Review by anonymous reviewer 1, 31 Jul 2024 09:09

> In this revised manuscript, the authors have addressed most of the comments raised in the previous round of reviews. I still have some major concerns about the study and believe that further work is needed, although the authors can address many of these concerns by adding further discussion and caveats to the text.

> The revised manuscript presents some results that are potentially useful, but still needs to acknowledge previous work more comprehensively and needs clearer differentiation from previous analyses. A lot of previous work has already been done on the factors affecting the performance of molecular dating. In particular, there have been many studies based on simulated data, not just the earlier work by Duchene, Ho, Schrago, but also the more recent work by Carruthers et al. that focused on among-lineage rate variation and tree incongruence (which has not been cited in the present study).

Thank you for this second round of review. We agree with the concerns expressed and we have cited the explicitly suggested articles and also additional relevant studies based on simulated data. We have however tried to keep this selection narrow and focused, while referring to the existing reviews for more information. We think that our study specificity lies in studying genes independently instead of concatenations. Among other additions, the following has been added in the introduction to make this point:
«Assessing the adequacy of different models of rate variation has been investigated by a number of simulation-based studies (Aris-Brosou & Yang 2002; Ho et al. 2005; Rannala & Yang 2007; Battistuzi et al. 2010; dos Reis & Yang 2014; Duchêne et al. 2015), but with a focus on multiple-loci datasets and speciation dates.»
The Discussion has been expanded as well, with regard to all reviewers comments.
We provide the manuscript file with track changes for easier screening.

> ABSTRACT

> L14. There are still several mentions of gene duplications and transfer, but these are not very relevant to the present study. It would be better to omit these to make the purpose and scope of the study clearer.

We removed some mentions, but without mentioning them at all it seems harder to justify using single gene trees separately. Although our study does not consider duplications or transfers, one interesting use case would be to date such events that are gene-tree specific. For this reason we slightly updated the last abstract sentence by inserting “events only observed in single gene trees”.

> L31. I am not sure that “constrained genes evolve more constantly” can be reported as a general result. It seems counterintuitive because we would expect selection to cause rate variation among lineages.

We were stating this as an hypothesis but we cannot indeed clearly identify a mechanism. It seems however relevant to note that these categories are «expected to be under strong negative selection.» so we inserted that phrase.

> L34. The authors should tone down the claim that “relaxed clock inferences are mainly driven by the tree prior when calibrations are lacking and rate heterogeneity is high”. The authors have not compared different models of rate variation, which is likely to be a more important factor (as shown in previous studies).

We now specify that we observe it in the case of the uncorrelated log-normal model.

> L36. The authors need to temper their conclusion that “Our study finally provides a general scale of parameters that influence the dating precision and accuracy”. The study does not investigate what are arguably the two most important factors in molecular dating: the calibrations and the model of rate variation.

The phrasing was too vague as we were only referring to the measurements of gene tree features (in table 1). We updated to «Our study finally reports the scale of the gene tree features [...]».

> INTRODUCTION

> L96. This description of the white noise model needs some revision. In this model, the variance of the rate increases linearly with branch length. So the variance is larger, not smaller, on longer branches. Although Lepage et al. (2007) stated that the mean rate under the white noise model is “expected to have a smaller variance over longer branches”, they meant “smaller” in comparison to the uncorrelated gamma model (under which the variance increases quadratically with time).

To be honest I am confused and I would love to have a discussion on this for clarification. I will stick to a phrasing that is close to the actual paragraph of Lepage et al. (2007) we are both referring too. The updated text is:

«In the uncorrelated case, the branch-wise rates are most commonly modeled as a lognormal, gamma or exponential distribution, and the branch lengths in units of substitution are usually obtained by multiplying these rates by the absolute time differences. As a consequence the variance of the branch length scales quadratically with the time difference. Alternatively the “white noise” model is totally uncorrelated at all times (i.e. within branches) and has the interesting property that the variance of the branch length scales only linearly with the time difference (Lepage et al. 2007).»

> L125. “accurate dating” relies most importantly on the calibrations, more so than the level of sequence information. In viruses, the sampling times can be much more informative as calibrations than the fossil calibrations in analyses of vertebrates. So even though viruses usually have very small genomes, molecular dating can sometimes produce precise estimates of virus divergence times.

I removed the mention of viruses and added “as well as calibrations”.

> RESULTS

> L161. The authors should discuss the impacts of forcing the gene tree topologies to match the species tree topology. I think this is a rather big problem in the study because the species tree has a few polytomies (Fig 1 and L433). Forcing the gene trees onto the species tree topology can distort branch lengths (Mendes and Hahn 2016; Carruthers et al. 2022). The authors hint at this problem on L235 “incongruence being masked by the reconciliation step in our dataset”.

We added a discussion of this. However we think that polytomies are on the contrary a good way to lift topology constraints on difficult nodes, because the gene tree inference will have the choice to select a better fitting bifurcating topology at these nodes. In simulations though, the gene trees are truly polytomic. One thing that we would like to nuance is that we did not “force” topology beyond what the reconciliation algorithm computes. The trees we selected do not show any duplication *node* (it’s not just that they have exactly one homolog in each species, instead it’s the reconciled tree that matches the species tree).

> Fig 1. Gene trees are likely to differ in topology as well as branch lengths, because of differences in lineage sorting and coalescence times. I am concerned about the effects of rescaling all of the gene trees to the same height. The authors should comment on the limitations of this approach.

Regarding the height rescaling, it is a way to mask gene-specific average rates (“gene effects”). I am not sure I see particular problems with doing this, except that genes with lower average rate might show a greater dispersion after rescaling, but this is a parameter investigated in the subsequent part (regression). I am adding the following text in the part 1 of the Results:

«Because the gene trees were independently subjected to dating, and the dated output trees were all scaled to 43.2 My, their differences in mean rate (ie the “gene effect”, [Ho et al. 2014]) are not considered in this part. The impact of genes specific features is investigated in the subsequent part.»

To complement this I added the following sentence in the analysis of the gene specific features (regression results):

«also because of rescaling all trees to the same height of 43.2 My, trees with the lower mean rate are expected to display a higher dispersion.»

> L170. Branch rates can be inferred even when there is a single calibration, as is the case in the present study.

We meant “insufficient” in practice, with regard to the accuracy of variable rates. We modified with “is unlikely to lead to accurate variable rates”.

> L186. I am not comfortable with the claim that “we can expect that the average of gene ages should fall on the correct value”. We actually expect different genes to have different node times, although they should be constrained by the species divergence events (assuming no subsequent gene flow).

We meant that heterotachy should not cause a bias, but indeed using “the correct value” was maybe naïve. We rephrased in a hopefully better way:

«However, these gene and site heterotachies should just cause dispersion but not loss of accuracy, as we expect that the distortions on branch lengths should compensate themselves on average.»

The next sentences are intended to discuss that point, we hope it is sufficient as it is:

«Another simplification of our inference is to consider instantaneous segregation of genes at speciations. In reality, speciation takes up to several million years and segregated genes can correspond to older allelic divergence. This phenomenon of deep coalescence should cause speciations to appear older than they are, and not younger as is the case in our results. [...]»

> L225. What is meant by “internal calibrations” here? The analyses here included a calibration at the root node, which is sufficient for estimating the branch rates.

We meant calibrating internal nodes in addition to the root. We rephrased as “only one calibration”, which we think does not allow to infer variability in rates (except as influenced by the tree prior).

> L255. The standard deviation of root-to-tip path lengths is used as an approximate and imperfect measure of among-lineage rate variation. It is much better to measure among-lineage rate variation using the branch rates themselves, because root-to-tip paths are mutually non-independent.

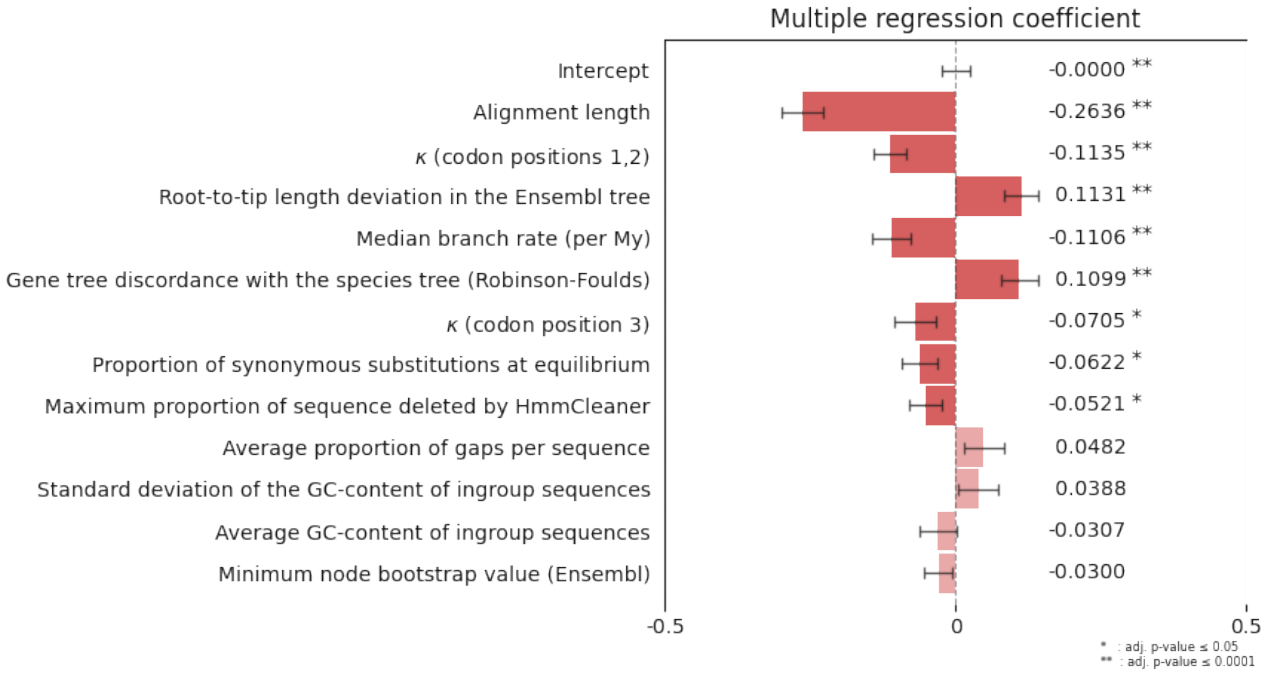
We agree.

> It does not make sense to include both of these measures in regression (also see my next point below).

We agree with this, however our safety checks for collinearity did not show them as strongly correlated, at least not in a way that destabilizes the Lasso regression.

> L259. Discussing the standard deviation of root-to-tip path lengths, the authors find “its predictive power is limited”. But this is only because they included a better measure (the standard deviation of branch rates) of the same property of the data. If they had omitted the standard deviation of branch rates, the standard deviation of root-to-tip path lengths would become one of the top predictors of precision.

We ran a separate regression with only the root-to-tip estimate and added the figure in the supplementary (S3). It shows up in third rank (figure below). This is good but not as good as the finer estimate. We updated the corresponding sentence in the manuscript: «we can expect a better predictive power by using a more refined measure of rate heterogeneity instead; we also performed the regression with only the root-to-tip variance as the sole measure of rate heterogeneity, and it has a lower coefficient than the more refined measure (supp. info. S3)».



DISCUSSION

> L340. I am not sure that “non-neutral substitutions” are “usually more clock-like in absolute time”. I think the authors mean “nearly neutral substitutions”.

My bad, I have indeed misread/approximated the relevant sentence in Ho et al. 2014. I am thankful for the patience of the reviewers and I hope that the following rephrasing is accurate:

«[…] our dating method employs a nucleotide substitution model which cannot distinguish neutral from non-neutral substitutions. According to the nearly neutral theory the majority of amino-acid changes is slightly deleterious which causes molecular divergence patterns to be more clock-like in absolute time, whereas strictly neutral substitutions should show a generation time effect (Ho 2014)»

> L373. The present study does not investigate the impact of calibrations, so it is unclear how the “study also reinforces the notion that fossil calibrations are quantitatively more important to accurate dating than sequence data alone”.

We have replaced the sentence with: «However fossil calibrations are quantitatively as important as sequence data alone to accurate dating (dos Reis et al 2015)»

> L391. Change “Extremely large gene families” to “Extremely large datasets”?

Done.

REFERENCES CITED IN THIS REVIEW

•Carruthers et al. (2022) The implications of incongruence between gene tree and species tree topologies for divergence time estimation. Syst Biol 71, 1124-1146.

•Mendes and Hahn (2016) Gene tree discordance causes apparent substitution rate variation. Syst Biol 65, 711-721.

Both of these references are now included.

Review by David Duchêne, 30 Jul 2024 16:46

The authors have made a substantial effort in addressing the reviewers' comments. One remaining point of confusion is the use of the term "precision", which seems to be mixed up with "accuracy" at times. I suggest that the authors replace their usage altogether for the actual definition of the terms, so "the distance from the true value" for accuracy, and the "width of the confidence/credible interval" in the case of precision. After addressing this point the article will be a nice contribution to the field of molecular dating with genomic-scale data.

Thank you for this second round of review. We have screened all uses of “precision” and “accuracy” words in order to replace them where appropriate. We have more rigourously used the word “dispersion” (or “consistency”) when referring to average deviation from the median, and “deviation” when referring to the deviation from the median of a single estimate. We have restricted the use of “accuracy” to the context of the simulations or in theoretical discussion. However we kept some instances of “accuracy” and “precision” in the abstract and introduction to make it more synthetic. As it is probably more convenient this way, we would like to redirect the reviewer to our file with track changes for an exhaustive listing of updates.

Review

My sincere apology for the late response as I have been spending too much time on my new born. I thank much the efforts of the authors in revising the manuscript. While my view will probably not affect the editor’s/recommender’s decision I feel that further comments may be **helpful for improving the study**, so I hereby provide two more.

We completely understand and are very grateful for the time you managed to spent reviewing the manuscript one more time.

2. L89: I could be wrong but I don’t think heterotachy is related to across-site difference. It in my memory specifically refers to the heterogeneity among branches.

There is an across-site aspect, but our phrasing is maybe too simplified, so we extended the definition of heterotachy to make it as unambiguous as possible: «At the scale of a single sequence, the heterogeneity of the rate across branches does not necessarily follow the same pattern between sites, e.g. different sites accelerate or decelerate in an independent manner. This is called heterotachy [...] »

This is incorrect. As I suggested before, heterotachy is not about across-site difference. It would be good to look at the wiki where it's indicated in the first sentence "Heterotachy refers to variations in lineage-specific evolutionary rates over time", and the cited literature there.

Sorry. We inserted the modified definition of heterotachy: «In particular, when considering a single site in an alignment, the across branch rate variation is called heterotachy (literally "different rates"; Philippe et al. 2003).». We kept some of the original text that discusses combinations of site and lineage variations but do not call it heterotachy anymore.

Also, the authors have carefully revised the ms regarding the weakness of using single genes in dating. I however still encourage the authors to strengthen and to expand related discussion a bit more, which would make the results and conclusion more stringent, in my view. Places where such discussion might be about could be "*To start with, calibrating only one node is insufficient, but this is precisely the purpose of our analysis, since we study gene trees for which nodes lack calibrations*". Certainly, it's the authors' liberty to take it or not.

We have expanded another part, in the discussion with arguments as to why we would want to study individual gene features. The new block is:

«Beyond identifying outliers, it would be interesting to understand why they are so, in terms of function, selection pressure or genomic context. In this high-throughput era, homologous sequences from a wide variety of taxa are accessible. A few genes have received considerable attention because of very distinctive evolutionary dynamics, such as PRDM9 which is generally evolving under positive selection and is a "speciation gene" in mammals (Oliver et al. 2009) or MHC immune genes that display elevated polymorphism in populations (Piertney et al. 2006). Since our evaluation used only one calibration, it is a worst case setting that could occur in gene trees with many duplications or transfers, events for which fossil calibration is less informative (but see Davín et al., 2018, who use horizontal transfers as relative time constraints). Key adaptations or transitions may result from these gene specific events, in particular duplications (Aguileta et al. 2006; Vosseberg et al. 2021), horizontal and endosymbiotic transfers of genes (Ochman et al. 2000; Koonin 2016), or movements of transposable elements (Boissinot et al. 2000; Ovchinnikov et al. 2002; Khan et al. 2006).»