Response to decision letter

2023-05-01

Dear *Editor-in-Chief* Isabel Sanmartín,

Please find below response to each point raised by the *Associate Editor* Daniele Silvestro and the two *Reviewers*. We greatly appreciate your time and consideration for our resubmission.

Kind regards,

Luna L. Sanchez Reyes, on behalf of all coauthors.

> 19-Oct-2022
>
> Dear Dr Sanchez Reyes:
>
> Decision on USYB-2022-152, DateLife: leveraging databases and analytical tools to reveal the dated Tree of Life:
>
> Accept pending minor revisions.
>
> Thank you for your Systematic Biology submission. It has been reviewed by Associate Editor Dr Daniele Silvestro and two reviewers with the relevant expertise. Their comments are listed at the end of this letter. Both reviewers and the AE provide some excellent constructive suggestions that I am sure you will appreciate. The general consensus is that this software is of general interest to and expected to be widely used by the systematics and evolutionary biology community, providing capabilities that will help in the construction of supertrees. I agree with this assessment.
>
> I think that your paper will be a valuable contribution to Systematic Biology once the detailed comments provided below are addressed, especially in relation to data curation and the English style. Thank you very much for your submission.
>
> Sincerely,
>
> Dr Isabel Sanmartín Editor-in-Chief, Systematic Biology isanmartin@rjb.csic.es

Thank you for your decision! We made sure to address all comments carefully and hopefully you find the manuscript improved.

> Please acknowledge (by email to sysbio.editorialoffice@oup.com) receipt of the reviews and give a probable time frame for the return of your revised manuscript. Your revision should be submitted as soon as possible. Any revision not received in a timely manner may have to be considered a new submission.
>
> Your revision must comply with our instructions to authors (in this letter, with additional instructions at https://academic.oup.com/sysbio/pages/General_Instructions). This applies even if your original submission deviated from Systematic Biology style and corrections were not included in the notes from the editors or reviewers. Therefore, READ AND APPLY OUR INSTRUCTIONS BEFORE SUBMITTING A REVISION. Failure to do so will result in significantly delayed

processing of your paper. If anything remains unclear after reading the instructions, use recent issues of the journal to find examples. If necessary, feel free to contact the Managing Editor at sysbio.editorialoffice@oup.com for help.

Noted and checked.

Log into https://mc.manuscriptcentral.com/systbiol and enter your Author Center, where you will find your manuscript title listed under "Manuscripts with Decisions." Under "Actions" click on "Create a Revision." Your manuscript number will be appended to denote a revision. Your original files are available to you. Delete the old files and replace them with your modified versions. Files that did not require revision can simply be retained

You will be unable to make your revisions in ScholarOne Manuscripts. Instead, revise your paper using a word processing program and save it on your computer. HIGHLIGHT the changes to your manuscript within the document using the track changes mode in MS Word or by using bold, strikethrough, or colored text; any method you choose which indicates all changes made. If a large number of changes were made and you feel the document is too cluttered to read easily with all of them shown, feel free to submit a "clean" copy as well, in which changes are not indicated. If you decide a clean copy would likely be helpful to reviewers and editors, please upload it in the category called Related Files.

Noted and checked.

When submitting your revised manuscript, address each point made by the Editor, AE and Reviewers IN THE SPACE INDICATED (under "Response to Decision Letter"). Your revision cannot be processed if your responses to reviews are given only in a cover letter. The best way to address each point would be to copy this letter and insert your comments after each point made. Please do not change the order of or delete any of the comments because this makes it difficult to review again and would slow the process. The format of your responses must be compatible with ScholarOne Manuscripts text fields. For example, colored text is not an option, but you could use asterisks (and numbers, spacing, etc.) to clearly distinguish your responses from the text of the reviews.

Feel free to argue your case, with careful consideration and documentation, if you disagree with any of the suggestions. If you feel a reviewer did not understand a point you made, in your response keep in mind that as an author it is your responsibility to make your points clear to the readers.

Noted.

DO NOT submit your revision in .pdf format. This applies to the main text, tables, and appendices. Individual figures may be in .pdf format.

In the specific case of papers written using LaTeX, in addition to the .tex (and associated style, bib, or etc. files) please include a .pdf generated from those files. Upload the .pdf in the ScholarOne Manuscripts category called Related Files.

Figures must be uploaded separately, with each file name including the figure number. No figures may be imbedded in or tied to the .pdf or .tex files.

After you've uploaded your revision, carefully view the ScholarOne Manuscripts version to verify that all figures and other files display correctly, and that you've followed all of our author instructions.

Figures: Each figure must be submitted as an individual file. The name of the file should include the figure number. Numbers and letters on figures should be as large and clear as possible, without overlapping any text. No figures may be imbedded in the document. Captions should not be placed on the figures. Figure portions should be referred to in the text and figure captions using lowercase letters. On the figures themselves, portions should be labeled by a lowercase letter followed by a single parenthesis (e.g., "a)"), located in the upper left area of the figure portion.

Line thickness (including graph axes) should be a minimum stroke weight of 0.5pts, and 1.0pts is recommended for most lines. We prefer vector rather than bitmap figure formats. If desired, see http://systbio.org/?q=node/138 for an explanation of the difference. If bitmapped figures are necessary, they should be created at a minimum of 300 dpi and should be about 8 inches wide. We accept a wide variety of figure file types, as long as the figures are of sufficiently high resolution.

Noted and checked.

The cost of printed color figures is $600 each. Authors are normally expected to cover this cost, but we do have limited funds available for authors who are SSB members, and cannot pay the full amount. If you feel your circumstances create an unusual financial need, please explain in your cover letter. The maximum allowance is one color figure per paper. All color costs can be avoided if you decide to have the figures printed in black and white, but be shown in color in the online version of the paper. If you do this, the captions must be worded such that they are appropriate for both situations (e.g., descriptions should not name colors), because the captions in print and online will be identical even if color is used online only. Do not submit two versions of any figure. Instead, make sure the color figure is also easily legible in grayscale, then submit only the color version. In your cover letter, please make your intentions regarding the use of color clear.

Noted.

Journal Covers: Please consider submitting a suggested cover image. They can be illustrations of theory, photos of organisms, a combination of the two, or alternatives. These can be uploaded with a revision of the manuscript or may be sent later. Only images for which you have copyright permission or that are not under copyright may be submitted. The color figure fee does not apply to images chosen by the Editor to be on the cover of the issue. If you have questions about possible cover designs, please email sysbio.editorialoffice@oup.com. We encourage and greatly appreciate cover image suggestions.

Data and Online Appendices: All such files should be deposited in the Dryad data repository. Also, before the manuscript can be published, data accession numbers must be in the text. Nucleotide sequence data and alignments must be submitted to GenBank (http://www.ncbi.nlm.nih.gov/genbank) or EMBL (http://www.ebi.ac.uk); alignment, input file and tree files must be submitted to TreeBASE (http://www.treebase.org); morphological data must be submitted to either Morphbank or MorphoBank. Check over your online appendices or supplemental material (if any) carefully, because they will not be copyedited or proofread, and cannot be changed later. The first time you mention online-only material in the text of your paper, give the doi provided by Dryad as the location where the material can be found. Then, in a separate section after the main text, include a statement such as:
SUPPLEMENTARY MATERIAL. Supplementary material, including data files and/or online-only appendices, can be found in the Dryad data repository (put your Dryad doi in parentheses here).

We added DOIs for Dryad package with Supplementary Material, L551.

Make sure all section headings conform to Systematic Biology style for first, second and third levels. Use of incorrect styles is potentially confusing and, in any case, is likely to delay processing of the manuscript.

Noted and checked.

Tables should have a single-sentence informative title above the table, with any other descriptive information located below the table in the form of notes and/or specific footnotes.

Noted and checked.

When references are grouped together in parentheses in the text, they should be listed in ascending chronological order. Multiple references in a single year should be alphabetized.

Noted and checked.

All funding used for this work should be listed in a "Funding" section preceding the Acknowledgements. Please give the full official name of each funding body.

Done on L556-558.

Our accepted abbreviation for millions of years ago is Ma. The abbreviation for millions of years duration is myr.

Checked.

Our full instructions are accessible via the link in the upper right area of the ScholarOne Manuscripts page (the link goes to https://academic.oup.com/sysbio/pages/General_Instructions). In case of a discrepancy between the information at that site and this letter, follow the instructions in this letter.

Noted.

If you are the first author and your manuscript is based on work done while you were a student, you would be eligible for the "Publisher's Award for Excellence in Systematic Research." When submitting your revised version be sure to indicate student work in the checkbox provided. If two student researchers were heavily involved in the manuscript, please briefly describe the situation in your cover letter.

This work was done while the first author was a Postdoctoral researcher.

Please note that the accepted version of your manuscript will be published on Advance Access prior to typesetting and copyediting; it is therefore important that all files, including tables, figures, and supplementary files, are properly labelled and with the editorial office at this stage of the process - it is the author's responsibility to ensure that the correct files are submitted. Please also keep in mind that files will be published as submitted (ie, please review table formatting and labelling and position of figure images, etc). For further information please see the Advance Access Publication section at: https://academic.oup.com/sysbio/pages/General_Instructions.

Noted and checked.

Associate Editor: Dr Daniele Silvestro

Recommendation #1: Accept with minor revisions

Comments to the Author: Dear Dr Luna L Sanchez Reyes and co-authors,

many thanks for submitting your manuscript to Systematic Biology. Your study was reviewed by two highly qualified reviewers who provided an overall positive assessment of the paper while listing a number of things that should be revised and clarified. Based on their assessment and my own reading of the manuscript I invite you to resubmit it after carefully revising to address each and all points raised by the reviewers.

Thanks for your reviews and comments! They were most helpful and we believe they greatly improved the ms. We carefully revised the manuscript to address all points raised by the reviewers and yourself, and we hope you find the revision suitable.

In addition to that, please also address the following points:

1. how and why is a parsimony method used to estimate branch lengths and how is this used in combination with a likelihood method (line 174)?

Thanks for this question. We elaborate on this on the Description section, subsection "Dating a tree with branch lengths" L221-228. Briefly, we implement the parsimony algorithm ACCTRAN using functions from the R package phangorn. We chose this algorithm as a quick way to obtain

initial branch lengths that can be optimized afterwards using ML. The DNA matrices generated from data mined from the Barcode of Life Database (BOLD) do not have too much variation, so ACCTRAN resolves ambiguous character optimization by assigning changes along branches of the tree as close to the root as possible (Agnarsson & Miller, 2008, doi:10.1111/j.1096-0031.2008.00229.x).Once parsimony branch lengths are estimated, they become parameters that are then optimized using Maximum Likelihood (ML), given the alignment, the topology and a simple JC model. The branches that are returned are ML branches. We added a vignette to the R package in which we showcase the whole workflow of branch length optimization, from parsimony to ML. This vignette is available at http://phylotastic.org/datelife/articles/making_bold_trees.html.

---

2. Why are the node ages evenly distributed between calibrations? I would expect an exponential distribution of node ages under a standard birth-death process.

Thanks for this questions. We elaborate on it on the "Description" section, at the end of subsection "Dating a tree topology with no branch lengths", L 202-210.Note that we expanded the subsection "Dating a tree topology" into three sections: "Applying secondary calibrations", "Dating a tree topology" and "Dating a tree with branch lengths".

We want to point out that the main goal of the summarizing step in DateLife (when we distribute ages evenly in the chronogram) is to provide a single chronogram that can quickly show at a glance, and in the most agnostic way possible, the distribution of node ages obtained from published studies. Using an exponential distribution for missing node ages would make additional assumptions about the underlying evolutionary model of the chronogram, which can bias results in downstream analyses, as we elaborate in the discussion section, "Effects of phylogenetic sampling on downstream analyses" (L445 of manuscript without differences; L527 of manuscript with differences).

In the DateLife package we have implemented a function called "make_mrbayes_tree", that runs the MrBayes algorithm using the option in which you can use a birth-death strict-clock model to sample branch lengths in the absence of genetic data. Inherited from MrBayes implementation, the function requires setting parameter values for birth and death rates. If the amount of missing genetic data is large, the birth and death rate values applied will determine the shape of the chronogram branch length distribution and bias the tree shape (Rabosky 2015, doi:10.1111/evo.12817).

We discuss the effects of using this type of chronograms with missing data that have been completed at random following a certain diversification model, especially when used for analyses that require estimating a diversification rate, which might introduce circularity (Rabosky, 2015. Evolution doi:10.1111/evo.12817).

This is why we chose an algorithm that distributes node ages evenly between calibrations and does not require any assumptions on the underlying model of branch length distribution. The Branch Length Adjuster (BLADJ) algorithm allows this and is the one set as default for the summarizing step. However, users that wish to use MrBayes instead con do so with the "make_mrbayes_tree" function.

---

4. I think the use of an arbitrary root age set by default as 10% older than the oldest age is unjustified and dangerous. If no root age is provided by the user, I think the function should return an interpretable error message and refuse to run.

Thanks for raising this point. We agree that this practice could be dangerous if a user is not aware that the age of the root was set in an arbitrary way. We changed the behaviour of the code to return a very conspicuous warning message when the root is arbitrary to make sure that users are aware of the fact (L196-201).

In line with our thoughts on the previous point, we want our program to be able to return a chronogram if there is at least one age data point available from the literature. Unfortunately, it is impossible to return a chronogram if there is no age for the root, so in the absence of that information, and to be able to automatize the algorithm, the only option available to us is to randomly chose an age that will allow to generate a chronogram using the real data that is available. We agree that choosing the 10% is probably too arbitrary, so the function now adds one standard deviation unit of the mean of ages available to the max age value, if there is more than one age data point, and uses that as the root age.

The warning message that we implemented for this version of datelife now suggests ways on how users can provide an age for the root that is informed in the literature. The main goal of the message is to make sure that users are aware that there is no age data available for that root in the datelife database.

Please make sure to carefully revise the text to remove typos.

We revised the text throughout to remove all typos we could detect.

As one of the reviewers pointed out, it is good to provide links to permanent repositories for your code. I see that DateLife actually is already hosted in a Zenodo repository, so maybe you can add the link to your Availability section to make it more visible.

Right! We added Zenodo links for our code and other materials used for this research in the Availability section (L505) and Supplementary Material (L518-520).

I hope you will be willing to revise and resubmit your paper and that you'll find these and the Reviewers' comments useful.

Best regards, Daniele Silvestro

Yes, thanks!

Reviewer(s)' comments to author:

Reviewer: 1

Comments to the Author The manuscript introduces DateLife, an R package and web service that provides time-calibrated phylogenies across a number of organismal groups. It integrates with a number of other services including the Open Tree of Life project. The authors provide some benchmarks and walk through a worked example of using their service.

I don't agree with the characterization that stochastic polytomy resolutions methods such as PASTIS (used in Jetz et al 2012) is "making up" these branch lengths (line 455). My understanding of the phrase "making up" implies that these are solely inventions of the researcher, rather than generated through the use of well-tested statistical models. The manuscript also claims that there are no thorough analyses of phylogenies generated in this way (lines 462-465), which is not the case, and I suggest the authors revise this section in light of some of the relevant literature in this area.

- Cusimano et al Syst Bio 2012
- Thomas et al MEE 2013
- Rabosky Evol 2015
- Chang et al Syst Bio 2019
- Title and Rabosky MEE 2019

- Sun et al AJB 2020

Thanks for pointing us to these references. We elaborated on this subject more thoroughly by adding results from these studies to the corresponding section on the Discussion, "Effects of phylogenetic sampling on downstream analyses" (L481 of manuscript without differences; L527 of manuscript with differences), as well as changing the wording from "making up" to "in the absence of genetic data, simulating following a birth-death process" (L489; L540).

> There were many typographical errors in the manuscript which should be corrected prior to publication.

We corrected typos across the text.

> Reviewer: 2
>
> Comments to the Author As I have checked the box that I don't need to remain anonymous, there is also no point in being mysterious about this: I have been aware of DateLife for a good long while because I've seen its earliest prototype develop at a workshop at NESCent ages ago. I've loved the idea ever since - combined with some healthy reservations that I am happy to share here.
>
> DL synthesizes results from previous research. On the one hand that's great, but on the other, it invites the 'garbage in - garbage out' problem. Although OTOL has its own curation and QC facilities, the fact that users can provide their own garbage trees makes it so that the service might end up decorating nonsensical data, tainting its own reputation in the process. It would be good if the authors could emphasize this a bit more.

This is a good point. We are aware of this issue, and we think that any biological software is subject to this potential problem. For example, if someone generates bad DNA sequences, no matter how good the aligner software they use is, they will end up with a really bad alignment and subsequently a very bad phylogeny.

We agree that users of all biological software should keep in mind that the results they get are only as good as the data they provide, and that they should implement some quality control. We mention this in the abstract and 3rd paragraph of discussion.

> A separate but related point that I would also like to see discussed is that synthesizing services such as DL and OTOL seem capable of ending up in loops where bad trees with bad calibration points provide the skeleton for further bad trees based on the former - with their own seemingly well-supported but in fact dodgy secondary calibrations. Is that a risk? What can be done about it?

Thanks for mentioning this. We think DateLife actually helps preventing this issue. As shown in Figure 5, DateLife allows to compare all age data available for a group, and users can immediately identify chronograms that are outliers, and explore whether it is related to an artifact of the chronogram or something else about the data or methodology used to generate that chronogram. Users can then choose to drop those calibrations from the final analysis.

We agree that it is a responsibility of the users to check the trees and chonograms before using them as secondary calibrations, in the same way researchers curate fossils to use as primary calibration points.

> Also related: will we gradually start developing a body of literature with trees where the root always just happens to be ±10% older than the oldest nodes? Might that be bad?

The Associate Editor also pointed this out, and we agree that this feature of DateLife is not ideal, so we corrected it and address it in the code and in the manuscript. First, adding the 10% is too arbitrary. If there is more than one age data point, the function now adds one standard deviation unit of the mean of published ages to the maximum age value, and uses that as the root age. Second, the function now returns a conspicuous warning message when the root is not based on published data to make sure that users are aware of the fact (L196-201).

We want to highlight that the main goal of DateLife is to provide a single chronogram that can quickly show at a glance the distribution of node ages based on published data, so we want our program to be able to return a chronogram if there is at least one age data point available from the literature. Unfortunately, as you pointed out, when the root age is absent, it is not possible to return any dated tree. In the absence of that information, and to be able to automatize the algorithm, the only option available to us is to randomly chose an age that will allow to generate a chronogram using the real data that is available.

The warning message that we implemented now suggests ways on how users can provide an age for the root that is informed in the literature. The main goal of the message is to make sure that users are aware that there is no age data available for that root in the datelife database and that the one used for the chronogram is arbitrary.

Apart from these general points that might be touched upon a bit more in the Discussion, here now some specifics about the manuscript:

- The Abstract looks like an extreme afterthought. I understand how that works, but please have another look. I see verb disagreement on line 21 and on line 23. Probably needs a comma after databases on line 25. On the same line, 'timeframe' is spelt as one word (fine by me), but elsewhere it's two words. Line 27: 'incetivizited' is not a thing. Line 29, 'finding' scans weird, maybe use 'discovery'? Line 36 probably needs 'use' instead of 'using' but the sentence is hard to parse. Line 38, 'awereness' is wrong. In this way, the Abstract is quite different from the rest of the MS, which is otherwise well written.

Thanks for your comment, the abstract was indeed "a bit" forsaken on the version of the ms that you first revised. We rewrote most of the abstract, checked for spelling, and added more detail about the findings.

- In the first paragraph of the Intro you might want to add something like 'comparative analysis' (Harvey & Pagel, yada yada yada). It's clearly something that's on your mind because in the Conclusions, 'trait evolution' is the first research area you mention as needing chronograms.

Thanks for pointing that out. We added that topic to the intro along with a couple of references, L48-49.

- On page 7, second paragraph, you state that subspecies are ignored. What do you mean precisely? My guess is that you ignore the subspecific epithet and collapse to species level. Maybe state that more clearly.

Thanks for pointing this out. We actually only ignore subspecies when retrieving data from a more inclusive taxonomic group. When provided by the user, subspecific taxa are processed and searched fully. We clarify this in the Description section "Creating a search query".

It would be possible to drop the subspecific epithet and perform a more general search, so, we added this possibility to our software development plan, so that it can be implemented in a future iteration of the software.

- On page 7, third paragraph: how does the TNRS deal with homonyms? Given that we are in the tree realm it should be possible to infer intelligently whether some label is zoological or botanical code. Or is Aotus simply always the monkey, which is much cooler than that Australian legume genus?

Yeah! this is a pretty cool quality of TNRS that we implement in datelife. It is possible to provide or identify the "biological context" of a toxonomic name, so that if a list of names has mainly monkeys, then it will pick *Aotus* the monkey, but if it has mainly plants, it will pick the Australian legume.

Moreover, homonyms in the Open Tree taxonomy specify the group they belong too, so users can easily identify if the numeric identifier they obtained from TNRS processing belongs to the group they wanted or not. For example *Aotus* the legume is referred to as "Aotus (genus in kingdom Archaeplastida)" and the monkey as "Aotus (genus in Opisthokonta)".

We added examples for the two points above in a new vignette for the package, available at http://phylotastic.org/datelife/articles/make_datelife_query.html

---

- On page 8, fourth paragraph, it's not quite clear whether DL's database syncs automatically with Phylesystem or whether you have volunteered yourself for this task. Which would be noble, but hard to sustain.

It is indeed currently synced "manually". We specify this in the Description, L125. We comment in the discussion the benefits of having syncs done automatically, and ways to do so.

- On page 10, second paragraph: mining BOLD and aligning the sequences automatically is very cool functionality but I did not see it exposed on the website at all. How can users get at those alignments? Also, might there be performance issues? MAFFT can be quite greedy with larger data sets.

Ah, yes, thanks for the comment! Well, the BOLD workflow is faster than a classic workflow, but it still takes considerable time, specifically computational time required for the database search, which the server can't afford yet. The functions are currently only available on DateLife's R package and not on the web application, and we think it would be good to make them available there in the future. Regarding possible performance issues, the alignments from BOLD data are not very long, so we expect both MUSCLE and MAFFT to perform approximately the same. We chose MUSCLE as the default aligner.

We comment on this on L231-232.

- On page 25 you mention the fossilcalibrations.org initiative. Maybe that's a good opportunity to go a bit into what we need as a community. I suspect that, in general, most people in this field think that doing it by themself is 'better', i.e. do a bunch of sequencing (hybseq right now, I guess?) and then get good primary calibration points. Natural history collections must have many more of those, both as fossils but also from geology (i.e. vicariant events having to do with tectonics, orogeny, etc.). Shouldn't we want *that*?

Indeed! The golden standard for obtaining divergence time estimates entails the use of primary calibrations (fossil or from geologic events). Fossil preservation bias makes it impossible to get calibrations to constrain all nodes in a phylogenetic tree, hence, we will always have nodes that we need to infer using a model. Moreover, uncertainty in placement of fossil calibrations and phylogenetic hypotheses makes it so that there are many different hypothesis supported by the golden standard that can be even conflicting. The goal of DateLife is to provide the user

with an insight on the current state of knowledge of time of lineage divergence for any group of taxa, as well as a tree summarizing the data.

We addressed that in the discussion.

- Page 26, line 416 has some typos.

Fixed :)

- On page 26 you discuss some criteria for scoring quality of chronograms. One additional criterion might be where the calibration points are placed. Nodes that have a calibration point between them and the root have less freedom of movement and hence narrower confidence intervals. Ages ago, I did a bit of simulation work on that (Vos & Mooers, 2004 - definitely no need to cite). Maybe someone else has discussed this a bit better?

Thanks for pointing us to that paper! It gave us some nice ideas for discussion. We added a paragraph on the Discussion.

- Page 27 line 448, chronogram should be plural, I think.

Fixed!

- Page 28 line 473: I think it should be either 'public-funded' or 'publicly funded'

Went for "publicly funded", thanks!

- Page 29, Supplementary Material: it's probably better to sync the repos with Zenodo and cite the DOI, just so that it's guaranteed unchanging.

Thanks for the suggestion! We created stable versions for all three repositories on Zenodo, and refer the doi in addition to the GitHub addresses, L522-525.