**Responses for *Assessing the effect of model specification and prior sensitivity on Bayesian tests of temporal signal*  in Plos Computational Biology**

As with all papers reviewed by the journal, your manuscript was reviewed by members of the editorial board and by several independent reviewers. In light of the reviews (below this email), we would like to invite the resubmission of a significantly-revised version that takes into account the reviewers' comments.

We cannot make any decision about publication until we have seen the revised manuscript and your response to the reviewers' comments. Your revised manuscript is also likely to be sent to reviewers for further evaluation.

**Response:** *We would like to thank the Editors and the four reviewers who gave feedback for our manuscript. We have attempted to address all their comments, as explained in our point by point response below. Our main changes are highlighted in red in the submitted version with tracked changes.*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

REVIEWER 1

The paper is well presented and the arguments are easy to follow. I believe the problem studied is also an important one because the community needs guidance on how to set priors in Bayesian phylogenetic models. I have two major comments and a few minor ones.  
  
A. I understand the explanation of tree extension as the tree height of the posterior tree of heterochronous model on isochronous data is abnormally high, and the reason is that the prior on effective population size \theta has different unit for heterochronous model and isochronous model. However, I do not understand why Exponential prior on \theta could achieve perfect accuracy in Table 4. If mismatch of unit is the problem, shouldn’t this problem apply to all of Exponential, gamma, and Log-normal prior? If exponential distribution works well, maybe a simpler conclusion would just be “use exponential prior” instead of doing prior sensitivity and predictive analysis? Also, if there is a mismatch of units, I wonder if you could scale \theta during posterior sampling to the correct unit and solve this problem.

**Response:** *This is an important point, which we have clarified in the manuscript. It is true that we observed less tree extension with the exponential prior. However, there may be data sets where we expect very old divergences, meaning that this prior would be inappropriate. For example, we conducted an additional set of simulations under a population size of 100, where the exponential prior with a mean of 1.0 might be inappropriate. We explain this in the Discussion:*

In our simulation study, the exponential prior on θ mitigated the problem of tree extension, but this particular prior might not be reasonable for some data sets, such as those where we expect very old sequence divergence.

*It is also true that instead of setting a different clock rate to fix the problem of mismatching units one can rescale θ. We explain this point in the Discussion as follows:*

A more tractable approach is to conduct isochronous analyses by fixing the evolutionary rate to a value within the expected order of magnitude, as we have done here, or rescaling the parameters of the tree prior, such as θ accordingly (in units that are the proportional to time−1; see [29]).

B. In the abstract, the authors mention that one suggestion to make BETS robust to priors is to include internal node constraints or informative molecular clock rate distributions. However I don’t see an example of this in the current manuscript. It would be good to see an example of how internal node calibration helps correcting BETS error.

**Response:** *We have rephrased our Abstract to clarify this point about additional calibration information, which is not a recommendation for BETS, but rather for estimating evolutionary rates and timescales when there is not sufficient temporal signal in the data. The corresponding section in the Abstract now reads:*

In consequence, we recommend: (i) using prior predictive simulations to determine whether the prior generates a reasonable expectation of parameters of interest, such as the evolutionary rate and age of the root node, (ii) conducting prior sensitivity analyses to assess the robustness of the posterior to the choice of prior, and (iii) selecting a molecular clock model that reasonably describes the evolutionary process. preferring molecular clock models that reasonably describes the evolutionary process.

*Moreover, in the Discussion we clarified this argument and point the reader to in depth studies of the use of internal node calibrations:*

Our use of hard bounds on the root height is effectively an internal node constraint, but similar calibration priors in the form of a parametric distribution on any other internal node or nodes could be used [47, 52].

Minor comments:  
1. I wonder why the shape of the gamma prior is set to a small 0.001 and scale is a large 1000. Is this standard practice as opposed to, say, shape=1 and scale=1?

**Response:** Yes, this is one of the recommended priors for the population size parameter of coalescent models, as explained in Vogels et al. 2023 and cited in the text. We added a sentence in our Results section to justify this choice:

… we selected three prior distributions, an exponential, Γ (Gamma), and log-normal, that have been used in recent literature [26] (this Γ prior on the scaled population size is the default in the BEAST1.10.5), and as as shown in Table 1.

2. I understand that coalescent process is used as tree prior because it has fixed number of taxa, but is it possible to use birth-death process as a prior?

**Response:** *This is an important point. The birth-death is indeed a commonly used tree prior. However, the birth-death model with sampling through time (i.e. heterochronous) uses the sequence sampling times as information, whereas the isochronous version of the model is conditioned on a sampling proportion (or a compound parameter that includes the sampling proportion). Thus, it is not clear whether these two models would be comparable. We have added a paragraph in the Discussion and added relevant references to elaborate on this matter:*

[...] Although birth-death tree priors could conceivably be used in the context of BETS (they are usually fully parametric and have proper prior distributions), these models explicitly use sampling time information as data [47], such that more work is needed to determine whether comparing the log marginal likelihoods of the isochronous and heterochronous versions of the model is valid.

3. One suggestion in Discussion is to use relaxed model instead of a strict one for the clock model. However, if you allow mutation rate to differ among lineages, is it possible that a true temporal signal is explained by a higher mutation rate on one lineage than the other, and the model conclude the data as isochronous?

**Response:** *This is correct and we have amended this recommendation in the Abstract and Discussion. In particular we suggest careful model selection, which includes verifying whether the data include very divergent sequences or other potential confounders. In the Abstract:*

[...] (iii) selecting a molecular clock model that reasonably descirbes the evolutionary process. ~~preferring molecular clock models that capture rate variation among lineages~~.

*And in the Discussion:*

However, careful molecular clock model selection is advisable for tests of temporal signal and molecular dating in general.

Writing errors by line  
l.21 “i.e. that the the interval of time”  
l.29 “whether a sampled population is measurably evolving behaviour”  
l.82 “but it does is not”  
l.201 “branch lengths are in not in”  
Fig. 6 “the grey histograms for correspond to”  
Fig. S6 reference missing

**Response:** *These typographic errors have been corrected.*

REVIEWER 2

Major comments:  
  
Methods section:  
About the model selection: consider adding a detailed rationale for choosing specific priors like the exponential, gamma, and log-normal distributions. For example, discuss why these particular priors were chosen and how they relate to biological realities in microbial evolution.

**Response:** *We included additional text and references for the interpretation of these parameters. Briefly, they have different interpretations in epidemiology and in population genetics. The prior distributions are the default in the latest version of BEAST1.10 and have been adopted in recent studies, which we reference here. Our edit is in the Results section as follows:*

*[...]* we selected three prior distributions, an exponential, Γ (Gamma), and log-normal, that have been used in recent literature [27] (this Γ prior on the scaled population size is the default in the BEAST1.10.5), and as as shown in Table 1 (for the epidemiological or demographic interpretations of the parameters of the tree prior we refer the reader to [30, 33]).

For the three empirical data sets, why were these particular datasets chosen, and how do they represent a broad range of scenarios in microbial evolution?

**Response:** *We included a justification in the Methods section to explain our choice of empirical data sets:*

We chose these data sets because they included ancient samples (*T. pallidum* and *V. cholerae*) and very divergent sequences (*Powassan virus*), such that they have the complexity of many real-world analyses.

For simulation, please specify the parameters used for the JC substitution model. Also, how were the sequence alignments generated?

**Response**: The JC substitution model has no free parameters as rates are equal - a fixed parameter of 0.25. In the Methods section we added the following in parentheses:

We used a JC substitution model (all exchangeability parameters are equal).

What are the assumptions during the modeling? How do they relate to real-world biological assumptions? Besides, the potential limitations of the assumptions and how they might affect the generalizability of the results should be considered.

**Response:** Our simulations represent a typical microbial data set with samples collected over epochs (e.g. sampling blitzes), although we make simplifying assumptions for statistical tractability, such as a simple substitution model. We explained this in the Methods section:

Our simulations simplify the complexity or empirical data (e.g. simple substitution model, no recombination), but they represent a realistic situation where data were collected across four discrete sampling periods, which resembles sequencing blitzes [61], or archaeological sampling of strata [62].

Results section:  
For the empirical data analyses, the authors should provide more interpretation of the results in the context of existing literature. For example, how do their findings compare with previous studies? Are the observed impacts of different priors on BETS results consistent with what has been reported in other contexts?

**Response:** *We make several references to previous analyses of these data:*

‘Some empirical data sets yielded unclear evidence for temporal signal according to BETS, such as the *T. pallidium* data set that we reanalysed.’

and

‘Our empirical data set of V. cholerae, which had evidence of temporal signal under all prior conditions, consistent with previous analyses [25], also displayed strong evidence for temporal signal with a hard bound on of 500 years before present on the root height (all log Bayes factors were at least 200 in favour of temporal signal).’

Since the authors have concluded the recommendations for temporal signal analyses, it might be useful to include specific examples and/or case studies where following these recommendations could significantly impact the outcomes of molecular clock analyses. This could help to corroborate their recommendations and make them more relatable for researchers within the field.

**Response:** *The reviewer makes a valuable point. However, to determine the quality of previous studies that assess temporal signal would we would need to conduct critical evaluations of their methods used, which is beyond the scope of our study.*

Discussion section:  
It could be helpful to compare the Bayes to non-Bayes methods briefly.

**Response:** *This is an important point. We have added a sentence to explain that the date-randomisation test, which suffers from similar problems, is also available in a non-Bayesian setting. However, in the context of BETS, we are only aware of a Bayesian implementation:*

Here, the sequence sampling times are permuted a number of times, the evolutionary rate is re-estimated, and the data are considered to have temporal signal if the estimates from the permutations do not overlap with that from the correct sampling times. This test exists in a Bayesian and maximum-likelihood setting, where point estimates are used instead of posterior distribution [45]. Notably, when the data do have temporal signal, the estimates from the permutations are substantially lower, implying older times to the most recent common ancestor [14]. Therefore the inclusion of incorrect sampling times, whether the data set is truly isochronous or not, may result in tree extension as a means of compensating for the likelihood penalty imposed by incorrect sampling times.

It would be better if the authors provided more specific guidance on how researchers should approach the proposed recommendations. For example, what range of priors should be considered in typical analyses, and what criteria should be used to determine whether a prior is "reasonable"?

**Response:** *We added some text in the discussion to explain how one might determine what constitutes a ‘reasonable’ prior for an empirical data set:*

As a case in point, for a data set collected from a recently emerging pathogen the exponential prior on θ will likely generate more reasonable trees than the log-normal. Conversely, the log-normal prior, as parameterised here, might generate more reasonable trees for a data set involving ancient DNA samples.

Apart from these, the authors could consider adding an appendix or supplementary material that explains the key concepts/terminologies used in this manuscript. Also, a schematic diagram/flowchart could be helpful for illustrating the effects of different prior choices on tree topology and temporal signal detection.

**Response:** *We have clarified the terminology throughout the manuscript and included references to foundational text books and reference articles. For example, in the Introduction:*

[...] under the earliest and simplest molecular clock model, known as the strict clock, substitutions are assumed to accumulate at a constant rate over time and across lineages [1] (for an in-depth introduction to molecular clocks see [2, 3]).

For clarity, an isochronous tree is one where the sampling times are identical, also known as an `ultrametric' tree (the distance from the root to each of the tip is the same) [21, 22], and thus a heterochronous tree is one where the sampling times are different.

Minor comments:

Abstract section:  
consider briefly mentioning the practical implications of the study's recommendations, especially for researchers working with microbial evolution data.

**Response:** *Please see our responses to the other reviewers. We have substantialy edited the Discussion section to provide more practical advice. For example:*

In our simulation study, using hard bounds on the root height alleviated the problem of tree extension when the prior on θ favoured very old trees. In empirical data, however, one may not want to make such strong statements, and prior predictive simulations can help determine whether the tree prior and associated parameters produce trees with sensible root heights (e.g. the exponential prior on θ had low type I and type II errors in our simulations). As a case in point, for a data set collected from a recently emerging pathogen the exponential prior on θ will likely generate more reasonable trees than the log-normal. Conversely, the log-normal prior, as parameterised here, might generate more reasonable trees for a data set involving ancient DNA samples.

Typo: last line “ensuring the the”

**Response**: *We have fixed this typo.*

Introduction:  
the explanation of the "temporal signal" and its significance could be expanded for clarity. In detail, define "temporal signal" in simpler terms, and perhaps with an example of when a temporal signal might be absent or present.

**Response:** *We have improved the terminology in several sections, particularly the Introduction. Please see our detailed responses above. With respect to this particular point, we state:*

‘[...] verify that the data carry sufficient information for molecular dating, a practice referred to as evaluation of temporal signal’

It could be helpful to expand on the background, especially for readers who may not be deeply versed in the technical aspects of Bayesian inference and molecular clocks. E.g., including a brief primer on key concepts such as "molecular clocks," "tree priors," and "Bayes factors" would improve the accessibility of the manuscript.

**Response:** *We have substantially expanded our list of references and our terminology. Please see our detailed points to other comments above. We also refer the reviewer to ‘Primer’ papers on a range of topics. For example,*

*[...] and as shown in Table 1 (for the epidemiological or demographic interpretations of the parameters of the tree prior we refer the reader to [36, 39]).*

Literature review: The review of existing methods for evaluating temporal signals (e.g., root-to-tip regression, date randomization tests) is comprehensive. However, it would be helpful to include a brief comparison of these methods with BETS. Also, explain the rationale behind choosing BETS as the focus of this study. This could help highlight the novelty of this study.

**Response:** *We included some text in the introduction to explain this point:*

Although date randomisation tests are computationally efficient [21], they make it difficult to exploit the full extent of Bayesian phylodynamic models. For example, it is not clear how to incorporate uncertainty sampling times. A more powerful alternative that is fully Bayesian is the Bayesian Evaluation of Temporal Signal (BETS) [22].

Typo: line 21 “that the the interval”

**Response:** *Fixed.*

Results section:  
When discussing the impact of the tree prior to BETS results, consider explaining the log Bayes factors in more detail. For readers who may not be familiar with Bayesian statistics, consider including a brief reminder of what "strong" or "weak" log Bayes factor means.

**Response***: We have added an explanation of log Bayes factors from Kass and Raftery as seen below:*

When considering log Bayes Factors, a value of 3 or greater is deemed ‘substantial’ evidence, while any value below is deemed as ‘not worth more than a bare mention’ (Kass and Raftery). Further, any value greater than and including 5 is considered “strong” evidence.

Type: line 83 “the the biological process”

**Response:** *Corrected.*

Figures:  
Line 475: latex rendering error

**Response:** *We have corrected this error.*

In conclusion, this study provides valuable insights into the effects of model specification and prior sensitivity on Bayesian tests of temporal signal, with significant implications for the accuracy of molecular clock estimates. While the study is robust and offers practical recommendations, minor revisions to improve clarity could enhance its impact. Overall, this is a strong contribution to the field of Bayesian phylogenetics and molecular evolution.

**Response:** *We would like to thank the reviewer for their positive feedback.*

REVIEWER 3

In Table 1, the selection criteria for the parameters of the prior distributions are not clearly defined. Could you elaborate on how these parameters were chosen?

**Response:** *We have added additional explanations for our choice of priors. E.g.*

Instead, we selected three prior distributions, an exponential, Γ (Gamma), and log-normal, that have been used in recent literature [32] (this Γ prior on the scaled population size is the default in the BEAST1.10.5), and as as shown in Table 1 (for the epidemiological or demographic interpretations of the parameters of the tree prior we refer the reader to [35, 38]).

Regarding the simulation study, it would be insightful to consider simulations based on real data parameters, for instance, by using parameter estimates from empirical data to simulate the tree. This approach might provide a more realistic assessment of the model's performance. Furthermore, the use of only 10 replicates in the simulations raises concerns about the robustness of the results. Increasing the number of replicates could enhance the reliability of the findings.

**Response:** *The reviewer makes an important point about using more realistic parameters for simulations. We included additional experiments for data that resembled the empirical* Vibrio cholerae *data. We chose to conduct these simulations under isochronous trees only, which is the situation that resulted in most false positives (incorrectly classifying isochronous data as heterochronous). Our simulations are described in Methods:*

We also generated a set of simulations under conditions aimed at mimicking the empirical *V. cholerae* data set. We focused our attention on isochronous simulations, which resulted in the highest type I errors in all other conditions. We simulated phylogenetics trees with 50 taxa under a coalescent process with a constant population size, θ, of 100.0. The root heights of these trees ranged between 60.0 and 248.0 units of time. To specify an artificial set of sampling times we sampled numbers from an exponential distribution with mean of 1/10th of the root height. This distribution results in sampling times that spanned about 1/3rd of the height of the tree, with most samples collected close to the present. We set the clock rate and genome length to produce a similar number of site patterns as previous estimates for these data [30]. We analysed these data using the same methods as the rest of the simulations above, except that in this case we only use the exponential prior on θ with a mean of 100.0, with and without hard bounds.

*The resulting analyses are described in the Results section:*

To investigate the phenomenon of tree extension in more realistic conditions, we conducted ten simulations of data that resembled our *V. cholerae* empirical data (similar numbers of variable sites and distribution of sampling times), and evaluated it using BETS under an exponential prior on θ. We only considered isochronous trees under a strict clock, because this is the situation where we obtained the largest number of false positives in other simulation conditions (type I errors). In the absence of hard bounds on the root height only five out of ten data sets resulted in false positives (‘substantial’ support for temporal signal), with those that were misclassified displaying tree extension. The use of hard bounds (five fold older than the true root height) reduced the number of false positives to one.

*For clarity, our simulation study involved 1070 analyses in total, with 20 simulation replicates (ten under a strict and 10 under a relaxed clock) and with each necessitating 18 analyses as follows: two clock models, three configurations of sampling times (isochronous and heterochronous with and without hard bounds on the root height), and three priors on the population size. Moreover, for a subset of simulations we used a different prior, the exponential-growth coalescent, that entailed 360 additional analyses. We also included a new set of simulations under a different sampling regime (ten replicates) for a total of 1260 analyses.*

The performance of the gamma and log-normal distributions in the simulation study, particularly under the strict clock model in isochronous trees, suggests high type I error rates. It may be beneficial to consider additional distribution such as beta distribution. This could be particularly useful for variables bound within a specific range of positive values.

**Response:** *This is an interesting alternative to the hard bounds we specify here, which result in priors that are compound distributions. However, in the context of the population size, we would need to transform the Beta distribution, such that its domain is not 0 to 1. For other variables, it is indeed the case that Beta distributions are used (e.g. sampling proportions in birth-death sampling models).*

The interpretation of the polygon figure is unclear. A better explanation or an alternative visual representation is needed to clarify what it implies when a point is close to the center or a corner. Figures 2 and 3 should be merged if the representation of the isochronous case in Figure 2 is absent, as mentioned in the caption.

**Response:** *We have combined these figures and included additional sentences in the first polygon figure (Fig 1) to explain what they mean as follows:*

The corners correspond to models (a combination of molecular clock model and the inclusion or exclusion of sampling times). The outermost dashed lines are for the highest marginal likelihood recorded and the extent to which the polygons are deformed denotes relative model support (e.g. a perfect square would imply that all models are equally well supported). A corner that falls close to the centre would correspond to a model that has low support.

Similarly, Tables 3 and 5, and Tables 4 and 6 should be merged or clarified to avoid redundancy and confusion. The color representation in Figures 2 and 3 needs adjustment since there is no blue color present where it is indicated.

**Response:** *We improved the legend in the corresponding figures to improve clarity. We used purple in some of our figures to facilitate visualisation for a colour blind audience. We also merged data from the corresponding tables.*

The manuscript indicates that the exponential prior outperforms others in all scenarios. Given this, a detailed discussion on the necessity and impact of using other priors would be beneficial. This could help in understanding the trade-offs involved in the choice of priors and their implications on model outcomes.

**Response:** *It is true that the exponential prior outperforms the others in a range of settings, but we have included new simulations, at the request of another reviewer, that demonstrate that this prior is not ideal for all situations. Instead we recommend that the prior be chosen depending on the data set. Please see our new section under* Simulation experiments *for this discussion.*

Incorporating a discussion on the use of phylogenetic trees in statistical inference (e.g., differential abundance testing [1] and mediation analysis [2]) for microbiome analysis would provide a broader context and applicability of the methods discussed.

**Response:** *We included these two references in the first paragraph of the introduction and state the wide range of applications of phylogenetics:*

Phylogenetic methods are particularly useful and have a wide range of applications, from resolving the timing of evolutionary relationships [1], to quantifying differential abundance testing and mediation analyses [2, 3].

Minor typographical correction is needed in the caption of Figure S6 for better clarity.

**Response**: *We have corrected this.*

[1] Zhou, C., Zhao, H., & Wang, T. (2021). Transformation and differential abundance analysis of microbiome data incorporating phylogeny. Bioinformatics, 37(24), 4652-4660.  
[2] Hong, Q., Chen, G., & Tang, Z. Z. (2023). PhyloMed: a phylogeny-based test of mediation effect in microbiome. Genome Biology, 24(1), 72.

REVIEWER 4

[1] While I certainly appreciate the efforts to clearly outline the background and the rationale behind each analysis, the explanation of how BETS operates remains somewhat unclear, even for someone with a general Bayesian background. In typical Bayesian model papers, it's common to include a plate notation that specifies the prior, data generation processes and the observed data. In this manuscript, however, the textual descriptions of the prior and data generation processes are difficult to follow, especially for readers who may not be familiar with the domain-specific terminology. For example, concepts such as constant-size and exponential-size coalescence, tree priors, and the definition of a 'tree' are not clearly explained. While the later sections suggest that the tree involves parameters such as tree height, evolutionary rate, and root height, the absence of plate notation or mathematical equations makes it challenging to fully grasp the underlying logic.

**Response:** *We have included substantial mathematical notation in our Results section to illustrate the interactions between parameters as follows:*

To illustrate the interactions between parameters in the posterior expression for a full Bayesian phylogenetic model, consider expression (1), [...]

p(T, θ, κ, r|D) ∝ p(D|T, r, κ)p(T |θ)p(θ)p(µ)p(r)

[...] known as the phylodynamic likelihood and here it is the probability of the tree given θ. Note that p(r)~Γ(0.5, tree length, implying that p(r) is also dependent on T and therefore θ, via p(θ|T).

*We also point to the reader to a foundational paper on plate notation in phylogenetics (reference 34 in this section).*

[2] Another related concern is the lack of notation regarding the shape of each random variable; they appear to be univariate and independent of one another based on the textual descriptions but I am not entirely sure. This issue could be resolved by providing clear mathematical likelihood functions.

**Response:** Please see our point above where we include an expression for the posterior that illustrates the interactions between parameters.

[3] When the author refers to a simulation dataset as one 'where the data-generating process was well-understood'. What does the data generation process look like? Is it biased toward a specific type of prior choice?

**Response:** *We edited this sentence as follows:*

where the model used to analyse the data matches the data generating process.~~data generating process was well understood~~

[4] One aspect that's difficult to grasp is why certain prior choices are more likely to result in tree extension than others.

**Response:** *We included a more detailed explanation of this phenomenon and a supplementary figure that shows that the inclusion of sampling times may not be penalised in the phylogenetic likelihood, resulting in incorrect model classification. Please see Supplementary material Fig S6 and the following text in the Discussion:*

We find that tree extension is the most probable reason for type I errors in BETS, because it reduces the sampling window relative to the root height, such that the trees with sampling times have phylogenetic likelihoods that are very similar to those of ultrametric trees (Supplementary material). Because the phylogenetic likelihood has substantial weight on the marginal likelihood, this phenomenon can mislead model selection. Verifying that the prior does not favour trees that are implausibly old can mitigate this problem.

[5] I'm a bit confused about the prior predictive simulation section, but I suspect this might relate back to my earlier point [1] about the lack of explanation of the data generation process. A broader question for this section is: what is the conclusion after conducting this analysis?

**Response:** *We added substantial text in the Discussion to explain the value of prior predictive simulations in more detail. E.g:*

In our simulation study, using hard bounds on the root height alleviated the problem of tree extension when the prior on θ favoured very old trees. In empirical data, however, one may not want to make such strong statements, and prior predictive simulations can help determine whether the tree prior and associated parameters produce trees with sensible root heights (e.g. the exponential prior on θ had low type I and type II errors in our simulations). As a case in point, for a data set collected from a recently emerging pathogen the exponential prior on θ will likely generate more reasonable trees than the log-normal. Conversely, the log-normal prior, as parameterised here, might generate more reasonable trees for a data set involving ancient DNA samples.

Minor:  
  
[1] At present, the GitHub repository contains a collection of input files and XML files, but it's unclear how someone could use them to reproduce the results. It would be helpful to include instructions on how to run the code and obtain the results described in the paper.

**Response:***We have updated Github repository with code snippet, software versions, and tutorial link.*