Response to reviews

We would like to thank the editor and the two reviewers for their comments as they have enabled us to improve the manuscript. We respond below to specific points that they raised and hope that the changes we made will be found to be satisfactory.

Bastien Boussau

Round #1

Decision

by Cécile Ané, 2020-12-22 11:52
Manuscript: https://doi.org/10.1101/2020.10.17.343889

In this preprint, Szollosi et al. present an method to date a tree using relative age constraint, such as implied by horizontal gene transfer events, and a two-step approach to ease the computational burden. The usefulness and the ease of using the method are exciting.

Both reviewers are positive. The first review is high-level. The second review made excellent suggestions. In particular, one concern is that the simulated trees had modest rate variation and are close to being ultrametric. Looking at the materials on github, one simulated tree looks far from ultrametric to me. The authors could clarify, compare with non-ultrametricity in real trees, and perhaps consider the addition of simulations in which rate transformations are more drastic.

RESPONSE*: The simulated tree indeed was not ultrametric. We include now a figure S1 in the supplementary material to clarify how the species tree was simulated and altered to introduce deviations from the clock. We also measured the amount of non-ultrametricity and compared it to trees from the Hogenom database, which resulted in fig. S2 and an additional section in the supplementary material. We concluded from these analyses that the amount of non-ultrametricity was realistic.*

Reviewer 2 made valuable comments about some results interpretation, such as the marked improvement from 4 to 5 constraints, and the value added by proximal vs distal constraints. I very much agree. About distal constraints: I find the authors' conclusion that distal constraints are more informative than proximal constraints counterintuitive. Intuitively, a distal constraint

corresponds to a proximal constraint after information loss. For example, a proximal constraint implies distal constraints between the "older" node and any descendant of the "younger" node. As another example, the donor and recipient of a HGT need to have the same age (proximal event), but would provide a distal constraint due to extinction or a lack of speciation events (or lack of sampling) along the lineages "around" the HGT. Like reviewer 2, I invite the authors for more discussion, and I wonder if some other factor is at play in the authors' simulation.

RESPONSE: *We thank the reviewer and the editor for pointing this out as indeed this section required more work. We have added a section in the supplementary material to discuss theoretically what can make a constraint informative. We also added more simulations, by shuffling the order of the constraints. We computed several statistics on the constraints and used a linear model to better understand what makes a constraint informative.*

Another comment is about the sparse documentation of the empirical data analysis. I concur: documentation needs to be greatly expanded, to help understanding and increase reproducibility. For example, the lack of documentation made it hard to understand some information and annotations in figures 6 & 7 (e.g.: is the "95% HPD for Viridiplantae" in fig. 7 based on the authors' analysis, or from some other source?).

RESPONSE: *We have improved the documentation of the empirical data analyses, and included trees for the Archaea data set, in Fig. S8.*

I attached technical / minor comments and suggestions from my own reading.

Technical / minor comments and suggestions

Formal description. This section should be expanded, starting with a clarification of notations (example: what are Ca and Co mathematically? Is Co the indicator function that age_node1 $\leq$age_node2 and node1 $\in$ T and node2 $\in$ T, or a product of such terms for various choices of two nodes?). In the second equation at the bottom of page 4, should the terms be switched to show the prior P(Co|T) instead of P(T|Co)? In the sentence that follows, does the indicator function δ(T,Co) also require that the node of interest be displayed by the tree T; and should this definition be expanded to cases when “Co” includes multiple relative constraints?

RESPONSE: *We have rewritten the description of the model and believe it is now much improved.*

Implementation. “two additional functions”: additional compared to what? “Scripts are available”: will users need to write low-level scripts? It looks like a tutorial instead. The wording could be improved in this section, to better describe the authors’ contribution.

RESPONSE: *We have changed the wording here. RevBayes does require one to write scripts.*

Two-step inference. An equation or a more explicit description of the composite likelihood should be included. As stated, this composite likelihood is a factor of two terms: (Gaussian distribution) × (posterior means and variances). Surely this is incorrect! Also, are branch lengths assumed to be independent, with 0 covariances in the Gaussian distribution used to build the composite likelihood? “per-branch” suggests independence, but should be clarified.

RESPONSE: *We have rewritten this section.*

The simulations need more details: how many replicates were run (how many trees, and alignments per tree)? It looks like there was only 1. Why not do several, if only 10? Could the marked improvement from 4 to 5 constraints result from idiosyncratic to the one simulated data, in which a "large" rate change was spanned by constraint number 5?

RESPONSE: *We performed more replicates of our simulation in which we introduced the 15 constraints in random orders. This enabled us to study what makes a constraint informative.*

The rates of small and large changes are said to be 33 and 1, but which process was used: Poisson? under the original or rescaled branch lengths?

RESPONSE: *It was indeed a Poisson process, and we have made this explicit in the manuscript. Fig. S1 should also clarify our simulation protocol.*

I could not find the description of the simulations underlying Figures S1-S4.

RESPONSE: *We have now added description of these simulations in the Material and Methods.*

Archaea: could a figure be added to show the 62-taxon tree, the relative-age constraints, and visualize the hypotheses being tested?

RESPONSE: *We added Fig S8 with this information.*

Figures could be improved, like their legend ("n" and "y" is not explicit), and the x axis label "0" in fig. 5.

RESPONSE: *We have changed the figures and believe they have improved.*

Legends for figures 3-5 could recall which constraints were used (relative? calibrated? both?).

RESPONSE: *This information is now included in the legends.*

The legend for figure 2 says that constraints are numbered according to the order in which they were used. Is that applicable to figure 5?

RESPONSE: *Since we used 10 random orders in the revised manuscript, we no longer focus on one particular order.*

Could the relative constraint(s) be visualized in figure 6?

RESPONSE: *We cannot visualize the constraints on the figure, as there are too many of them: 431 for Archaea, and 144 for Cyanobacteria. The constraints are provided on Dryad.*

The legend for Fig. S1-S3 should explain why there are 3 violins for each condition (How many taxa, calibrated and relative constraints were there?).

RESPONSE: *We now provide information in the Material and Methods section and have improved the legends. The three violins corresponded to triplicates.*

##########

##########

Reviews

Reviewed by David Duchêne, 2020-12-04 23:11

In their manuscript, Szollosi et al. report an implementation and detailed exploration of a new approach of time-calibration based on relative node times. The approach is intuitive, and to my knowledge has not been described or tested in previous research. The description is very clear and the explorations using simulations and empirical data are thorough. In particular I commend the authors for exploring various widths of calibration and for using such realistic simulation schemes. The method is valuable and I believe that any comments on the methods or manuscript would be a matter of personal preference, rather than academic rigour. For these reasons I wish to recommend this piece in its present form.

RESPONSE*: We thank the reviewer for his comments.*

##########

##########

Reviewed by anonymous reviewer, 2020-12-14 20:44

In this work, the authors present a follow-up to their 2018 paper "Gene transfers can date the tree of life" in Nature Ecology & Evolution. This work expands upon previous efforts, demonstrating how the relative age constraints imposed by horizontal gene transfer between lineages on a tree are compatible with absolute age estimates constrained by the fossil record, for a test dataset of Cyanobacteria and Alphaproteobacteria. The authors also show that "true" relative age constraints between nodes improves the precision of age estimates for subsets of HGT types. This implementation is also added to RevBayes, so that HGT relative age constraints can be combined with other molecular dating approaches, as made available by the authors.

The authors also take a novel approach in estimating uncertainty in branch length estimations, which underpin estimates of evolutionary rates in ultrametric trees. Essentially, they generate a distribution of branch lengths from a population of trees generated from a MCMC analysis. These are then used to estimate rate parameter variances across branches. This appears to be an

efficient approach as an alternative to more computationally costly methods typically employed, as stated by the authors.

This work constitutes an incremental improvement in our understanding of the utility of HGT in refining divergence time estimates. The statistical analyses and phylogenetic/bioinformatics methods appear to be sound and appropriately used. The major limitation of this work, both practically and conceptually, is the starting point of published time-trees in the case of both simulated and empirical investigations; in short, it is not surprising that many HGT constraints, be they simulated as "true" or extracted from actual HGT events, will be redundant with respect to the observed relative ages within the calibrated tree; it is also not surprising that some of the HGT constraints will improve precision of age estimates if these happen to be "active" with respect to branches of high rate variance. That being said, this proof of concept is clearly validated here, which is important. The work would be far more impactful if application of simulated or empirical HGTs to rooted phylogenies independently converged on similar relative age estimates to those that were obtained by calibrated molecular clocks alone.
RESPONSE: *We clarified the wording in several places of the manuscript, and changed Fig. 1 to better show the dating information that constraints can provide thanks to our new method. We want to insist that this manuscript presents a new method, which allows dating a phylogenetic tree using relative order constraints, possibly many of them, in a statistically rigorous way, which was not possible previously. We validated our approach using both simulated and empirical data, which we believe is state of the art, and had not been done previously.*

Major comments:
(1) The authors test their approach on simulated trees, reconstructed from sequences simulated under the published timetree of Betts et al. As such, their simulated trees are already going to be very close to ultrametric, even after their shuffling approach (this is acknowledged, as the authors state the tree height is the same in both cases). That is, if the relatives ages of groups proposed by Betts et al are generally reasonable (being informed by fossils and evolutionary rate models), it is not surprising that the resulting simulated trees will also be largely compatible with a set of "true" transfers.The rate transformations applied as part of the simulation are extremely modest, with the vast majority of branches only having a 10% variance in rate, with very few branches having a 20% variance. It is unclear if these rate variances were selected to fit those observed within empirical datasets (e.g., Betts et al.). If so, then this is reasonable. However, if not, then the conservative choice of rate variances seems to favor their intended result: compatibility between valid HGT constraints and timetree estimates. Regardless, since this assumption is so important to their conclusions, the authors should perform additional tests, increasing the branch rate variances and observing at what level of branch rate heterogeneity the congruence between relative HGT constraints and the molecular clock model is "broken". In the introduction, the authors make clear that the utility of

this approach is driven by the absence of reliable absolute age constraints within most microbial lineages; therefore, it would be far more convincing if the simulated studies were performed on phylograms rather than chronograms, that more reasonably reflected the information we generally know about microbial groups (in the absence of rate models or fossil constraints). Ideally, the authors would apply their simulations to both the chronogram-derived trees (as they currently do), and the rooted phylogram generated by Betts et al. without any inferred rates, to show that the sets of simulated HGTs reasonably interact with trees that do not contain any dating information within their branches.

RESPONSE: *We believe there has been a misunderstanding. We clarified our protocol in the text, and added Fig. S1 to show how our simulations are done. We further quantified the amount of non-ultrametricity (=deviation from the clock) in our simulated tree, and compared it to empirical phylogenies, and found that rate heterogeneity in our simulations is realistic.*

(2) In discussing the results shown in Figure 3, the authors state that " Results improve markedly with 5 or more constraints, with a strong effect when moving from 4 to 5 constraints, and then a slower improvement. There is no obvious feature of constraint 5 that would make it substantially more helpful than other constraints for dating." It appears that this "step" effect is largely determined by the order in which HGT constraints were applied. If relative age constraints have a low probability of being "active" as the relative ages of nodes they constrain are already well-resolved by the timetree, then it is expected that stepwise addition will make little change to the error in node ages, until one highly active constraint is applied. This will produce a step down in error (as is observed). Following this step, application of additional HGT constraints will generally not further increase precision, if these are consistent with the first "active" constraint encountered (in the case of the authors approach, constraint #5). In fact, HGT #5 is the first constraint to be imposed thats pans the full diversity of the tree, away from the root. This in itself may be a meaningful observation. The authors should shuffle the order in which HGTs are applied in this stepwise approach, or jackknife using different subsets of HGT constraints (much like is often done with fossil calibration) in order to show that increasing numbers of HGT constraints generally increases precision, and the result is not being driven only by one or a handful of particularly active constraints.

RESPONSE: *We have added more computations, applying constraints in random orders. This additional experiment allowed us to study what makes constraints informative. We have also added a section in the supplementary material to discuss theoretically what can make a constraint informative.*

(3) The results of Figure 5 are not particularly surprising from a traditional divergence time estimate perspective; for proximal age constraints, the uncertainty in the age estimates of each node will be more likely to overlap, and thus be consistent with the presence or absence of the constraint; only distal HGT constraints have the potential to "violate" the rate models strong enough to result in narrowing the age distributions on the constrained nodes. This seems trivial, and it is unclear why the authors have presented this analysis, as it does not further their case for the other hypotheses they test.

RESPONSE: *We no longer focus only on proximal vs distal and instead consider several statistics of the constraints that can affect their informativeness.*

(4) Analysis of Empirical data: The authors applied HGT-based age constraints to cyanobacteria and archaea in order to show that these constraints improved the precision of age estimates independently of the fossil calibrations. However, the list of applied HGTs, what groups they constrained, and how they were obtained is lacking from the manuscript or SI (or, at least in a form that is not recognizable). These constraints are a critical part of their test and should be very clearly documented. If it is the case that these HGT constraints are similarly "simulated" , this should be made much more clear in the methodology.
RESPONSE: *The constraints were not simulated but come from Davin et al., 2019. There were 431 different constraints for Archaea, and 144 for Cyanobacteria. We have clarified this in the Material and Methods section. Files containing the constraints are available on Dryad.*

(5) Discussion:
" We further found that constraints between nodes of similar ages were less useful than constraints between nodes of differing ages. This is encouraging since it should be easier to find transfers between nodes whose ages differ widely (distal transfers) than between nodes with similar ages, because large age differences give more time for transfers to occur and to leave a detectable footprint in extant genomes." This encouragement seems to be misplaced. While it is certainly true that HGTs with donor-recipient ages that vary widely are easier to detect and likely more abundant, these would seem to consist largely of HGTs where the recipient is clearly much younger than the donor clade (e.g., "forward in time" transfers, such as one genus of bacteria being the HGT recipient from within a different order of bacteria), and thus, provide little or no additional information, as branch lengths alone will clearly resolve the recipient as being much younger than the donor. Rather, it seems the most active HGTs will be ones where the relative age distributions for the donor and recipient nodes are highly uncertain, and the HGT constrains the relative ages in a way that conflicts with the rate information—that is, pushes the recipient to be younger than the donor, when the evolutionary rate models would be most consistent with the inverse. Thus, "active" HGTs in the way described by the authors may actually be quite rare. A more nuanced discussion of this point should be provided.
RESPONSE: *We have changed the discussion and no longer make this claim.*

Minor comments:
(1) the manuscript contains many grammatical errors and appears to have been hastily drafted. It should be carefully copy-edited.
RESPONSE: *All co-authors have re-read this new version.*

(2) Introduction: The authors provide a useful discussion of previous work using HGT to constrain the relative ages of groups in a tree: " Recently it has been shown that gene transfers could help date species trees, because they contain information on the chronological order of speciation nodes(Szöllosi et al. 2012; Davín etal. 2018)." They should also include references to Wolfe & Fournier 2018 and Magnabosco et al., 2018, which also show the application of the same concept with similar methodologies.

RESPONSE: *We have now included the missing reference and discuss the differences in the methodologies.*

(3) Figure 1 is of low quality and should be re-drafted.
RESPONSE: *It has been replaced.*

(4) Betts et al. should not be cited for the ete3 library references.
RESPONSE: *This has been fixed.*

(5) In the "Constraints improve dating accuracy" section, the Maximum A Posteriori tree is incorrectly cited as Figure 2A.
RESPONSE: *This has been fixed.*