Review response

Revision needed - Federico Hoffmann, 10 Dec 2023 20:10

The reviewers find value in the research presented but also list several valid points that should be addressed in a revised version. Reviewer 3, in particular, is very critical and raises issues about some omissions when referencing prior research on this subject and comments on the apparent lack of novelty of the work. A revised version would need to address these concerns. In addition, I think that the authors need to build a stronger case for why this study would fit in PCI Genomics. It seems that analyses based on a single gene would be more appropriate for other sections of PCI such as PCI EvolBiol or PCI Math and Comp Biol. Reviewer #1 has some good suggestions regarding this.

We are thankful to the recommender for its positive appreciation of our manuscript, and to the reviewers for their constructive comments. We have performed a major update of the article to address the technical concerns. In particular, we have sampled from the prior as suggested, and replaced the Calibrated Yule tree prior by the Birth-Death. This does not change the major conclusions of the article, except for one simulation parameter that produced overly young ages in the original manuscript and that is now showing a bias in different directions depending on the node. We have introduced a new variable suggested by reviewer David Duchêne, which is a Robinson-Foulds measure of the incongruence between an estimated gene tree and the species tree.
Given these updates, we have also rerun the regression, and we identify the same top 3 covariates followed by the new Robinson-Foulds variable. Consequently the original figures 2, 3 and 4c were updated (the 1 has been removed), as well as table 1 and 2. We only rerun simulations for the figure 4c (Simulated rate variability) because the other panels appeared unaffected by the impact of the prior. We specify it in the Methods section with:

“We then dated Primates speciations from these simulated sequences with Beast 2, as for the real Primates dataset above, except that the tree prior is “Calibrated Yule” for panels b and d of the fig 3. This tree prior was updated to Birth-Death in panel c because it was showing a bias towards younger ages.”

This also led to update sentences from the Results at L298 with the following:

“However, we find shifts for the highest rate heterogeneity between branches (4d, $\sigma/\mu = 1$). Cebidae appears younger while Catarrhini appears older than in reality, an effect that we find when sampling from the prior (supp. info. S6).This shows that in presence of very high across-branch rate variation and uninformative calibrations, the prior on the timed tree (Birth-Death) strongly influences the ages.”

and to entirely rewrite the Discussion section “Limiting the bias when dating single gene trees”.

In their major points, all reviewers have questioned the relation of this work with gene-specific events such as duplications and suggested to change the focus that we made in the introduction. We have changed the manuscript to address this, and we will give our general answer here. In the introduction, we have taken a more general view on the interest of disentangling gene-specific evolution. We therefore now refer to specific rates of evolution in addition to events such as duplication/transfer/ILS. Also, as suggested by reviewer 1, we now mention that molecular clock analyses historically relied on single genes. As suggested

by reviewer David Duchêne, it may be interesting to insist on approaches that filter genes, so we added the following in the Discussion after citations of the “gene-shopping” approach:
«Our results may be extended to a similar gene filtering approach as it pinpoints gene features that are correlated with the dating uncertainty, although its predictive power would need to be improved.»

Finally, we think that our analysis fits in PCI genomics because it analyses a genome-wide selection of genes and characterizes their features and their variation at this level.

In addition to reviewers requests, we have made the following minor change to the manuscript: the total number of trees under study is in fact 5205 and not 5204, because the number of HmmCleaner failures is in fact 30 and not 31.

Review by anonymous reviewer 1, 25 Oct 2023 05:52

However, the study has a number of major shortcomings. Although the study is partly motivated by the need to understand processes that specifically affect gene trees (as opposed to species trees), the stated premise of “dating single gene trees” has already been well visited in previous work. Molecular dating was carried out using single gene trees for several decades (until multilocus and genomic data sets became widely feasible), and is still commonly done using single-locus data sets such as organellar genomes. In fact, the analyses in the present study do not explicitly address the processes that differentiate gene trees from species trees (incomplete lineage sorting, gene duplication, horizontal transfer). Consequently, the study primarily investigates factors that have already been the subject of numerous studies and comprehensive evaluations. This past work needs to be taken into account and discussed in the present study. There is also a rich literature on incongruence between gene trees and species trees that should be discussed (e.g., Carruthers et al. 2022).

Please see our general response above.

ABSTRACT

L15. Does “time of appearance of genes” refer to gene duplications?

No, it refers to any gene branch: I replaced “genes” by “gene lineages”.

L19. This statement is somewhat confusing. Variability in rates is not generally addressed through concatenation, and the measures taken to model rate variation in multiple-gene data sets can also be applied to single-gene data sets.

We recognize that our phrasing is a bit vague: we meant that concatenation smoothes out intergene variability, and that fossil calibrations deal with interspecies variability. We changed the sentence to reflect this.

L28. It would be helpful to mention why the best precision is associated with core biological functions. For example, is it due to lower rate variation among branches?

The set of genes with best precision is derived from the regression fit. It is therefore mainly linked to three parameters, the rate variation (lower), the alignment length (larger) and the

average rate (higher, but it is only weakly associated). Based on your question, we have retrieved the enriched functions from the genes with longest alignment (supp info S4) and it is a superset of the functions we originally found, indicating that this is the main driver (we added a mention of this in the main text).

INTRODUCTION

L73. Change “laps” to “lapse” or “interval”. Done.

L85. Change “mechanisms at the origin” to “causes”. Done.

L101. The white noise model effectively models rates in a branch-wise manner, as with the uncorrelated models (but with variance being linear with time).

Thank you for the clarification. Our original phrasing may indeed cause confusion. After verifying in the article, we have modified the text to highlight the property that seems most important, and inserted: “has the interesting property that the variance of the rate is smaller on longer branches”.

L103. There is some uncertainty over whether rate autocorrelation can be detected (Ho et al. 2015; Tao et al. 2019).

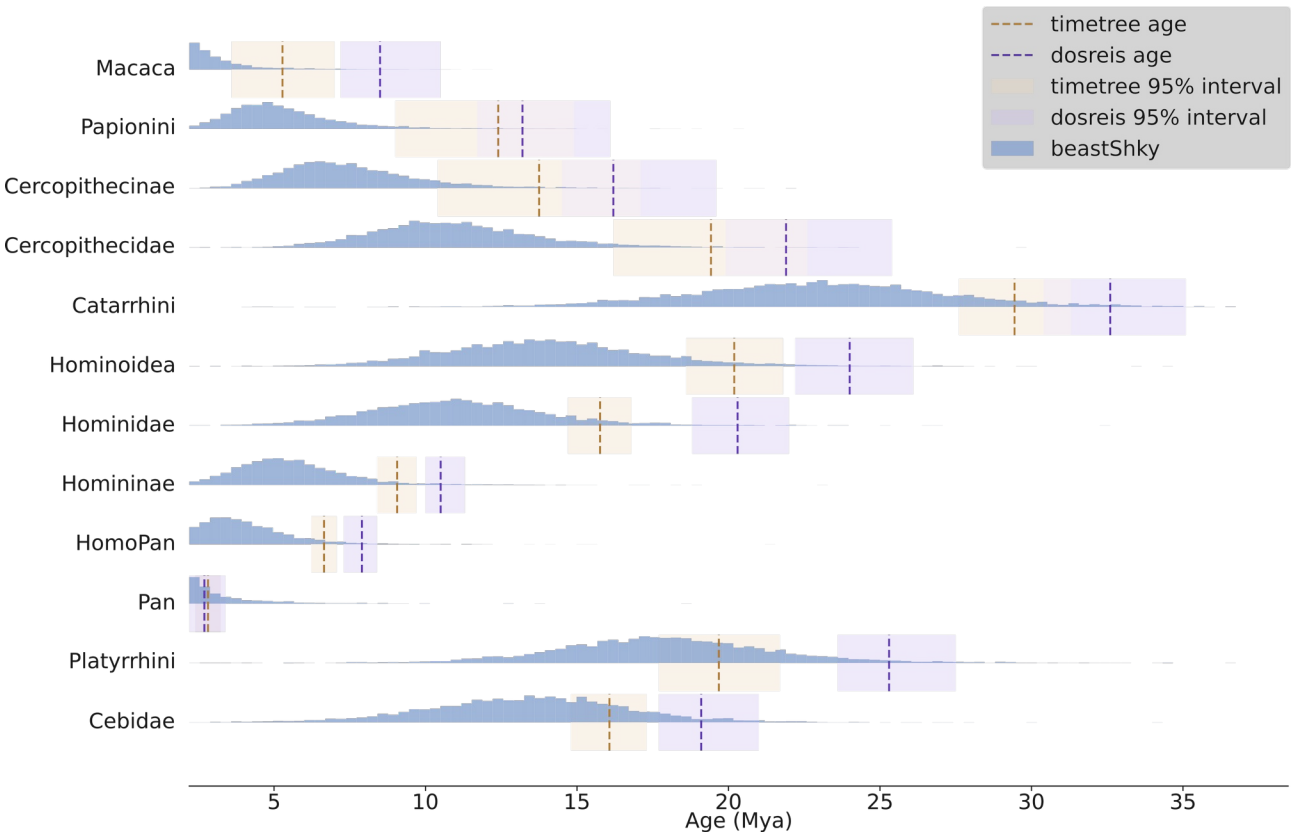
Remark inserted.

L114. Increasing the amount of information should lead to an increase in precision, but not necessarily accuracy. Correct, fixed.

RESULTS

L162. The age estimates from TimeTree are not necessarily reliable, given that they come from a wide range of sources, so they should not be used as a benchmark for accuracy.

It is useful to point out this limitation. It is one of the reason we perform simulations in the following. Also, we originally included a comparison with the study of dos Reis et al. (2018): *Using Phylogenomic Data to Explore the Effects of Relaxed Clocks and Calibration Strategies on Divergence Time Estimation: Primates as a Test Case*, which obtained significantly older ages than TimeTree (see figure below). However to clarify the message (as suggested by a previous round of reviewers), we only retained one comparison.



L169. Even when there is among-lineage rate variation, using a single calibration can be sufficient in some cases, although it is better to use multiple calibrations (Duchene et al. 2014).

Thanks for the reference. We gather that in the majority of situations, it is better to use more calibrations, so we did not change the text here. We inserted this reference in the appropriate part of the Discussion.

L200. The term “heterotachy” is normally used to refer to changes in site-specific or region-specific rates across the tree, not to among-lineage rate variation alone (Lopez et al. 2002). Please replace with a different term or phrase, to avoid potential confusion.

We have replaced “heterotachy” by “branch rate variation”.

L230. This can be confirmed using tests of saturation.

Sorry, we have decided to skip this extra analysis by lack of time.

L278. Would it be better to combine “alignment length” and “mean rate of substitution” into a single factor “number of variable sites”?

In the multiple regression, both variables are taken into account in a combined manner already, so we do not think merging them would further improve the regression results.

L298. The direction and size of any shifts would probably depend on the positions of the calibrations in the tree.

This result (original fig. 4) is based on simulated sequences on top of a fixed species tree, where we have not modeled heterogeneous species evolutionary rates. We used a single calibration at the root. I am not sure that additional calibrations would impact much this result, apart from reducing the shift.

DISCUSSION

L303. This section seems unnecessary; it mostly repeats parts of the Introduction.

We agree and moved some elements of this paragraph into the introduction.

L324. This can be evaluated using tests of model adequacy (substitution model adequacy and clock model adequacy).

This would be interesting, but since it would require running one or two extra MCMC of a more complex model, we cannot undertake this work in the timeframe of this review.

L339. I am not sure that this is the case. The nearly neutral theory was partly inspired by evidence that noncoding DNA showed a generation-time effect (causing rate variation among lineages) while coding sequences appeared to be clocklike over absolute time. Generally we expect a generation-time effect in the evolutionary rates of neutrally evolving DNA.

We corrected our sentence to reflect this: "[…] non-neutral substitutions, the latter being usually more clock-like in absolute time, whereas neutral substitutions tend to show a generation time effect (Ho 2014)."

L353. Independent rate variation among gene trees can be addressed using multiple clock models (dos Reis et al. 2014; Snir 2014; Duchene et al. 2016).

Since the updated simulation does not reproduce this bias, we have withdrawn this factor as a possible cause of bias, and removed the sentence.

L374. But genome-scale data sets seem to provide an ideal opportunity to discard any loci that have evolved too slowly/quickly or that show too much rate variation among branches (Klopfstein et al. 2017; Vankan et al. 2022).

We agree with the pragmatic necessity of discarding loci, as proposed with the "gene-shopping" approach, but our point here was to say that this might not be appropriate when a general picture is wanted. We extended the sentence after "unsatisfactory for genome scale analysis" with "that look for the most general picture".

L380. It would be worth noting that molecular dating can be performed on some large data sets using approximate likelihood calculation in MCMCtree (dos Reis and Yang 2011).

We have chosen to stay quite general regarding the details of the programs here, so we don't think that adding this information would change the main message of the paragraph which is not intended to sound pessimistic. If this is the case we will be happy to

reformulate. We incorporate this reference in the introduction instead, after the description of types of softwares.

METHODS

The Methods section seems to be a collection of points and needs to be reorganised and reformatted.

We have tried to improve the connections between Method sections in the revised manuscript. We have reordered some of them: the “Simulating alignments” section has been moved after the “Regression” section, to follow the presentation of the Results. The subsection “Retrieving mean rate and rate heterogeneity” of “Regression” has also been moved before presenting all the other features in “Collecting features of the gene families”.

L438. How were the three rates selected?

We had a sentence later in the text saying that we used the values from the real gene trees, that represent well the possible range. We moved the sentence before listing each value, and added again the reference to Supp. info. S5 which shows the real distributions.

L475. It would be better to select a few factors judiciously and focus on those that are most likely to have an impact on molecular dating. Also, many of the 71 characteristics overlap or essentially reflect the same features of the data.

We understand the critic, but we wanted to be as agnostic as possible regarding the influencing factors, so we adopted this “data mining” approach. We took great care in removing autocorrelated variables in the early steps of the analysis, through manual inspection of PCA results, computation of “VIFs” (variance inflation factors for each variable when regressed), step-wise removal of the variables that increase the most the “multicollinearity condition number” until it becomes less than 20, and finally applying the Lasso regression that only selects a limited number of features. We explain this procedure in the subsubsection “Reducing multicollinearity”.

L503. It is not clear how this mean and variance are different from the mean and standard deviation of the uncorrelated log-normal clock model mentioned on L501.

We added the following explanation:
«The latter estimation differs in that it is not a parameter of the model, but a statistic that is computed at the end of each iteration on the proposed tree. We monitor both estimates because they yield quite different values, although being correlated (supp. info. S9)»
The mean parameter in particular is apparently taking higher values than the a posteriori summary statistic.

L507. What is the purpose of computing the rate of substitutions per codon?

Regarding the results, it would not change anything to divide by 3 to show the average rate by nucleotide. I have added the following justification for our choice:

«This rate by codon equals three times the average rate by nucleotide, but we choose the codon metric for consistency with the simulation parameters based on a codon model in INDELible».

L509. This seems to be an unusual measure of rate heterogeneity. What information does this tell us beyond the metrics described in the paragraph on L500?

Since it may have been unclear, we inserted the following before L500:

"The clock model that we fit in Beast is unlinked between codon positions {1,2} and position {3}, meaning that the rate parameters are infered separately for each of these site partitions."

Consequently we had to design a measure that aggregates these site partitionned rate standard deviations and we did a weighted sum. The sentence was updated by replacing "obtained" by "summed", and not calling the total a standard deviation.

FIGURES

Figure 1b. For consistency with panels a and b, indicate that the tips are also 'calibrated' (assigned an age of zero).

Following the general concern of reviewers that the topic of gene duplications should be less important, we have removed this figure.

Figure 2. The tree should be oriented to face right, for consistency with the trees in Fig 1.
Figure 4a. The tree should be oriented to face right, for consistency with the trees in Fig 1.

After removing fig 1, all remaining figures now display the same orientation.

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.

dos Reis and Yang (2011) Approximately likelihood calculation on a phylogeny for Bayeian estimation of divergence times. Mol Biol Evol 28, 2161-2172.

dos Reis et al. (2014) The impact of the rate prior on Bayesian estimation of divergence times with multiple loci. Syst Biol 63, 555-565.

Duchene et al. (2014) The impact of calibration and clock-model choice on molecular estimates of divergence times. Mol Phylogenet Evol 78, 277-289.

Duchene et al. (2016) Estimating the number and assignment of clock models in analyses of multigene datasets. Bioinformatics 32, 1281-1285.

Ho et al. (2015) Simulating and detecting autocorrelation of molecular evolutionary rates among lineages. Mol Ecol Resour 15, 688-696.

Klopfstein et al. (2017) More on the best evolutionary rate for phylogenetic analysis. Syst Biol 66, 769-785.

Lopez et al. (2002) Heterotachy, an important process of protein evolution. Mol Biol Evol 19, 1-7

Snir (2014) On the number of genomic pacemakers: a geometric approach. Algorithms Mol Biol 9, 26.

Tao et al. (2019) A machine learning method for detecting autocorrelation of evolutionary rates in large phylogenies. Mol Biol Evol 36, 811-824.

Vankan et al. (2022) Evolutionary rate variation among lineages in gene trees has a negative impact on species-tree inference. Syst Biol 71, 490-500.

Review by David Duchêne, 19 Nov 2023 09:18

This article explores how the signal for molecular dating varies across individual genes, and tests how some of the features of these genes might lead to biased inferences of excessive uncertainty. The article will be a useful piece for future dating studies using genome-scale data.

The term 'gene duplication' is being used to refer to lineage divergence; however, the term means the appearance of a new gene copy within a genome, in the form of a paralog. Since the authors do not consider paralogs or duplications within genomes, I suggest the authors replace the term throughout the manuscript with 'divergence event' or similar.

One important factor that might drive gene-specific variation from the dated tree is the distance between the gene tree topology signal and the species tree topology. This distance might reflect incomplete lineage sorting or a limited signal in the data (e.g., the combination of a short gene with low rates). I suggest the authors consider making gene tree inference and adding this distance to their regression. Even fast inference would be sufficient, straightforward to implement, and will likely reveal a critical factor in driving dating error.

The introduction and text suggests that researchers are interested in the age inferences from single genes. Instead, the authors should consider focusing on the possible gene filtering for molecular dating, or on approaches to further scrutinise genome-scale inferences.

As mentionned in our general response above, we have introduced the robinson-foulds distance between the species tree and a reestimated gene tree using IQtree with model GTR+G+FO. It indeed appears to be an important predictor of inaccuracy, so we have discussed it in the results at “The fourth largest association [...]”.

Minor comments.

Abstract. Consider replacing the 'estimation of time' with 'inference of divergence times'. Done

Abstract. Consider removing the second sentence since the study does not explore gene duplication, and other factors such as mutation and genomic rearrangement also play a role (far more than only duplications and horizontal transmissions). Done

Abstract. The term 'speciation dating' is unconventional, consider revising. 'Such solutions' is a vague term and no real solutions have been mentioned.

We rephrased the whole sentence from:

“While speciation dating can cope with this variability by concatenating multiple genes and using fossil calibrations, such solutions cannot be applied to date gene-specific events.”

to:

“When dating speciations, per-lineage rate variability can be informed by fossil calibrations, and gene-specific rates can be either averaged out or modeled by concatenating multiple genes. However, when dating gene-specific events, fossil calibrations only inform about speciation nodes, and concatenation does not apply to divergences other than speciations.”

Abstract. Consider revising the emphasis on 'relaxed log-normal clock dating' since there are many other factors that can be as important or more in the model (substitution model, tree prior, calibrations, MCMC sampling settings, among others). We have partially rephrased, by saying that the simulations were done under a relaxed clock, and by simply stating “Divergence times were estimated with the bayesian program Beast2”. However, among the factors you mention, the main one that we have tested in the paper is really the across branch rate heterogeneity.

Lines 42-56. This paragraph seems to refer to divergence events rather than gene duplications within a genome. Gene duplications are not really relevant to this study.

We removed text starting at “In particular, gene duplications”, and rewritten the introduction to address this general concern (see our general answer).

Lines 83-94. There is a important gap of literature here. Consider citing the following literature (and related articles):

- Gillespie, J. H. (1991). The causes of molecular evolution (Vol. 2). Oxford University Press, USA.
- https://doi.org/10.1016/j.tree.2014.07.004 (Ho 2014, The changing face …)
- https://doi.org/10.1093/sysbio/syu020 (dos Reis 2014, The impact of rate prior …)
- https://doi.org/10.1038/nrg.2015.8 (dos Reis 2015, Bayesian mol. clock. … genomic era)

We now cite Gillespie 1991 in this paragraph. The subsequent paragraph adds some more details and already cites dos Reis 2015. We added a citation of Ho 2014 in the Discussion.

Line 98. Relaxing rate constancy cannot be settled since rate patterns will vary across taxonomic groups, timescale studied, genes sampled, calibrations used, among other factors. Remark added.

Lines 112-122. Two topics are noticeably missing from this paragraph. One is the scale-dependance of concatenation versus coalescent methods, where the error that each method addresses varies across data sets (https://hal.science/hal-02535651). The other is the impact of the substitution model on molecular branch lengths and divergence times (https://doi.org/10.1080/10635150500354647).

In our understanding, the “concatenation question” has been mostly studied with regard to topology inference, which is why we did not mention it. As we do not test the impact of the substitution model, we left the topic aside, and we are of the opinion that it is well described in the literature elsewhere and is not necessary here.

Line 134. Does higher precision mean narrower uncertainty intervals? Clarify.

It is the intent, but not with uncertainty intervals. Higher precision here means a smaller deviation from the median age. We now have inserted “(as measured by the deviation from the median age)”.

Figure 1. If the events of divergence do not lead to paralogs (two copies in one genome), then the authors are not referring to gene duplication, but rather divergence.

We removed this figure.

Lines 162-188. Consider emphasising in this paragraph the fact that calibrations are also about informing rate variation. Issues in inference can arise from unaccounted for variation in rates (missing calibrations).

It is true and we feel like we discuss this problem at length starting with the sentence “To start with, calibrating only one node is insufficient”. We have added 2 extra sentences:

«In fact, molecular clock “dating” is as much about estimating rates using calibrations than estimating dates from rates, this shift in focus being due to the high variability of molecular rates. Without many calibrations, variation in rates cannot be faithfully infered, and in turn dates remain uncertain.»

Lines 169-172. Consider citing Gillespie or Ho (mentioned above), and referring to these forms of variation adequately (gene effects, lineage effects, etc.).

We inserted the “lineage effects” and “residual effects” and cited Ho 2014.

Lines 298-300. The younger ages are likely also driven by the root calibration, which is a constrain on all ages that is often not available. Consider mentioning or even testing this.

We are not sure to understand how it can have an impact, because the same root constraint is applied to all simulated trees, so it is just the scale factor that should not impact the relative proportions inside the tree.

Line 421. The use of a Yule process is known to impact node age estimates (e.g., https://doi.org/10.1093/sysbio/syw095). Consider mentioning the use of birth-death, or testing its usage.

Thank you for directing us to this article. I gather from their conclusions that they only detected a problem with intra-species data. However, the other reviewers also expressed concerns with the Yule prior, mentionning the problems of the original implementation in Beast v1. In fact, we had used the updated implementation named “Calibrated Yule Prior” (Heled & Drummond 2012) which is not suffering from the issue of inconsistency between the input age distributions and the ones sampled by the joint prior. However, we had improperly specified (uniform) rate priors that caused problems specifically with the Calibrated Yule prior. We therefore rerun analyses with the Birth-Death prior instead and ensured proper priors.

Line 429. Were ESS values actually verified to be above 200? Consider checking whether convergence (ESS) is associated with uncertainty in estimates.

ESS values were computed with 'loganalyser'. After updating our pipeline to the latest Beast parameters, we had unfortunately omitted to check on all of the ESS but we have now done so, and for all parameters. In the original "Calibrated Yule" run of the 5204 real Simiiformes trees, 319 had an ESS below 200. In the revised manuscript we used the Birth-Death computations, leading to 390 trees with at least one ESS<200 after 20 millions iterations. We extended those once with another 20 millions iterations (this is now described in the Methods). After this, 13 trees have one ESS<200, and we removed them from the regression fit.

Review by Sishuo Wang, 06 Nov 2023 09:56

The authors conducted an interesting study in exploring the issues affecting dating single gene trees using molecular clock approaches. They appear to have a good understanding of molecular clock literature. The perspective is very novel and makes much sense. Strictly speaking, there are many issues to address some of which are even not covered in my following comments. However, to encourage open science and the new academic publishing way using PCI, I would suggest giving the authors an opportunity to revise.

In addition, reviewing this technically interesting but complex ms requires much expertise much of which may be beyond my knowledge. So my apologizes in advance for any misunderstanding and please correct me if that happens.

Major points:

1. In Fig. 1, the authors mentioned the ultimate goal, but this seems to me to make the present ms a bit confusing as duplication is not mentioned and dating gene trees with duplications also involve other issues for example the accelerated rate due to relaxed purifying selection, loss of duplicates in some lineages, and many others. In the current analysis, the authors removed all genes that showed duplications, which is clearly indicated. So, it seems a bit strange that Fig. 1 is presented in this way. Instead, I was wondering if it's better to move anything related to duplication to the end or SI.

   We have now addressed this point generally (see our general response).

2. The use of TimeTree to set calibrations should be very careful. See below.

   a. L156: The authors set the calibration at the specified node based on the recorded dates in TimeTree database but there might be some to address. Looking at the corresponding method description in L419-L431), the authors chose to fit the 95% CI and the point estimation using a gamma distribution. However, as some suggest, the times in TimeTree should be interpreted very carefully and they simply do not like the idea of using that as a reference. Hence, if the authors are going to take this advice, they might want to choose one or a few alternative calibrations for analysis.

We acknowledge this problem, but we think that our benchmark should be interpreted in a relative way, where the root calibration only serves to set the scale, which is the same for all replicates. This is why we then express dispersion (MAD and IQR95) as a percentage in paragraph 3 of the Results. To reflect this point of view we inserted:

«The choice of this root calibration and its associated uncertainty is arbitrary because all trees are then compared by this yardstick. Likewise, the choice of Primates for the source trees is arbitrary; the specific selection of species does not matter, what matters is that we collect natural replicates of the same tree.»

b. Also, setting the calibrations at other nodes or adding a few calibrations would also help. The reason is simple: only when people do not have good calibrations would setting a single calibration at the root makes much sense. So in this regard, the results are very interesting but the practical significance seems to me not very clear. This at least should be discussed.

We recognize that this is a major problem in the general case. We have also discussed this in the manuscript, in particular L166:

«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.»

as well as in the Discussion section "Limiting the bias when dating single gene trees", L345:

«In such circumstances the model with a relaxed clock requires calibrations on internal nodes to properly infer branch rates and times».

So we would like to defend this design where we keep nodes uncalibrated because it is by analogy with gene trees that could display duplications or other non-speciation events (this was the argument of the original figure 1). We therefore extended the above discussion with:

«Here, the choice of a single calibration was made precisely with the aim to measure the dating accuracy on uncalibrated nodes, such as those occuring in gene trees if we consider events other than speciations».

c. Speaking as a researcher not believing the above point so much, even if the calibration used by the authors is widely used by many people, it is almost always suggested to run the same analysis using alternative calibrations.

We will provide here the same justification as for point a.

d. Further, the detailed parameters of the Gamma distribution used as the calibration should be made clearer. Is it the following in L428? If so that corresponds to Gamma(4.6,0.656) but I see that α and β are the same for the two calibrated nodes, aren't they? Also, the specified Gamma distribution above has a mean at 4.6/0.656 which is apparently smaller than 70.8 and 40.9. Am I misunderstanding anything?

Thanks for spotting the imprecise parameter description, indeed it should be corrected to follow the accepted convention that the scale parameter is represented by θ. For our defense, the Beast and Beauti programs name their parameters 'alpha' and 'beta' even though they use the 'ShapeScale' mode by default (we knew we were using this mode and our scale parameter is appropriate). We have corrected the text to be fully explicit. Additionally, we replaced our word "location" by the parameter name "offset" as used by Beast. This gives the following mean ages for Primates: $70.8 + 4.6 \times 0.656 = 73.8$ My, corresponding to the TimeTree age.

Also please excuse us for the copy-pasting error, the parameters for Simiiformes should be $\alpha = 4.0$ and scale $\theta = 0.575$. The resulting mean age is: $40.9 + 4.0 \times 0.575 = 43.2$ My.

e. Another way to rely on TimeTree is to fit all recorded dates of the node of interest into a gamma distribution instead of basing the analysis on only a mean value and the CI. This particular subpoint is simply a suggestion so the authors can well take the liberty to accommodate it or not and actually I see no need to accommodate it.

For the same reasons as above (point a), we do not estimate necessary to perform this analysis.

f. Perhaps more importantly, the authors gave me a "wrong" impression that the dates recorded in TimeTree were somehow taken as a "reference" although this is neither case for setting the calibration nor when comparing their obtained posterior dates with TimeTree. I think this is partly because the authors mentioned at the very beginning the use of TimeTree and throughout the manuscript. I suggest the authors change this way of writing which can easily mislead readers like me.

We do both comparisons, against the median age and against TimeTree. Indeed the TimeTree comparison is only done in the first section of the Results. After that we rely on the deviation from the median as our measure of precision. To insist on the use of the median we have inserted a new sentence in the introduction before mentioning TimeTree:

«Using the median age as a point of reference produces a measure of the precision of dating.»

3. L207: In Fig. 3, why was the intercept set to zero? I think it is the result of OLS i.e., the second round of your regression analysis, isn't it?

The intercept was not constrained, but it was expected to be estimated at zero, because all input variables were centered (and normalized). I added this information there and in the methods section.

Also for Fig. 3, are these 9 items the only ones that were significant?

There are only 9 (10 in the revised manuscript) variables because of the Lasso regression with regularization parameter set to 0.02 which performs variable selection (the least redundant set of variables). Significance levels are indicated by the asterisks next to the coefficient values.

Was there any multiple testing?

We originally did not apply multiple testing correction because it is not considered necessary in a multiple regression, which already accounts for all variables in a joint manner. However we double-checked the literature and it appears that we need to do so in the context of our full procedure because of the preliminary variable selection by Lasso. We updated the p-values by multiplying them by the number of initial variables being tested.

4. I think the use of simulations are worth mentioning in the abstract. This part also seems to me to make more sense since by that people know the "true" values of the parameters to estimate. So I suggest expanding the simulation part more.

   We expanded a little bit the abstract. Instead of "as well as simulated alignments", we now have "We also simulated alignments based on characteristics from Primates, under a relaxed clock model, to analyze the dating accuracy".

5. One important thing to explore for single gene dating might be the impact of the time priors on the time posteriors. Some suggest that single gene trees contain not much information so their dating results are not trustable. This seems to me to be important, but I also understand this may be beyond the scope of the study. Nevertheless, I was wondering if the authors can discuss this a bit.

   Thank you for highlighting this point, we have now run a sampling from the priors. This showed us that the simulated ages were in some cases biased by the time prior (it's not noticeable in the empirical Primate gene trees). We have however not tested many time priors, only the Calibrated Yule from the original article and the Birth-Death in the revised one. In the revision we have added the following in the Results on simulations:

   «This shows that in presence of very high across-branch rate variation and uninformative calibrations, the prior on the time tree (Birth-Death) strongly influences the ages.»

   and in the Discussion:

   «However in the simulated case with very high rate heterogeneity, the ages are biased by the time tree prior (Birth-Death model).»

6. The ms involves many methods. In general, I would suggest referring to Methods when they are mentioned in Results. For example, L281: the authors should mention "see Methods", and the same applies to elsewhere. This would greatly help readers understand the work.

Minor points:

1. L74: per site?

We chose to not mention the number of sites explicitly, because the mathematical formulation holds for any number of sites.

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 [...] »

3. L100: I suggest citing the corresponding papers of the two models here, although I see the papers I would cite are cited later in the para.

We assume you mean Thorne et al. 1998 and Drummond et al. 2006. We added them here.

4. L122: The authors mention the following "the most critical limitations are the difficulty of characterizing the type of clock relaxation and the uncertainty in calibration points themselves (dos Reis et al. 2015)". So, I was wondering how they were dealt with in the present ms, particularly the choice of the clock model (as I see the authors chose to use an independent rate model assuming the rate across lineages *i.i.d.* ~ log-normal). This could be a bit challenging but if not possible to test at least the authors should discuss this limitation.

Regarding the uncertainty in the calibrations, we think that it is beyond the scope of this article because we are interested in relative dates, not especially absolute dates.

We did not explore much the type of clock relaxation, because the log-normal already appeared as the most flexible. We had however run one set of simulations (not shown in the manuscript) by generating trees with the geometric brownian rate relaxation (autocorrelated), and then dating with the independent log-normal. We do not show it because the two models are difficult to compare due to the input "diffusion" parameter not producing the same output standard deviation of the rate.

5. L164-L165: the statement is not wrong, but because the authors used a single calibration on some 5000 single gene trees, I would tend to believe those estimated by other studies. So in my view this sentence is not very meaningful to be mentioned here.

We agree that the error likely comes from our study design, and in fact we would like to keep this discussion, because it is precisely an important message of the study in our opinion: that the lack of calibrations in a single gene tree causes errors. We have added "less plausibly" to the hypothesis "that the reference ages themselves are inaccurate".

6. L203: any reference?

We are not sure for which part you are requesting a reference. The "absolute deviation from the median" metric, or its application for the dating imprecision? Such application is entirely a design of our own.

7. L215-L219: my perhaps not accurate understanding is that even for uncalibrated nodes, they have time priors derived from a birth-death process (in the present ms I think the authors mentioned it's a Yule process), and also there are effective priors caused by the truncation effects both of which could lead to time priors not as flat as a uniform distribution (see Barba-Montoya et al. 2017 MPE). That said, given proper priors, I am unsure if the non-identifiability mentioned by the authors holds. Again, I am unsure.

Thank you, we were indeed incorrect about identifiability (given proper priors). We also noticed it when sampling from the prior, as suggested by the reviewers. In the revised manuscript, we replaced the Calibrated Yule by the Birth-Death prior and we still observe that high rate variation mislead the dating inference, despite proper tree and rate priors. We now favor the idea that the time priors have large impact on the output that supersedes sequence data (in fact reviewed by dos Reis et al. 2015).

So although not strictly speaking a non-identifiability problem, we are dealing with insufficient information from the sequence combined with a tree prior that is too inaccurate. To reflect this, we have replaced "not completely identifiable" with "in practice difficult to infer" and added this sentence about the branch rate parameter: "it would take values dictated by the tree prior (the Birth-Death branch process) that cannot possibly recover the exact Simiiformes speciation dates [dos Reis and Yang 2013]."

8. L223: I think you may want to cite https://link.springer.com/article/10.1007/BF00160154 (Yang 1994b)

We would like to keep the Results section as clear as possible, so we prefered not to add another bibliographic reference here.

9. L415-L417: the authors used the codon alignments if I did not misunderstand anything. It should be made clearer here if so.

Because HmmCleaner does not consider codons, we applied it on amino-acid, and then backtranslated the filtered alignments. In the original publication as well (Di Franco et al. 2019), HmmCleaner performance was assessed on AA alignments.

10. L421: What's the single parameter for the Yule process to specify the priors for uncalibrated nodes?

In fact we used the "Calibrated Yule Model" prior (Heled & Drummond 2012), parameterized by a birth rate (estimated), lower limit 1.0e-6, upper limit 10. In the revised manuscript, the Birth-Death tree prior was used instead, with estimated birth and date parameters. These birth and death parameters have hyperpriors following a Gamma distribution with shape and scale (0.001, 1000). We provide the template file in the data archive so we were initially trying to stay concise about the parameters. Since we now also show the sampling from the prior in the Supp. info. S6, we would argue that specifying these hyperpriors might not be essential.

11. L428: what software did the authors use for assessing the ESS?

ESS were computed with 'loganalyser' from Beast. There was a preburnin of 5% (i.e. did not save the first 1,000,000 out of 20,000,000 iterations), saving one sample every 2000, and then we removed 1% burnin (of 10000 saved samples).

12. L452: Isn't $\kappa = 4$ a bit large?

We chose this value based on inference from the Primates gene trees, initially with Codeml and then with the Beast HKY model used in this manuscript. Both show a median close to 4 We inserted the histograms based on the Beast inference in the supplementary information S13.

One study that supports this value in Mammals is Rosenberg et al. 2003 (https://doi.org/10.1093/molbev/msg113) that used the fourfold degenerate codon sites.

13. L452-L453: If I interpreted it correctly, the authors did not involve an across-rate heterogeneity in the analysis by setting the α parameter. Is that correct? This is fine as this part of analysis is not the main one, but the authors should clearly indicate this.

This is a good observation, sorry for not being explicit here. Modeling a Gamma distributed across-sites rates variation in INDELible with a relaxed clock is a bit more involved as it would have required partitioning the sequence into different rate categories, rescaling the relaxed clock tree for each partition, and then providing these branch lengths to INDELible. We added "No across-site rate heterogeneity was modeled".

14. L509: do you mean "across-branch rate heterogeneity"? If so, I'd also suggest showing the equation that calculates v. I also do not quite understand the equation in L511. I wondered if the authors mind showing how it is derived.

We now use the explicit term "across-branch rate heterogeneity". Reviewer 1 also asked about this formula. It is in fact an arbitrary proxy, and not the exact standard deviation for all codon sites, as was incorrectly implied by our original phrasing. We have only access to $v_{1,2}$ and v_3 separately, and no access to the covariance between both so we cannot compute the variance $v_{1,2,3}$. We now only indicate that we *summed* both standard deviations $\sqrt{v_{1,2}}$ and $\sqrt{v_3}$, and again, this is a proxy.

15. L516: grammar mistake

We deleted the extraneous " a ".

16. The figures are duplicated at the end. Already inserted when they are first mentioned. This is not an error and could be due to the requirement of bioRxiv in uploading the files so actually no need to address.

We removed the duplicated figures.