Dear Hilmar, Matt, Jake, & reviewer #2.

I very much appreciated the detailed comments that you all provided on this manuscript describing the "2.0" version of my widely-used phylogenetics R package, *phytools* – particularly in light of its exceptional length. I have taken considerable measures to improve the article in response to this extensive feedback. In addition to my detailed comments below, I'm including a full record of all manuscript updates in the tracked-changes document created using *latexdiff*.

I believe that reviewer #2's comments was slightly cut-off as the last remark that was recorded appears to end midsentence. Nonetheless, this reviewer had already provided very detailed feedback on nearly the entire manuscript, so I suspect that their review was almost complete. I made considerable changes to the manuscript, including substituting a complete example, in response to this reviewer's thoughtful & detailed commentary.

I hope that this substantially revised manuscript can be published with the *PeerJ* journal.

Sincerely, Liam

Editor's Decision MINOR REVISIONS

I agree with the reviewers that this is a much welcome and timely update on an important software package in the phylogenetic comparative methods ecosystem. The reviewers make a number of insightful and constructive comments for improvement. I'll add the following suggestions, mostly regarding the code itself.

1. Open Source licenses are not all equal, and therefore in Software and Data Availability, I suggest to give the actual license (GPLv2+, based on what the DESCRIPTION file says?). Given that there's no LICENSE file in the repository, I would also suggest to add that information to the README file there, so visitors don't have to guess (or know to look into the DESCRIPTION file).

I have now added this (GPLv3) to the GitHub as suggested.

2. I assume that for reproducing the manuscript's computations and figures, the author is referring not to the root directory in the liamrevell/Revell.phytools-v2 GitHub repository, but to the "peerj" subdirectory there? I suggest making this clear in the manuscript (Software and Data Availability section). Also, this repository seems to lack a license, and thus one needs to be assigned.

Yes. The root directory refers to the whole project, including prior versions of the manuscript and associated files. The files formatted for submission to PeerJ are in the folder "peerj." Since I want to make all the files of this project available, I have indicated in the manuscript that the project files and data are in the indicated repository "and folders therein," and I've also updated the GitHub README file to point visitors in the right direction.

I've also added a license (CC-by-4.0) to the repository as suggested.

4. To rerun the Rmarkdown source in this directory, upgrading phytools to the latest release on CRAN turns out to be insufficient. (It results in version 1.5-1, which for example lacks the butterfly dataset.) Upgrading instead to the GitHub repository's version results in a version considerably higher (1.7-7) than the one reported in the manuscript. It is thus a little unclear what the author's CRAN release policy is. It seems at least it would help if the Rmarkdown included a minimum

version check. If the author's CRAN release policy is indeed to make much less frequent CRAN releases than version increments in the GitHub repository, then this seems worth noting (at least in the README of the repo).

Excellent point. Thanks for observing that. Yes, I added a number of new features to *phytools* in the course of preparing this manuscript with the intention to have a new *phytools* version fully-consistent with the manuscript on CRAN well in advance of publication. The current CRAN version of *phytools* (1.9-16) will run all of the code of the manuscript as shown *except* the second to last plotting example, which uses data that I have added to the *phytools* package in response to reviewer 2 feedback (see below).

5. There is a bug in the data file for betaCoV.tree, which I posted as issue #131 on the phytools Github repository. (This bug causes the Rmarkdown source of the manuscript to fail rendering at line 971 -- even though the actual problem is in line 956 --; perhaps the author uses a filesystem with case-insensitive filenames?)

This should now be fixed. Thanks for catching this error.

Comments from the reviewers

Reviewer: Matthew Pennell

Basic reporting

This paper, which describes major updates to the highly used and highly cited (and probably more used than cited), R package phytools. This package has been under continuous development over the last decade, with new features added regularly. The paper then is a combination of "status update" and "tutorial". It is written in the informal style typical of a blog, rather than that of a conventional paper; I am honestly fine with this -- it is clearly written and references the literature where appropriate. I note that there is a lot of overlap between the material presented here and that which appears in the author's recent (excellent) book (Revell and Harmon 2022, Princeton). The author addresses this overlap, and I will leave it to the judgement of the editor as to whether this is problematic.

Thanks for your kind comments about our book. It's true that there is some overlap; however, I've done my best to ensure that this article both stands alone and usefully extends what we cover in Revell & Harmon (2022). No analysis or figure directly reiterates any of the analyses or figures of the book. The book covers (or attempts to cover) all phylogenetic comparative methods in R, whereas in this article I focus exclusively on *phytools*. Finally, the manuscript highlights a number of new features of the *phytools* package that were not yet available when we wrote the book.

Overall, this is a useful publication documenting the evolution of a very useful R package. The phylogenetics community is indebted to the author for implementing so many various methods into an easy-to-use and well-documented piece of software and thereby facilitating so many analyses that would have been done otherwise.

Thanks for saying this. I appreciate it.

Experimental design

I appreciated the detailed walkthrough. I think it is helpful to have all the little bits (e.g., the plot aesthetics, etc.) spelled out explicitly rather than left as an exercise to the reader. The code all runs (see caveat mentioned in 4) and the analyses work as advertised. I also appreciate that both the codebase and the article itself are version controlled.

Thanks for this comment.

Validity of the findings

I understand that the gamma test is technically done and described correctly but in the literature: 1) the null model is almost always a CR birth-death model and not a pure-birth model; 2) when the null is a CR birth-death, this is almost always considered a one-sided test (i.e., only values that are significantly negative are of interest). Of course, I understand that this is beside the core point that is being made; nonetheless, I think this introduces some unnecessary confusion about the use of this statistic.

Thanks for this comment. I generally agree; however, I might go one step further and argue that gamma is just a phenomenological characterization of the shape of the lineage-through-time accumulation curve (and, thus, a significant value of gamma merely indicates if the LTT is significantly concave or convex on a semi-logarithmic scale, independent of the cause). I have added text to this effect in the manuscript. Specifically, I now say (something to the effect of) "Note that since the pull of the present (Nee et al. 1992) means that our lineage through time plot is expected to curve upwards towards the present day for any non-zero rate of extinction, some have argued that γ should only be interpreted when *negative*. (I.e., that statistical tests of γ are properly one-tailed.) I don't subscribe to that view, inasmuch as I see γ as a phenomenological measure of lineage accumulation in our reconstructed tree whose positive or negative deviation from the statistic's expected value under pure-birth could have multiple underlying causes." In addition, I've updated the example to one (elapid snakes from Lee et al. 2016) that is significantly negative when incomplete sampling is not taken into account – but becomes non-significant using the MCCR test. I think that this example will be more interesting to the reader and I thank the reviewer for prompting me to identify it.

Additional comments

There is some inconsistency in the naming of objects -- both "." and "_" are used throughout. I would advise that these be standardized. Related to this, I encountered a strange problem (and perhaps this is on my end that when I copied and pasted from the PDF into RStudio (on Mac), the "_" character was not recognized and the code didn't run. I ran the code directly from the GitHub page and everything worked fine.

Regarding the varying use of "_" and "." in object names, this was done intentionally – but I since I realize this was inconsistent, I have strived to do it with even more intention in this revision. In general, when naming objects, I used a general identifier referring to the particular dataset e.g., "butterfly" followed by a period, then followed by the (e.g.,) model name (e.g., "ER_ordered" or "ARD_unordered", etc.). If the second or first parts consisted of two or more separable components, I would split them by a "_". Having said this, however, I realize that I was not totally consistent throughout the manuscript. I have updated various code chunks accordingly.

With regard to the second part of this comment (copying & pasting code chunks including the "_" character), I can't explain it; however, we had a similar issue with the typesetting of our book even though we (likewise) composed it using Rmarkdown language and the *bookdown* R package. In that case, double-quotes were rendered incorrectly. I'll be attentive to this issue when the final article is typeset for publication in *PeerJ* or elsewhere.

Reviewer 2

Basic reporting

This paper meets all the basic reporting requirements. Although one might argue language is not entirely professional more conversational in tone but it makes it more approachable.

I really appreciate this sentiment. This description ("more approachable") exactly captures my intention. Thank you.

Experimental design
NA this is a software description

Validity of the findings
NA this is a software description

Additional comments
I agree that a new paper describing phytools is timely, as there have been many additions since the 2012 paper including several important and exciting developments highlighted in this paper. The new book by Revell and Harmon is a comprehensive overview of PCM and as such there is still a need for a paper focused on phytools. Although an unusual form for a software description, it is useful, if a little self-indulgent at times. It is essentially a curated collection of vignettes focusing on the more recently implemented methods with some general information that links them together. The sections: discrete traits, continuous characters, diversification, and visualization provides a very clear general structure. Although, you could consider removing the diversification section, as there are other packages that provide much greater functionality for exploring speciation and extinction and I am not sure the functions in phytools provides novel methods. While the target audience is both new and returning users, my recommended changes focus on the likely needs of new users. I have thus suggested where additional explanations may be necessary to make this paper as helpful as possible to someone just getting started.

I agree with the reviewer's characterization of this article as "self-indulgent" – however, I hope that it can be both self-indulgent *and* useful. (Perhaps that *belief* is self-indulgent. I'll leave that to others to decide.) I appreciate the reviewer's attention to the usefulness of this article for new *phytools* users. I agree with the reviewer's assessment that diversification analysis is better-developed in other contributed R packages, such as *diversitree* (in fact, I note exactly this in the preamble of section 6 of the manuscript). Nonetheless, there is significant functionality in *phytools* (such as LTT plotting and birth-death tree simulation) that is often used in the literature and not precisely duplicated elsewhere. Furthermore, I think it is valuable to have important methodologies implemented independently in multiple R packages and I intend to extend some of the diversification method functionality of *phytools* in the future. As such, I prefer to keep this short section (in relative terms – nothing about this article is genuinely "short") in the manuscript.

I appreciate it is a hard balance between being exhaustive in the list of tools and providing examples and explanation, as well as between providing too much detail and not enough. However, if your target audience includes new users then I think some details are missing. For example, folks would very much appreciate a table of the major functions for each section. These additions shouldn't add to the length of the paper, as you can cut out some of the unnecessary aspects of the code, such as breaking halfway through an analysis to examine the output and judiciously editing the more expansive, conversational-style language.

I appreciate the suggestion of this reviewer; however, the *phytools* user manual index of function names *alone* currently spans five full double-column pages. *phytools* presently exports nearly 300 unique names (most of which are functions) in its name space, and that doesn't include class generic methods that are defined but not exported. Even though I exercised relatively little restraint in composing this manuscript, I would resist adding a table

– though I've tried to detail some additional relevant functions not explicitly demonstrated (with references, as necessary) in the preamble to each section.

Most software descriptions include context and comparisons to other available software and discussions of limitations. Phytools is a well-established, widely used, and extensive package and so I understand the wish and need to focus on its implementations. However, I think readers would benefit from some additional context and discussion of limitations of the methods as well as alternatives.

Scientific software descriptions take a wide variety of different forms and (if I'm being honest), I've never seen one quite like this article. (None quite as 'self-indulgent,' perhaps, to use the words of the reviewer!) That being said, some software papers, for instance, those purporting to provide a "faster" or "better" method than existing competitors do include explicit comparisons to existing software implementations. Since *phytools* doesn't make that kind of assertion, I would prefer to avoid adding additional explicit comparisons. (We do include a bit of this in our recent book, Revell & Harmon 2022, and I've done similar things on my *phytools* blog in the past: http://blog.phytools.org.) On the other hand, in response to both this reviewer's comments, and to those of reviewer 3 (Jake Berv) I have tried to be more attentive to at least *citing* other relevant software to different problems treated by the methods of *phytools* in the current manuscript revision.

Structure and language, there are several parenthetical paragraphs which I recommend changing, as well as some odd paragraph transitions and figures that could be placed closer to code provided to generate them.

I've reduced the number parenthetical statements. This is a (bad?) habit I acquired by footnoting a lot of parenthetical comments in our recent book (Revell & Harmon 2022)!

With regard to figure placement, the reviewer makes an excellent point. "Floating" figures is a feature of the LaTeX rendering from markdown. (It's possible to "force" them to be immediately following each code chunks, but this often results in an awkward layout – large white spaces or really obnoxious paragraph formatting.) If the article is accepted at *PeerJ*, I will be submitting the original LaTeX files and I'll be very attentive to how figures are placed during typesetting by the journal. (FWIW, I tried to both force the figures to land in place *and* force them to "float," but land *after* each corresponding code chunk. Neither option looked very good, so I can only hope that the *PeerJ* formatting for publication works better!)

It might also be worth considering bolding the warnings you provide e.g. lines 780-781

Much as a I appreciate this comment, aesthetically I would rather not have the published article populated with bolded sentences. Instead, I've moved this specific note into a more prominent position (above the code chunk, instead of in parentheses below it) so as to help ensure that it isn't missed by the reader. I have also gone through the entire text to be careful that other similar warnings are noted as prominently as possible.

Note: I did not run the code examples to check that these worked as expected.

That's understandable. Other reviewers report running some or all of the code of the manuscript, and I appreciate this reviewer's very careful attention to all of the details of the manuscript text!

Line 34: "Modern phylogenetic comparative methods are not new" – this is picky but recommend changing to phylogenetic comparative methods are not new, I think most would agree that PCM were catalyzed by Felsenstein 1985 but I am not sure it should be referred to as modern, which means present or recent past 40 years is a long time!

This is reasonable. Changed.

Lines 84-104: I am not sure so much space needs to be devoted to installing, if you are assuming some familiarity with R just tell folks to download from CRAN, as this is a very basic R task.

The reviewer makes a very good point. On the other hand, this section includes some information (e.g., checking package version numbers) that I'd like to keep in. As a compromise, I have kept the section but substantially shortened and simplified it.

Lines 105-118: the list of methods and models that you have chosen to highlight would be easier to read in a table and I suggest adding a column with a brief description of the use of each, so that new users don't have to look up for example what an extended Markov model is used for.

I think this is a wholly sensible idea. Indeed, in the original *phytools* paper (Revell, 2012) I included a table exactly of this nature. Since that time, however, *phytools* has grown enormously in size and currently exports nearly 300 unique names (most of which are functions) in its name space. As noted previously, this doesn't include class generic methods that are defined but not exported – and the number of such functions is very large and growing. As such, I would resist the suggestion to add one or more function tables. I hope the reviewer will forgive me.

With regard specifically to the "extended M*k* model" – this is just what Luke Harmon (2019) has called the standard discrete character evolution model used in phylogenetic comparative analyses. It's often called the M*k* model (without "extended"), following Lewis (2001); however, close attention to Lewis (2001) reveals that his model is defined more restrictively than it's typically used in phylogenetic comparative biology – hence the "extended" part! I've now added an explicit note of this model in the section 6 preamble.

Line 119: if part of your target audience is new users it might be helpful to provide a description of what stochastic mapping is for e.g. "Reconstructing the history of a discrete character with stochastic character mapping". This wouldn't be necessary if you chose to add the table suggested above.

This is a good suggestion. I've now added more text to the introductory paragraph on stochastic mapping so that the method better stands on its own.

Lines 129-133: it is most unusual to have a whole parenthetical paragraph and I think it confusing. Perhaps help readers by saying there are two popular ways to implement stochastic character mapping either first fitting a character model using ML or to sample using MCMC. Readers might also appreciate if you explain what they might want to consider when choosing one method over the other. I also think it would be helpful to provide context, especially comparing it to SFREEMAP also available in R.

To avoid confusion, I have now moved this content to the preceding paragraph as suggested. (I tend to read parenthetical paragraphs like footnotes – but I appreciate that this is a matter of taste, and mine may be unusual in this way!) I have provided more details about stochastic mapping as implemented in *phytools*, as requested by the reviewer – however, I have no specific recommendation about so-called "simulation free stochastic mapping," as implemented in SFREEMAP. This is a different method that's only loosely (in my opinion) related to stochastic character mapping, as traditionally defined. (That's not said to in any way denigrate this method! It's just a very different approach.)

Lines 138-139: I also would recommend being explicit about whether the example you provide fits one of the aforementioned methods or is distinct – again it will really help readers not familiar with these methods.

I've now clarified that is meant here – that *simmap* (the new generic method) can take as input a single model or an arbitrary set of models. If the latter, then it samples stochastic maps from each of the models from the set in proportion to their Akaike weights. I also add a sentence of explanation as to why I think this is a good idea (integrating over model uncertainty, etc.), partly in response to this comment but also as a response to a comment of reviewer 3 (Jake Berv).

Lines 151-155: another parenthetical paragraph = confusing. I don't think you need the parentheses you start "in this example" so readers are aware you are discussing something specific to the example and it is not a general topic. I do like the inclusion of this point, as for new users the differences between characters, factors etc. can be very confusing.

Parentheses removed as requested!

Lines 197-200: recommend moving the parenthetical sentences to below the example code, as it will help emphasize the code implements the weighted analysis not the alternatives.

Done.

Line 201: why use 1000 simulations, how does the user know how many simulations to run to get a reliable results?

There's no magic number. I've now added a short paragraph describing a reasonable 'rule of thumb' for picking the number of stochastic simulations to run.

Lines 211-215: either explain why the added complexity of the viridis palette is useful i.e. improves graph readability especially for folks with colorblindness or perhaps make it simpler and just pass it some colors.

The reviewer is making me do all kinds of new research! Just kidding. I appreciate it. I've now added text identifying why the *viridis* palette might be preferred, as well as a reference to the pair that originally devised it! (Apparently it derives from a conference presentation in 2015!)

Lines 220 -233: "perhaps most often users…" and "users undertaking a stochastic character mapping..", phrasing is not very helpful just because it is done doesn't mean it should be done. Help the reader understand that when looking at maps it is useful to a and b as you learn x and y.

OK. This text has been updated as suggested.

Figure 3 – recommend a single graph with the two transition rates overlaid, as it is easier to compare the distributions on the same graph.

Interesting! The reviewer is on to something because this is the *default* visualization method for objects of this class in *phytools* (to overlay the distribution of backward and forward changes for a binary trait), so obviously, I'd typically agree with them on this point. In this instance, however, I found that overplotting the two distributions resulted in a graph that was too messy and not readable. I now note this in the text.

Line 271 – before moving on to the next topic I think it is important to add a brief discussion of the benefits/limitations to the methods implemented. Can one account for variation in tree structure for the analyses and visualization? Can one account for rate variation across the phylogeny? Can one build density maps when there are more than two states? Can you include polymorphism?

The answer to all of these questions is "yes" – and more! Indeed, in the very next section I show how *simmap* can be used for a polymorphic trait evolution model. I've now added a short paragraph to the end of the section guiding readers towards this additional functionality.

Line 271 – good choice to highlight this new model! But perhaps it should be first as this is a new model of trait evolution and then you discuss mapping as a tool that uses these models. You can briefly discuss the available models for discrete traits including ones not mentioned currently then segue into the new model. This would be a little more comprehensive, which would be helpful for new users. You could move the mapping of a polyMk to the second section on character mapping

Thank you for this kind comment. With regard to the specific order of topics, I see the reviewer's point, but moving the polymorphic trait evolution model forward would put it ahead of the standard M*k* model that it builds on…. I'm happy with the current arrangement of topics – although I also appreciate that some compromises had to made!

Line 377-382 – not really necessary to look at the results halfway through.

Good point. I've removed this.

Figure 6 – move to after line 409, so that it is closer to the code used to generate it.

As mentioned earlier, this article was written in Rmarkdown and built using the *bookdown* package. It's possible to "force" figures to appear where they are created in the code, but this often results in awkward formatting and large white spaces in the rendered manuscript. (I tried it!) When the article gets typeset by *PeerJ* (if they choose to publish it!), I'll be sure to be attentive to figure placement.

Line 423 – why use base color graphics when you have already introduced viridis, an explanation for the choice is useful especially as you say it is difficult for you to read, surely you don't want to encourage others to generate graphics that are difficult for folks with colorblindness to read?

That's a great question! I wanted to show the forest <-> fringe <-> open (plus polymorphic conditions) in a 3D RGB color space – whereas *viridis* and other color scale palettes are 2D color gradients. For what it's worth, however, I'm colorblind (M-cone deuteranomaly: the most common form of colorblindness, at around 75% of all cases) and the palette is relatively easy for me to read. (OK, forest+fringe & forest+fringe+open seem quite similar; and fringe+open & open are similar – but these are, by design, meant to be close to each other in the 3D color space that I selected!) I've added a tiny bit more detail to this effect in the text.

Line 430 – why not move the LTT plot to the relevant later section? Also, I think it might be important to highlight the info given at the end in parentheses at the beginning - this plot is only meaningful with complete taxon sampling.

The reviewer makes a good point. I had to make the decision to group this analysis with the polymorphic trait evolution analysis, rather than with diversification analysis – but I can appreciate that not everyone will agree with this decision! I've moved the text as suggested, and also noted that this phylogeny contains 85% of described species for the group (so taxon

sampling is not too bad, in this case).

Line 438 – before moving on to the next topic I think it is important to add a brief discussion of the benefits/limitations to the methods implemented. What are the benefits of a polymorphic model vs a regular Mk model where the polymorphic states are code as separate states? Can one account for rate variation across the phylogeny? Can one account for variation in tree structure for the analyses and visualization?

Once again, the answer to all of these questions is (generally speaking), "yes." I've now added an additional short paragraph describing this.

Lines 439-451: the list of methods and models that you have chosen to highlight would be easier to read in a table and I suggest adding a column with a brief description of the use of each.

Yes, the reviewer made a similar suggestion earlier. As noted then, *phytools* now contains nearly 300 exported functions so a table enumerate all of them is not feasible. I see the reviewer's point, however.

Line 515: I think it would be very helpful for many readers to expand on this point, to explicitly explain why the different procedures can provide very different conclusions sometimes.

I agree with the reviewer that this point is underappreciated; however, I'm not sure what else to add other than they are different measures! (I already have text explaining each measure in the preceding paragraphs.)

Lines 566-569: if you are going to explain them do it fully – it is not clear which parameter $a and $y etc are..

Done.

Line 572: this is all very well to say that the percentage of burn-in can be changed but many readers may not know why this is important and even if they do, how they go about checking to see if 20% is sufficient. It would be helpful to provide this information.

I wouldn't dare try to cover Bayesian MCMC comprehensively, but I've added some text explaining what burn-in (and convergence to the posterior probability distribution) means, as well as what kind of tools can be used to assess it!

Line 521: move figure 9 here out of the Bayesian ASR section.

Again, the specific location of the figures in this review manuscript have been allowed to "float" and were optimized by the LaTeX renderer – I expect them to change during typesetting for publication. Nonetheless, I'll try to be attentive to where they end up finding themselves!

Line 620: before moving on to the next topic I think it is important to add a brief discussion of the benefits/limitations to the methods implemented. Can one account for variation in tree structure for the analyses and visualization? Can you include rate variation?

OK. Bayesian ancestral state reconstruction for continuous characters in *phytools* is not quite as flexible as stochastic mapping – but I've nonetheless added an additional concluding paragraph to the section that highlights some of the advantages of this analysis workflow.

Lines 621-622: it might be useful to point readers to other packages that contain additional multivariate methods here.

Sure. I've now added references to several other methods and packages, also as suggested by reviewer 3.

Figure 12: move to line 654 to follow the code for plotting it

The specific location of the figure was set by the renderer. I expect that it will change again during typesetting by the publisher, but will be attentive to their final locations.

Line 672: provide brief explanation of the different models here, the reader will be confused if they can't interpret the results and have to search the out from tropidurid.fits to understand them.

Done!

Lines 694-719: I am not sure this is necessary as they are not discussed in the text.

I see that the reason for this detailed print-out of the model results is not clearly explained. I have moved around some text as well as added additional details to make this clearer!

Line 720: before moving on to the next topic I think it is important to add a brief discussion of the benefits/limitations to the methods implemented or alternatives. For example, how does this compare to the models used to identify integration and modularity?

I'm not sure where to start, but this approach *is* one that has been used to identify integration and modularity. I have added an additional sentence noting this and point to one recent example from the primary literature.

Line 720: similar to the discrete section, it might work better to highlight this new model of trait evolution first and put it into context with the other models of continuous trait evolution at the beginning then move on to methods that implement these models. For readers new to these methods, briefly explaining the basic models first would be helpful – this can also include OU, which is an important model not mentioned currently and would allow phylogenetic signal to be explained in the context of Brownian motion.

I appreciate the reviewer's comment, but (even though this article is very long) I'm not trying to be comprehensive of phylogenetic comparative methods here! We did that in our book, and this is not a substitute for that. I have added some text to the current manuscript version to better enunciate the goals of the article.

Lines 747-753: it is hard to get a feel for what is biologically realistic here and I suspect it might depend on the taxonomic scope of the analysis. Are these parameter values 0.1 and 10 really bounding what is feasible or is it far too broad or not expansive enough? How does one determine this? Given this is a relatively new method in phytools, some more guidance would be helpful.

This is a great question! I've added a bit more discussion about the penalty term and also point interested readers towards Revell (2021, also published in *PeerJ*) where this method was originally described.

Line 760: projection not project

Thanks! Corrected.

Lines 765-767: move to the explanation of how visual inspection will help to line 761 before the explanation of the type of plot.

OK. Done!

Figure 13 move to just before line 768.

See earlier comments about figure rendering.

Figure 14 move to just before line 783

See earlier comments about figure rendering.

Line 787: before moving on to the next topic I think it is important to add a brief discussion of the benefits/limitations to the methods implemented or alternatives. For example, how does this compare to the variable rates model in BayesTraits or RevBayes?

To be perfectly honest, I don't know too much about these softwares! Based on RevBayes' tutorials, it contains a rate shift model (a la Eastman et al. 2011), but not a continuous evolution of the Brownian rate through time. The closest method that I'm aware of was published recently by Bruce Martin et al. (2022). I have now added a citation to that article, as well as to Venditti et al. (2011) and some other relevant literature in this area.

Lines 788-920: Given there aren't that many methods for diversification in phytools I suggest you consider excluding this section, as the functions available are not unqiue and there are several other packages that provide more functionality for modelling speciation and extinction.

I disagree with this comment (as indicated above), but appreciate that the reviewer has no additional comments on this section!

Lines 924-932: this paragraph seems unnecessary; just say in this final section I'll illustrate a few popular plotting methods not covered in the preceding sections.

OK. I see the reviewer's point, but disagree and would prefer to keep the paragraph. My purpose is highlighting the plotting methods of the paper that use *phytools* since this is not always clear!

Line 921: again, readers I think would find a table listing the various visualization options very helpful.

Once again, this is a valid point and would be an excellent suggestion for a much smaller R package – but in this case, I respectfully disagree.

Line 947: this shouldn't be a new paragraph.

Fixed!

Figure 18: should be just before line 981 putting it next to the code used to plot it.

See earlier comments about figure rendering.

Line 980: why use RColorBrewer over the others you have already introduced like viridis? It seems unnecessary complexity and will potentially add confusion for new users.

The *viridis* palette is a color scale. *RColorBrewer* generates aesthetic divergent palettes. I've added a brief sentence to the manuscript explanation this decision.

Line 990: instead of focusing on how laborious it was, perhaps state the type of data which is needed – lat/long coordinates.

Based entirely on the reviewers' thoughtful comments, I decided to update this example so that it would be more practically applicable. I've used a phylogenetic and geographic coordinate dataset from Poulakakis et al. (2020) of Galapagos tortoises. This allows me to also demonstrate how to use a higher resolution base map (from the R package *mapdata*) and restrict the map area to a circumscribed region. Note that to reiterate this example *precisely*, readers will currently have to update *phytools* from its GitHub page; however, I expect that the CRAN version will be updated by the time this article goes to press.

Line 1050: why use randomcoloR over the others you have already introduced like viridis? It seems unnecessary complexity and will potentially add confusion for new users. Here you felt the need to explain how to download it but not for RColorBrewer – why?

The reason for using *randomcoloR* was pretty simple – it offers larger palettes! In the updated example this is no longer necessary.

Line 1055: how would one restrict the map? Seems like a useful tidbit to include as many folks won't have a global dataset.

Based on the reviewer comments I've updated this example so that it now demonstrates this functionality, as well as the use of higher resolution maps. I think this now more closely tracks the typical use-case for this method!

Lines 1078-1083: explain why you prefer to reconstruct ASR on log scale and then back-transform to original space.

Done.

Line 1117: it looks like all the data are logged already? Perhaps w

I believe this comment is incomplete. Yes, the data were already log-transformed.

Reviewer: Jacob Berv

Basic reporting

I have evaluated the basic reporting requirements and the article appears to meet or exceed the indicated standards.

Experimental design

I have evaluated the experimental design requirements and the article appears to meet or exceed the indicated standards.

Validity of the findings

I have evaluated the validity requirements and the article appears to meet or exceed the indicated standards.

Additional comments

Review for:
phytools 2.0: An updated R ecosystem for phylogenetic comparative methods (and other things)

Overall, I have evaluated the PeerJ reporting requirements, and the article appears to meet or exceed the indicated standards in all areas. I have a few general comments that may improve a few areas of the text if the author chooses to implement them.

First:
I have no concerns about the technical implementation of the provided examples. However, I wonder if the content about model averaging is presented in the best way. I am unsure if there is consensus within the field of systematic biology that model-averaging applied in this context (e.g., starting around line 205) can be considered a "best practice." I do not think the author means to imply that it is -- however, I think a novice reader may read it this way. Many students getting into PCMs do not understand what they are doing and follow example tutorials without thinking too much. I have no statistical argument against model averaging; I think it makes sense in some contexts. However, it is unclear to me whether it makes sense to average the results across models that make incompatible statements about the data-generating process (e.g., ER and irreversible models), even if those model outputs are weighted according to their model uncertainty. The fact that phytools 2.0 can do this notwithstanding, I suggest the author reconsider how a naive reader may interpret the presentation of this example. If, on the other hand, the author believes that this *is* a preferred way to account for model-fitting uncertainty, I think it may be worth being a bit clearer on that point and perhaps providing some context as to when this is advisable over standard fitting and model comparison procedures.

This a really good point. In fact, I think that there is a really strong argument *for* model averaging in stochastic mapping, as well as for ancestral state reconstruction in general – but the reviewer is correct that this is not (yet) the standard practice. In my over enthusiasm, I neglected to emphasize that. Let me briefly try to make the case here – and then I'll also add an abbreviated explanation to the manuscript (although a think a proper article about this is warranted).

Let's imagine the hypothetical scenario of a binary trait with (as described in the ms.) a total of four *possible* (extended) M*k* models: ER, ARD, and the two irreversible models. Now imagine that 50.1% Akaike weight falls on the 1->0 loss only model and 49.9% on the 0->1 gain only model. By model selection, we would say that the 1->0 model is "best" – and our ASR would prescribe a 100% empirical Bayes posterior probability (marginal likelihood) that the global root was 1 & not 0! In fact, the probability at the global root is closer to 50:50 than 100:0. (Indeed this should be precisely our interpretation if we treat model weights as probabilities that the model is true, e.g., Link & Barker 2006.)

Now, of course this scenario is imaginary, and the case is unlikely to ever be this stark – however, it's not difficult to find a real empirical dataset that can make the same case. Just look at our sunfish feeding mode example. The best-supported model (by AIC) is actually one of the two irreversible models (non->piscivorous). Under this model any internal node with at least one non-piscivorous descendant *must* have been piscivorous. Nonetheless, given the substantial ambiguity of which model is actually best (0.38 Akaike weight for the irreversible model vs. 0.34 for ER), it would be absurd to claim that this means we *know* (without an ounce of doubt) the state at virtually every internal node of the tree!

Once again, though, the reviewer is correct that model averaging in ASR is *not* a standard practice. I have now added text that notes this, as recommended.

In any case, a naive reader may interpret the provided example as a suggestion of an optimal workflow rather than a possible workflow and end up with results that are difficult to interpret (e.g., the bimodality in Figure 3). [an aside: I think the bimodality in the posterior estimates for the number of changes may break the HPD estimator because these are no longer posterior estimates under a given model (indeed, what is the "model" when the parameter estimates are model-averaged?) -- I also think there may need to be 2 HPD intervals in the case of some bimodal posterior distributions (e.g., in some cases, the HPD would no longer be similar to to the 2.5 and 97.5 quantiles)].

This is another terrific observation by the reviewer; however, I disagree with the interpretation. To the contrary, the HPDs accurately capture the broad uncertainty in the number of transitions of each type. Based on an interpretation of model weights as probabilities (Link & Barker 2006), and assuming all possible models are in our set, there's a 0.38 probability of *no* changes from piscivory to non-piscivory (etc.). (Actually, it's slightly higher than that – probably because in some realizations of the ARD model histories with zero piscivory -> non transitions were sampled!) I now point out that model averaging has allowed our HPDs to more accurately capture the genuine uncertainty in the number of changes in the trait on our phylogeny, and try to explain why I think that's true.

Second:
I think it may be helpful to include an example of examining the likelihood surface for the discrete and/or continuous character models (like Fig. 9). Does phytools provide tools for visualizing model convergence (e.g., difficulty noted on lines 401-410) for these models? Perhaps this should be part of a "good" workflow, even if such diagnostics are rarely reported in the literature (I note this only because exemplifying a "good" workflow seems to be a goal of this paper). Perhaps the section after the polymorphic character model would be a suitable place for such an example, as the author already notes the difficulty in identifying the ML solution for these models (e.g., such as Fig 17).

Arrgh. This is a great idea and I'd love to add it. Unfortunately, the only example that it would be practical for would be the binary trait model for Centrarchidae. The best-supported polymorphic trait evolution model for the butterfly dataset has a total of 12 parameters – so visualization of the likelihood surface is just not something I can bear to entertain.

A minor detail:
622-629 -- mvMORPH, RPANDA, bayou, SLOUCH, PCMFit, and PhylogeneticEM have sophisticated machinery for fitting multivariate models -- perhaps worth mentioning some of these. For example, mvMORPH can fit multi-regime models in which the VCV can vary across specified regimes and/or according to different models (e.g., BM, OU, etc.). Another minor and related point that could be clarified in this example is the underlying model assumption (e.g., this specific example assumes mvBM).

Thanks! I have added various of these references to the revision.