Peer Community In Genomics

Dating single gene trees in the age of phylogenomics

Federico Hoffmann *based on peer reviews by* ***Sishuo Wang***, ***David Duchêne*** *and 1 anonymous reviewer*

Guillaume Louvel and Hugues Roest Crollius (2024) Factors influencing the accuracy and precision in dating single gene trees. bioRxiv, ver. 6, peer-reviewed and recommended by Peer Community in Genomics. https://doi.org/10.1101/2020.08.24.264671

Submitted: 16 August 2023, Recommended: 28 November 2024

Cite this recommendation as:
Hoffmann, F. (2024) Dating single gene trees in the age of phylogenomics. *Peer Community in Genomics*, 100292. 10.24072/pci.genomics.100292

Published: 28 November 2024

Dating evolutionary trees is a critical task that allows us to connect biological history to ecological and geological events, helping us explore connections between environmental change and genetic innovations. The central idea behind these techniques is to link changes at the sequence level to divergence times, under the general assumption that substitutions accumulate steadily over time. So, sequences that diverged earlier are expected to be more different than sequences that diverged more recently. For a number of biological and statistical reasons, the relationship between sequence divergence and time is not linear, so it is not always the case that more divergent sequences have accumulated more substitutions than less divergent ones. In the case of organismal-level divergences, a natural approach to mitigate these challenges is to incorporate as many genes as possible into the analyses. However, this route is not available when we are focusing our interest on a single gene or a gene family. Thus, exploring how different features of single gene trees impact the accuracy and precision of divergence time estimates is of interest. In this study, Louvel and Roest Crollius (2024), select a well-studied group of mammals, primates, extract single copy genes from their genomes, and explore how different factors such as alignment size, evolutionary rate variation and discordance between the gene and species trees impact divergence time estimates.

There are many strengths of this study. The central ones are the number of factors considered and the transparent discussion of the limitations. In this regard, the study is an elegant combination of empirical and simulated data. Some of the results match intuitive expectations. For example, the authors find that longer alignments are more informative than shorter ones, that differences in evolutionary rate among branches lead to loss in precision, and that slow-evolving genes perform worse. Intriguingly, they also find differences in performance among genes with different ontologies. The empirical data used in this study is limited to a single group, and generally considers genes that have apparently remained as single copies. Accordingly, the

conclusions that can be drawn are somewhat limited, calling for future studies building on and expanding the concepts of the study by Louvel and colleagues. For example, including genes that have been lost or duplicated would be of interest because changes in gene complement are a prevalent source of variation at the genome level in mammals in general (Demuth et al. 2006), and particularly in primates (Hahn et al. 2007).

References:

Demuth JP, De Bie T, Stajich JE, Cristianini N, Hahn MW (2006) The evolution of mammalian gene families. PLoS One, e85. https://doi.org/10.1371/journal.pone.0000085

Hahn MW, Demuth JP, Han SG (2007) Accelerated rate of gene gain and loss in primates. Genetics, 177,1941-1949. https://doi.org/10.1534/genetics.107.080077

Louvel, G and Roest Crollius, H (2024) Factors influencing the accuracy and precision in dating single gene trees. bioRxiv, ver. 6 peer-reviewed and recommended by PCI Genomics. https://doi.org/10.1101/2020.08.24.264671

Reviews

Evaluation round #2

DOI or URL of the preprint: https://doi.org/10.1101/2020.08.24.264671
Version of the preprint: 4

Authors' reply, 29 October 2024

Dear recommender,
please find attached our reply to reviewers comments. Apologies for the delayed answer.
We also attach the tracked changes between the two rounds of review (the font I used can be downloaded from here: https://practicaltypography.com/charter.html).
Best regards,
Guillaume Louvel
Download author's reply
Download tracked changes file

Decision by Federico Hoffmann, posted 04 October 2024, validated 04 October 2024

Two reviewers have evaluated your revised manuscript and agree that you have addressed most concerns. Nonetheless, one reviewer in particular has identified a number of regions that warrant further discussion and/or clarification. In addition, they point out that there is a wider literature that should be cited and discussed in your manuscript.

Please let us know if you have any questions regarding the reviewers' comments. We look forward to receiving your revised article.

Sincerely,

The PCI Genomics Managing Board

On behalf of Federico Hoffmann

————

Note made on October 12, 8 days after initial decision.

In addition to the official decision made on October 4, the Managing Board has received minor comments from the third reviewer by who was unable to make the original deadline, but we believe sharing their comments will help the authors revise their manuscript. Those minor comments are attached as Word document to this message. Please let us know if you have any questions about this process.

Sincerely,
The PCI Genomics Managing Board

Download recommender's annotations

Reviewed by anonymous reviewer 1, 31 July 2024

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).

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.

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.

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).

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.

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).

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.

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”.

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.

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

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).

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.

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

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.

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”.

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”.

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

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.

Reviewed by David Duchêne, 30 July 2024

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.

Evaluation round #1

DOI or URL of the preprint: https://doi.org/10.1101/2020.08.24.264671
Version of the preprint: 2

Authors' reply, 06 June 2024

Download author's reply

Decision by Federico Hoffmann, posted 10 December 2023, validated 18 December 2023

Revision needed

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.

Reviewed by Sishuo Wang, 06 November 2023

Download the review

Reviewed by anonymous reviewer 1, 25 October 2023

This study examines the performance of Bayesian molecular dating using single-gene data sets. Dating analyses are carried out on 5204 sequence alignments from primates and on simulated data sets. Based on these analyses, the authors find that the most important factors affecting date estimates from single-gene data sets are sequence length, rate variation among branches, and evolutionary rate. The study is interesting because it considers a wide range of factors that potentially affect date estimates, and the analyses are framed in the context of studying gene duplications and other processes that have specific effects on gene trees.

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).

ABSTRACT

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

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.

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?

INTRODUCTION

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

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

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).

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

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

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.

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).

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.

L230. This can be confirmed using tests of saturation.

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

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

DISCUSSION

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

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

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.

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).

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).

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).

METHODS

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

L438. How were the three rates selected?

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.

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.

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

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?

FIGURES

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

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.

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.

Reviewed by David Duchêne, 19 November 2023

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 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.

Minor comments.

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

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).

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

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).

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.

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
- https://doi.org/10.1093/sysbio/syu020
- https://doi.org/10.1038/nrg.2015.8

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.

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).

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

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.

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).

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

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.

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.

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