window widths. The colours, labels and legends match those from fig 2. However, in these analyses we deliberate use misleading priors on two key parameters, with a exponential distribution with mean 5,000 for the coalescent population size (true value=5,000),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

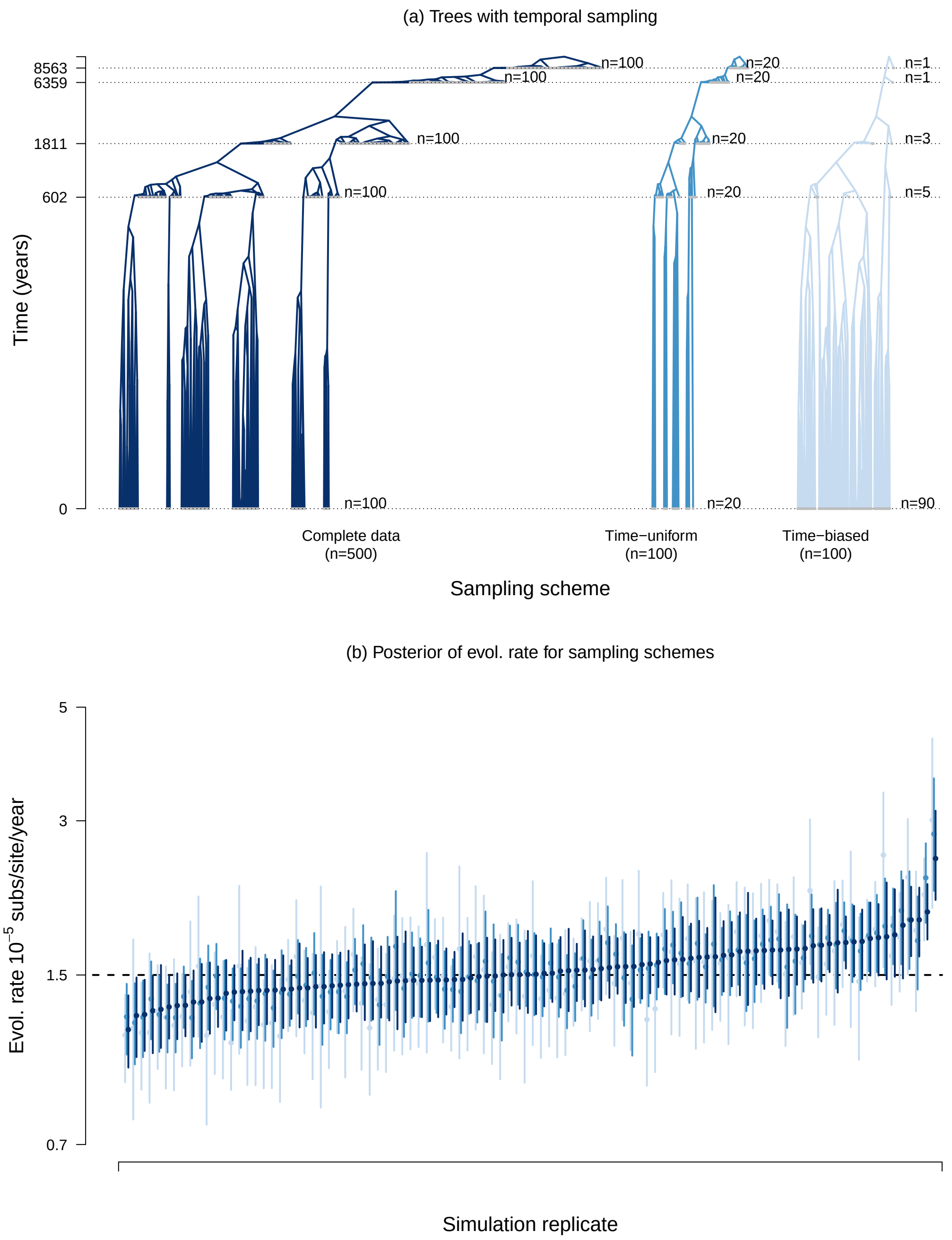
2.2 Temporal sampling bias

We investigated the impact of temporal sampling bias on the precision and accuracy on molecular clock estimates. For this purpose we simulated data with the same genomic characteristics as HBV and where genome sampling was conducted over five periods of time uniformly distributed between the present and the root of the trees (fig. 4(a)). The fully sampled trees contained 500 genome samples, with 100 for each of the five sampling times. Such stratified sampling is expected in ancient DNA studies, for example when a set of samples are drawn from archaeological sites (e.g. Spyrou et al. (2019b)). We sampled the complete data sets by randomly selecting 20 samples from each strata, which we refer to as 'time-uniform' sampling, and by sampling with a probability that is inversely proportional to the age of the strata, referred to as 'time-biased'. The time-uniform and time-biased sampling strategies both contain 100 samples ($1/5^{th}$ of the complete data), but the time-biased only includes a small number of ancient samples.

The coverage of the evolutionary rate estimate was comparable across simulation treatments, at 88% for the complete data set, 83% for the time-uniform, and 89% for the time-biased (fig. 4(b)). The somewhat higher coverage for the estimates from the time-biased analyses is likely because this sampling treatment has the lowest precision in the posterior, rather than an improvement in both accuracy and precision.

We also calculated a measure of bias for both sampling strategies by counting the number of simulations for which the posterior mean with either sampling strategy was higher than that with the complete data. We found that 50% of the estimates under time-uniform sampling had higher means than the complete data, while the same was true for 45% of those with time-biased sampling (fig. 5(a)). Although these numbers do not indicate substantial bias, such as a systematic over- or underestimation, we do note that time-biased sampling tends to produce lower mean evolutionary rate estimates than those obtained from the complete data or time-uniform sampling.

The most striking result of the temporal sampling strategies was in the precision of the posterior. Both sampling treatments resulted in posterior distributions that were wider than with the complete data, which is to be expected because they are effectively smaller data sets with less information. However, the time-biased sampling data sets almost invariably have posterior distributions that were less precise than those from the time-uniform sampling (Fig. 5(b)), implying that the distribution of samples, and not just the number, is important for estimation precision.


(a) Trees with temporal sampling

(b) Posterior of evol. rate for sampling schemes

Figure 4: Analyses under sampling treatments over time. In (a) we show an example of the trees for a simulation replicate, with branch lengths and time in $\log_{10}$ scale. The complete data set consists of 500 genome samples, collected in five points in time, with an equal number of samples per time point (n=100). The first sampling strategy is unbiased, where 20 samples are drawn from each time point, and is known here as 'time-uniform'. The 'time-biased' regime is where sampling intensity decreases over time. Note that the total number of samples in the time-uniform and time-biased treatments is identical. In (b) we show the posterior estimates of the evolutionary rate for each treatment. Each simulation replicate is represented by three error bars: dark blue for the complete data, and lighter shades of blue for the estimates from the time-uniform and time-biased sampling treatments. The width of the error bars denotes the 95% quantile range and the dots are the mean value. The dashed line shows the true value used to generate the data.

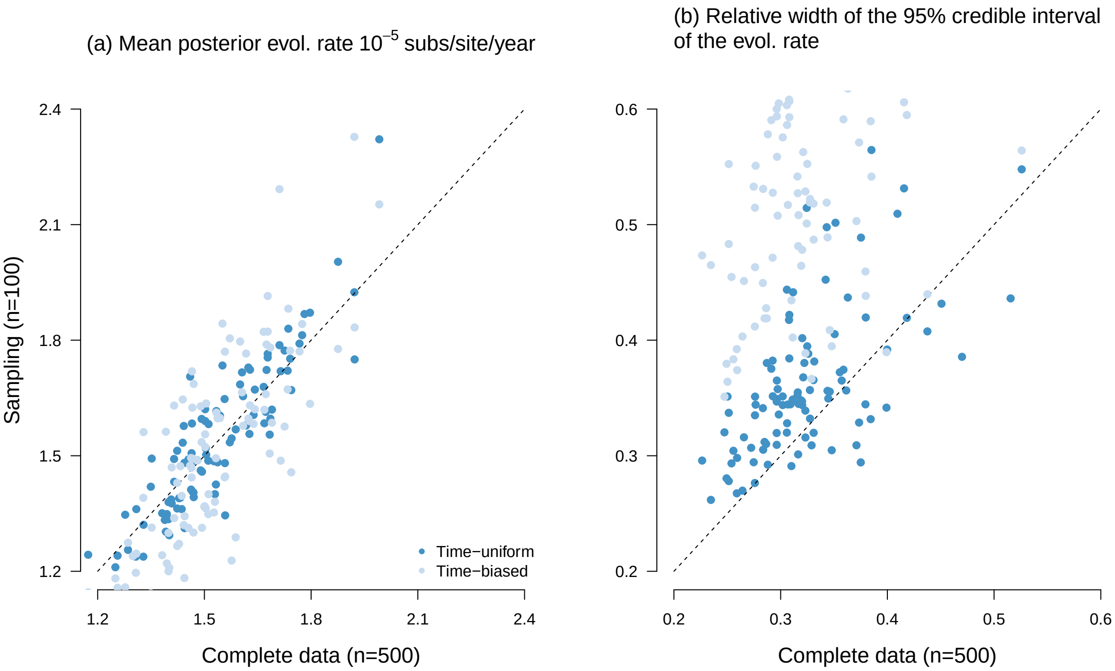


Figure 5: Comparison of posterior evolutionary rate estimates between complete data (x-axis) and two sampling treatments (y-axis): time-uniform (dark blue) and time-biased (light blue). Each dot is a simulation replicate. In (a) we show the mean posterior evolutionary rate estimate. Points that fall along the y=x line (dashed line) represent identical mean posterior for the sampling treatment and the complete data, while those above or below represent higher or lower estimates, respectively, relative to the complete data. In (b) we show the width of the credible interval (a measure of precision or uncertainty), calculated as the upper minus the lower 95% credible interval range divided by the mean value. Values that fall along the y=x line denote those for which the complete data and either sampling strategies are equally precise, while those above and below the y=x line are more or less precise, respectively.

2.3 Empirical analyses of Hepatitis B virus (HBV) ancient and modern genomes

To explore the impact of the width of the sampling window and the temporal sampling bias on the estimates of evolutionary rates and times, we performed analyses of a HBV data set that includes modern and ancient genomes, from Kocher et al. (2021). The complete data set consisted of 232 genomes of length 3,344 nucleotides and with a sampling window of 10,535 years. HBV is an ancient pathogen that has likely codiverged with human populations for thousands of years (Locarnini et al., 2021, Mühlemann et al., 2018, Paraskevis et al., 2013, Zehender et al., 2014), and thus its phylodynamic threshold has been reached while it has not been empirically established if it can be considered to be a measurably evolving population, as is the case for recent outbreaks, like SARS-CoV-2 (Duchene et al., 2020a).

For our first set of analyses we varied the width of the sampling window. We drew 100 genomes with different sampling window widths: 0 (only modern samples), up to 500, 1,000, or 5,000 years before present. Increasing the sampling window resulted in estimates of the evolutionary rate that were more precise and closer to the estimate from the complete data set (fig. 6). Here we find that the evolutionary rate is estimated to be higher for shorter sampling windows, with correspondingly older estimates for the tree height (see

Supplementary material). This pattern can be due to one or a combination of other factors influencing the inference, for example the vagaries of evolutionary rate variation in this virus, particularly time-dependency (Vrancken et al., 2017). Similarly, population structure that is unaccounted for has been shown to produce an overestimation of the evolutionary rate, because under the tree prior samples that are genetically linked are expected to have been sampled at the same point in time (Möller et al., 2018).

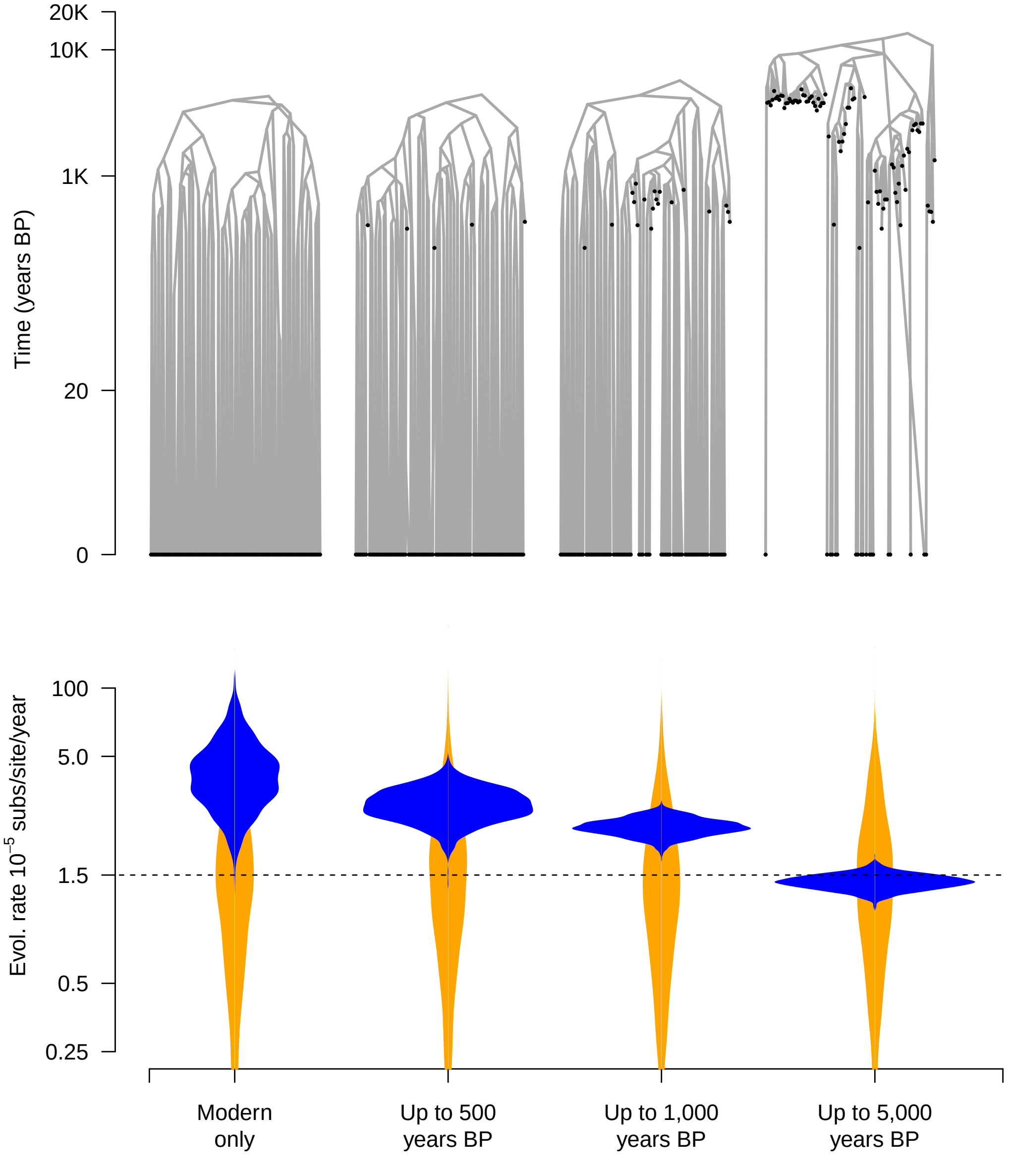


Figure 6: Results from empirical analyses of Hepatitis B virus (HBV) ancient DNA data. The phylogenetic trees correspond to highest clade credibility trees from three analyses where the data were subsampled to increase the width of the sampling window progressively. First, we only consider modern samples, then those up to 500, 1,000, and 5,000 years before present. In all cases the data sets consist of 100 genome samples. The violin plots show the posterior distribution of the evolutionary rate in blue and its corresponding prior in orange. The dashed line shows the mean evolutionary rate estimate from the complete data set.

For our second set of empirical analyses, we subsampled the data to produce a range of temporally biased sampling scenarios. We varied the proportion of modern samples, from 95% to 10%. Here the sampling window is constant because we always retain the most ancient samples. Each data set consisted of 100 genomes, such that they only differ in the distribution of modern and ancient samples (fig. 7).

The impact of temporal sampling bias was less clear than in our simulations above (figs 4 and 5). The data set with 95% modern genomes had the highest uncertainty in the evolutionary rate estimate, but uncertainty did not decrease monotonically with the proportion of modern genomes. Moreover, the posterior estimate of the evolutionary rate from the data set with 50% modern genomes deviated the most from the that obtained using the full data set. Critically, when we analysed these data using the 'misleading' prior setting we found that increasing the number of ancient samples resulted in a less influential prior (see Supplementary material). These results demonstrate that the exact impact of temporal sampling bias may be difficult to predict in practice, but that generally increasing the number of ancient samples results in data sets that are more informative, in that the difference between the prior and posterior is more pronounced than when only a few ancient samples are available.

3 Discussion

The concepts of the phyloydnamic threshold, measurably evolving populations, and temporal signal are helpful for our understanding of rapidly evolving organisms or data sets with ancient DNA. Our analyses help us disentangle the definition of these concepts and their practical implications.

The phylodynamic threshold and measurable evolution are not discrete bounds. Increasing the sampling window generally improves precision and accuracy, but there is no clear cut-off for when the estimates become accurate and objectively 'precise'.

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)). In our simulations with a reasonable prior the posterior always included the correct value, but when we set a misleading prior a sampling window of 10 times the time expected to accrue one mutation (i.e. the so called expected phylodynamic threshold) was necessary to obtain a posterior that included the correct value. As a consequence, the phylodynamic threshold and measurable evolution depend on the extent to which the data inform the posterior, which is ultimately a measure of the relative contribution of the prior and the data (via the likelihood function). Maximally uninformative priors, such as Jeffrey's prior, offer an attractive approach, but such priors can ignore useful expert knowledge, they can be particularly difficult to sample, and are not necessarily proper probability distributions (see Baele and Lemey 2014, Wang and Yang 2014 for discussions about the prior in Bayesian phylogenetics). This complicates the application of tests of temporal signal (Duchene et al., 2020c) and the analysis of data with low information content, which requires sampling from the full parameter space allowed by the prior.

Ideally, the prior and posterior of key parameters, particularly the height of the root node should overlap, while the posterior should be narrower than the prior (i.e. more informative), meaning that the data and the prior are not in conflict (for recent work on quantifying prior-data conflict see: Nott et al. 2020). In this vein, assessing the adequacy of the model and prior via predictive checks can be illuminating (McElreath, 2018).

Recent years have seen the development of a range of methods for assessing phylogenetic model adequacy (Brown and Thomson, 2018, Duchêne et al., 2018, Duchene et al., 2019, McElreath, 2018), for instance one can simulate phylogenetic trees under the posterior estimates to verify whether the height of the root node and the topology could have been generated by the model in question.

Measurably evolving populations are those for which the phylodynamic threshold has been attained and the sampling window is *sufficiently* wide. The criteria for determining the phylodynamic threshold and whether a population is measurably evolving are the same, and are typically assessed via temporal signal. Statistical tests for this purpose quantify the strength of the association between sampling times and genetic distance (Duchêne et al., 2015, Featherstone et al., 2024, Murray et al., 2016, Rambaut et al., 2016, Rieux and Balloux, 2016). That is, the degree to which the sampling times on their own constitute an informative molecular clock calibration. We contend that assessing prior sensitivity is more important than the outcome of tests of temporal signal for obtaining reliable molecular clock estimates. In fact, a poor choice of prior can mislead tests of temporal signal (Tay et al., 2024). If data are drawn from a sampling window that spans the expected phylodynamic threshold, the presence of temporal signal is likely supported by most tests, suggesting accurate estimates. Yet, if the prior used for the estimation is misleading and informative, it might actually obscure the 'correct' signal from the data. In contrast, if the data are drawn from a narrow sampling window but the prior is reasonable then the estimates may be still be reliable, despite a lack of temporal signal. It also has to be noted that an increasing sampling effort does not necessarily lead to increasingly correct inferences, because misspecification not only in prior distributions of hyper-parameters, but also in the underlying model, can introduce biase (Ferretti et al., 2024, Möller et al., 2018).

An obvious concern about molecular clock calibrations using sequence sampling times is sampling bias. We find that temporal sampling bias, where data are overwhelmingly collected at a particular period of time does not have a substantial impact in estimation accuracy on a simple coalescent model, but that increasing the number of ancient sequences can improve precision. An other form of sampling bias is when genetic diversity is not uniformly sampled or the underlying population is structured. Previous work has demonstrated that in such cases, the evolutionary rate and the age of the root node tend to be overestimated (Möller et al., 2018), a problem that diminishes when sequence data are are increasingly informative or by using a tree prior that explicitly models population structure (e.g. Kühnert et al. 2016, Müller et al. 2017).

Our study has a few limitations that have been partly addressed elsewhere. The number of sequences is fixed in most of our experiments, but it is well known that increasing the number of sequences generally means that data are more informative and thus the estimates are more precise (see Möller et al. (2018) and our fig. 4), and therefore it is likely that the width of the sampling window needed to obtain reliable estimates also depends on the number of sequences. Moreover, our simulation experiments involved a low degree of evolutionary rate variation among lineages. In this respect, it is expected that the width of the sampling window scales with the amount of dispersion in the molecular clock. In addition, our simulations are based on a simple population dynamic model, the constant coalescent. The impact of the sampling scheme and width is likely to be more complex for models with more parameters. We also assume the correct model of evolution and population growth in all simulation-based inferences. With the empirical analysis, we, however, highlight how conclusions drawn from these do not directly extend to real-world data. Rather, the isolated effects found therein describe only one of many elements impacting the inference from real data. Further scrutiny of these factors is warranted, but the main implication is that the necessary sampling time window is combination of the data set, the organism, and the model at hand.

Overall, our study elucidates some of the fundamental intricacies of molecular clock calibration strategies. We urge researchers to carefully question their model and its underlying prior assumptions, not only via tests of temporal signal, but also through careful choice of the prior, an understanding of the information content in the data, and the implications of model misspecification.

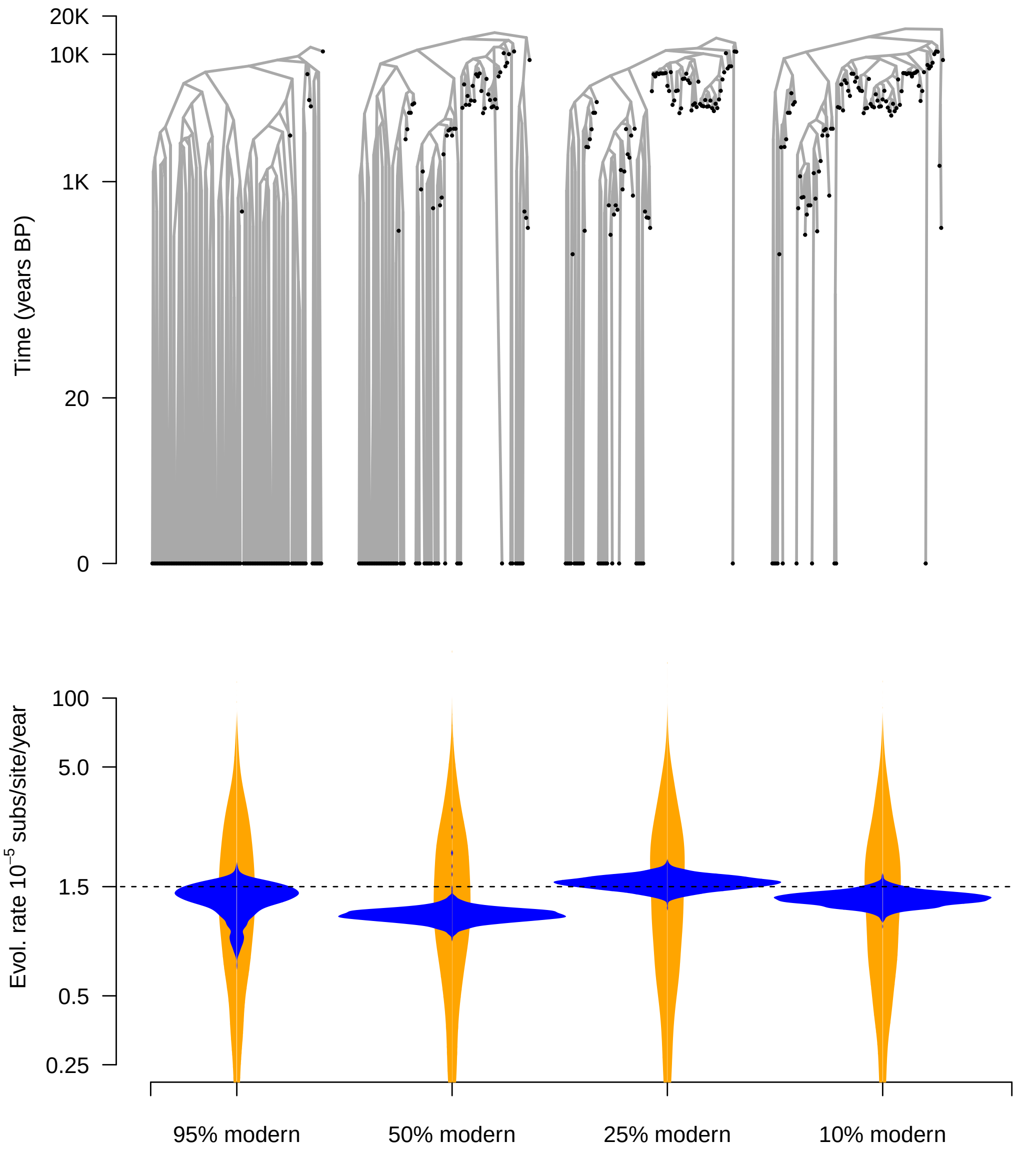


Figure 7: Results from empirical analyses of Hepatitis B virus (HBV) ancient DNA data. The phylogenetic trees correspond to highest clade credibility trees from three analyses where the data were subsampled to include an increasing number of ancient samples. First, we consider a data set for which the samples are 95% modern and the remaining 5% being the most ancient. Then, we reduce the number of modern samples to 50%, 25% and 10%, and the rest being ancient. Note that the sampling window is constant because we always retain the most ancient samples. In all cases the data sets consist of 100 genome samples. The violin plots show the posterior distribution of the evolutionary rate in blue and its corresponding prior in orange. The dashed line shows the mean evolutionary rate estimate from the complete data set.

4 Materials and methods

4.1 Simulations

4.1.1 Data generation

We simulated phylogenetic trees and the evolution of nucleotide sequences to assess the impact of varying the sampling window and temporal sampling bias. We parameterised our simulations to resemble an HBV population evolving over 10,000 years before present, as described by Kocher et al. (2021).

We generated phylogenetic trees under a coalescent process in which the population size has been constant over time using the package ReMaster (Vaughan, 2024), part of the BEAST 2 (version 2.7) software. We set the population size to 5,000, which results in trees with an average age of about 10,000 time units (years). The number of samples (i.e. tips) drawn from the trees and their ages was defined in ReMaster, according to the simulation scenario described below.

The simulation of sequence data requires trees with branch lengths in units of subs/site, instead of time. For this purpose, we multiplied the branch lengths of the simulated time tree with the rate of evolution, a lognormally distributed random variable with parameters $\mu = log(1.5 \times 10^{-5}) - \frac{0.25^2}{2}$ and $\sigma = 0.25$. This procedure equates to simulating an uncorrelated relaxed molecular clock model with an underlying lognormal distribution (Drummond et al., 2006) with mean of 1.5×10^{-5} subs/site/year and a standard deviation of 0.25 subs/site/year. Because we multiply the branch lengths in units of years by a variable in subs/site/year, the resulting trees have branch lengths in units of subs/site, formally known as phylograms, in contrast to chronograms where the branch lengths correspond to time. We obtained sequence alignments using the R package phangorn (v2.8.1) (Schliep, 2011), according to a HKY+Γ_4 substitution model, with parameters $\kappa = 2$, $\alpha = 4$, and equal base frequencies. The alignments consisted of 3,200 nucleotides to match the average genome size of HBV.

We considered the expected phylodynamic threshold of our data to be about 20 years. For our simulations where we varied the sampling window, we set the ages of 100 tips to be sampled at present (all have an age of 0), or to be drawn from a uniform distribution between 0 and 10 (1/2 of the expected phylodynamic threshold), 0 and 20, 0 and 200, or 0 and 2,000.

To investigate the impact of temporal sampling bias we initially simulated trees with 500 tips with sampling times distributed in 5 time points, with 100 tips per time point. The distribution of sampling times followed an exponential distribution with mean of 4,000, such that sampling times were concentrated towards the present. We simulated these complete trees in ReMaster. We followed the procedure above to simulate a molecular clock model and sequence alignments.

We conducted two sampling schemes for the trees with 500 tips: the 'time-uniform' scheme consisted of drawing 20 samples from each time point, whereas the 'time-biased' scheme included of mostly modern samples (90 from the present, and 5, 3, 1, and 1 from the remaining time points). For each simulation scenario we generated 100 replicates.

4.1.2 Analysis of simulated data

We analysed all data sets in BEAST 2 under a model that matched that used to generate the data: the HKY+Γ_4 substitution model, a uncorrelated relaxed molecular clock model with an underlying lognormal distribution, and a constant size coalescent tree prior. We used the default prior configuration in the program, except for the mean evolutionary rate and the population size of the coalescent (θ) for which we specifically set a prior that was reasonable but not overly informative (table 1). For the simulations with varying sampling window width we investigated prior sensitivity in detail and thus we also analysed the data under a 'misleading' prior configuration, where the prior density was deliberately concentrated on values that substantially differed from those used to generate the data.

Table 1: Prior configuration for the molecular clock model and tree prior. The substitution model parameters had the default priors in BEAST 2. Note that the mean of the Γ distribution here is $shape/rate$ and that the expected age of the root of a tree under a constant size coalescent is $2 \times \theta$.

Parameter	'Reasonable' prior	'Misleading' prior
Coalescent population size (θ)	$Exponential(mean = 5,000)$	$Exponential(mean = 50,000)$
Molecular clock mean rate (M)	$\Gamma(shape = 1.5, rate = 10^5)$	$\Gamma(shape = 1.5, rate = 10^6)$

Using the values from table 1, under the reasonable prior the mean of the lognormal distribution of branch rates has an average of 1×10^{-5} subs/site/year, with a 95% quantile width of 1.1×10^{-6} to 4.7×10^{-5}, and the expected height of the root node is roughly 10,000 years (expected time to coalescent=$2 \times \theta$ for an ultrametric tree, see Nordborg (2019)). In contrast, under the misleading prior the average evolutionary rate (mean of the lognormal distribution) is much lower, at 1×10^{-6} subs/site/year, a 95% quantile width of 1.1×10^{-7} to 4.7×10^{-6}, and with an expected height of the root node of 100,000 years. In all cases we set the sampling times for calibration. In the case of ultrametric trees, sampling times are set to the present, such that all calibration information is provided by the prior.

We used Markov chain Monte Carlo (MCMC) to sample the posterior distribution. We set the chain length to 10^8 steps, sampling every 5×10^4 steps. We deemed sufficient sampling by verifying that the effective sample size was at least 200, by using the R package CODA (version 0.19) (Plummer et al., 2006). When this criterion was not met we extended the chain length to 5×10^8 steps.

Our simulations with varying sampling window width used two possible prior configurations. To assess their impact we drew MCMC samples from the marginal prior of the evolutionary rate and the height of the root node. That is, the prior for a given parameter integrating over the prior in other parameters, the number of tips, and their heights. We obtained such samples by setting the option `sampleFromPrior="true"` in the input xml files in BEAST 2, which conducts the MCMC while ignoring the phylogenetic likelihood.

4.1.3 HBV empirical data

We selected a complete genome data set of HBV published by Kocher et al. (2021). The complete alignment included 232 genomes of length 3,344 nucleotides, with 1,807 variable sites, and 1,498 site patterns. The sampling times ranged from the present to 10,535 years before present. To investigate the impact of varying the sampling window and on temporal sampling bias we subsampled the data as described in our Results section. We analysed each data set using the same model and prior settings as in our simulations, including the use of the reasonable and misleading prior configuration.

5 Data availability

Computer code, analysis files, and data sets in this study are available at:

`https://github.com/sebastianduchene/phylo_threshold_code_data`

6 Competing interests

None.

7 Acknowledgments

Pending.

8 Funding

This work received funding from the Inception program (Investissement d'Avenir grant ANR-16-CONV-0005 awarded to SD) and the Australian National Health and Medical Research Council (2017284 awarded to SD),

References

A. Andrades Valtueña, G. U. Neumann, M. A. Spyrou, L. Musralina, F. Aron, A. Beisenov, A. B. Belinskiy, K. I. Bos, A. Buzhilova, M. Conrad, et al. Stone age yersinia pestis genomes shed light on the early evolution, diversity, and ecology of plague. *Proceedings of the National Academy of Sciences*, 119(17): e2116722119, 2022.

G. Baele and P. Lemey. Bayesian model selection in phylogenetics and genealogy-based population genetics. In M. Chen, K. L, and L. PO, editors, *Bayesian phylogenetics, methods, algorithms, and applications*, chapter 4, pages 59–93. CPC Press, Boca Raton (Florida), 2014.

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.

M. F. Boni, P. Lemey, X. Jiang, T. T.-Y. Lam, B. W. Perry, T. A. Castoe, A. Rambaut, and D. L. Robertson. Evolutionary origins of the sars-cov-2 sarbecovirus lineage responsible for the covid-19 pandemic. *Nature microbiology*, 5(11):1408–1417, 2020.

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.

J. M. Brown and R. C. Thomson. Evaluating model performance in evolutionary biology. *Annual Review of Ecology, Evolution, and Systematics*, 49(1):95–114, 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.

A. M. Devault, G. B. Golding, N. Waglechner, J. M. Enk, M. Kuch, J. H. Tien, M. Shi, D. N. Fisman, A. N. Dhody, S. Forrest, et al. Second-pandemic strain of vibrio cholerae from the philadelphia cholera outbreak of 1849. *New England Journal of Medicine*, 370(4):334–340, 2014.

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.

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.

D. A. Duchêne, S. Duchêne, and S. Y. Ho. Phylomad: efficient assessment of phylogenomic model adequacy. *Bioinformatics*, 34(13):2300–2301, 2018.

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, R. Bouckaert, D. A. Duchene, T. Stadler, and A. J. Drummond. Phylodynamic model adequacy using posterior predictive simulations. *Systematic biology*, 68(2):358–364, 2019.

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.

K. Eaton, L. Featherstone, S. Duchene, A. G. Carmichael, N. Varlık, G. B. Golding, E. C. Holmes, and H. N. Poinar. Plagued by a cryptic clock: insight and issues from the global phylogeny of yersinia pestis. *Communications Biology*, 6(1):23, 2023.

L. A. Featherstone, A. Rambaut, S. Duchene, and W. Wirth. Clockor2: Inferring global and local strict molecular clocks using root-to-tip regression. *Systematic biology*, 73(3):623–628, 2024.

L. Ferretti, T. Golubchik, F. Di Lauro, M. Ghafari, J. Villabona-Arenas, K. E. Atkins, C. Fraser, and M. Hall. Biased estimates of phylogenetic branch lengths resulting from the discretised gamma model of site rate heterogeneity. *bioRxiv*, pages 2024–08, 2024.

A. Gavryushkina, T. A. Heath, D. T. Ksepka, T. Stadler, D. Welch, and A. J. Drummond. Bayesian total-evidence dating reveals the recent crown radiation of penguins. *Systematic biology*, 66(1):57–73, 2017.

M. Ghafari, L. du Plessis, J. Raghwani, S. Bhatt, B. Xu, O. G. Pybus, and A. Katzourakis. Purifying selection determines the short-term time dependency of evolutionary rates in sars-cov-2 and ph1n1 influenza. *Molecular Biology and Evolution*, 39(2):msac009, 2022.

N. Gharbi, E. Rousseau, and T. Wirth. Clock rates and bayesian evaluation of temporal signal. In *Phylogenomics*, pages 153–175. Elsevier, 2024.

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.

S. Guindon. Rates and rocks: strengths and weaknesses of molecular dating methods. *Frontiers in Genetics*, 11:526, 2020.

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.

D. Kühnert, T. Stadler, T. G. Vaughan, and A. J. Drummond. Phylodynamics with migration: a computational framework to quantify population structure from genomic data. *Molecular biology and evolution*, 33 (8):2102–2116, 2016.

D. Kühnert, M. Coscolla, D. Brites, D. Stucki, J. Metcalfe, L. Fenner, S. Gagneux, and T. Stadler. Tuberculosis outbreak investigation using phylodynamic analysis. *Epidemics*, 25:47–53, 2018.

S. A. Locarnini, M. Littlejohn, and L. K. Yuen. Origins and evolution of the primate hepatitis b virus. *Frontiers in Microbiology*, 12:653684, 2021.

R. McElreath. *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC, 2018.

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

F. K. Mendes, R. Bouckaert, L. M. Carvalho, and A. J. Drummond. How to validate a bayesian evolutionary model. *Systematic Biology*, 74(1):158–175, 2025.

M. Merker, J.-P. Rasigade, M. Barbier, H. Cox, S. Feuerriegel, T. A. Kohl, E. Shitikov, K. Klaos, C. Gaudin, R. Antoine, et al. Transcontinental spread and evolution of mycobacterium tuberculosis w148 european/russian clade toward extensively drug resistant tuberculosis. *Nature Communications*, 13(1):5105, 2022.

S. Möller, L. du Plessis, and T. Stadler. Impact of the tree prior on estimating clock rates during epidemic outbreaks. *Proceedings of the National Academy of Sciences*, 115(16):4200–4205, 2018.

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.

N. F. Müller, D. A. Rasmussen, and T. Stadler. The structured coalescent and its approximations. *Molecular biology and evolution*, 34(11):2970–2981, 2017.

G. G. Murray, F. Wang, E. M. Harrison, G. K. Paterson, A. E. Mather, S. R. Harris, M. A. Holmes, A. Rambaut, and J. J. Welch. The effect of genetic structure on molecular dating and tests for temporal signal. *Methods in Ecology and Evolution*, 7(1):80–89, 2016.

M. Nordborg. Coalescent theory. *Handbook of Statistical Genomics: Two Volume Set*, pages 145–30, 2019.

D. J. Nott, X. Wang, M. Evans, and B.-G. Englert. Checking for prior-data conflict using prior-to-posterior divergences. *Statistical Science*, 35(2):234–253, 2020.

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.

M. Plummer, N. Best, K. Cowles, K. Vines, et al. Coda: convergence diagnosis and output analysis for mcmc. *R news*, 6(1):7–11, 2006.

A. Rambaut, T. T. Lam, L. Max Carvalho, and O. G. Pybus. Exploring the temporal structure of heterochronous sequences using tempest (formerly path-o-gen). *Virus evolution*, 2(1):vew007, 2016.

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.

F. Ronquist, N. Lartillot, and M. J. Phillips. Closing the gap between rocks and clocks using total-evidence dating. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1699):20150136, 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.

K. P. Schliep. phangorn: phylogenetic analysis in r. *Bioinformatics*, 27(4):592–593, 2011.

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, 2019a.

M. A. Spyrou, M. Keller, R. I. Tukhbatova, C. L. Scheib, E. A. Nelson, A. Andrades Valtueña, G. U. Neumann, D. Walker, A. Alterauge, N. Carty, et al. Phylogeography of the second plague pandemic revealed through analysis of historical yersinia pestis genomes. *Nature communications*, 10(1):4470, 2019b.

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.

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.

T. G. Vaughan. Remaster: improved phylodynamic simulation for beast 2.7. *Bioinformatics*, 40(1):btae015, 2024.

B. Vrancken, M. A. Suchard, and P. Lemey. Accurate quantification of within-and between-host hbv evolutionary rates requires explicit transmission chain modelling. *Virus Evolution*, 3(2):vex028, 2017.

Y. Wang and Z. Yang. Priors in bayesian phylogenetics. *Bayesian phylogenetics: methods, algorithms, and applications*, pages 5–24, 2014.

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.

G. Zehender, E. Ebranati, E. Gabanelli, C. Sorrentino, A. L. Presti, E. Tanzi, M. Ciccozzi, and M. Galli. Enigmatic origin of hepatitis b virus: an ancient travelling companion or a recent encounter? *World Journal of Gastroenterology: WJG*, 20(24):7622, 2014.

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