Dear Dr. Thomas and reviewers,

Thank you for your careful reading of my manuscript. All of you raised very helpful points that have assisted greatly in improving the manuscript. In this revision, I have used all of your comments to improve the clarity in the writing, and to provide more detail in important areas. I have also added another set of analyses by the suggestion of Reviewer #3 that analyzes the simulated data under both branch length priors implemented in the *cophymaru* package. This addition provides a clearer understanding of the behavior of this method, and also helps to clarify some of the discrepancies observed in the empirical analyses when using different branch length priors. Below, I give comprehensive responses to each point raised by the reviewers.

Thanks again for your consideration.

Sincerely,
Caroline Parins-Fukuchi

CPF comments written in this color

Comments to the Author:
This manuscript describes a novel approach to placement of fossil taxa on phylogenies using continuous data. Three referees have provided insightful reviews. All conclude that the approach is novel and a potentially valuable contribution to the field. In relation to the question of novelty, I would like to see some discussion of Revell et al (2015; Placing cryptic, recently extinct, or hypothesized taxa into an ultrametric phylogeny using continuous character data: A case study with the lizard Anolis roosevelti. Evolution 69-4: 1027–1035). Revell et al use continuous data in a maximum likelihood framework to place taxa and superficially at least there are overlaps. It would be helpful to consider, both in the introduction and the discussion, how the new method differs and/or is an improvement. Is it primarily in the explicit focus on fossil taxa?

I had somehow overlooked this paper (although was aware of the method) when preparing the manuscript. My approach clearly does share several things in common with that of Revell et al., but differs in some important ways. I have added some discussion of their approach in both the intro and the discussion in the revision.

To answer the question briefly here, the major differences from the framework used by Revell et al., are 1) as you state, the focus on fossil taxa, which are not accommodated by their approach, 2) my method's expression of branch lengths in terms of unit variance, which eliminates the need to time-calibrate the topology *a priori*, and potentially reducing error stemming from morphological dating (and thereby allowing the fossil placements to be used as node calibrations in molecular dating), 3) the extension to allow the placement of multiple tips simultaneously, and 4) the use of a predominantly Bayesian framework-- this might or might not yield immediate benefits in and of itself, but the implementation and theoretical framework might be more straightforward to extend than that of Revell et al.

More broadly, the reviews identify numerous specific concerns that are generally points of clarification or a need for greater detail. Since these points are numerous and disparate I will not repeat them here but it is important in any revision to address each of these explicitly.

One more minor edit to add to the comprehensive lists provided the reviewers: line 505, The citation is incomplete.
.

Dr. Gavin Thomas

Reviewer(s)' Comments to Author:

Reviewer: 1

Dear Reviewer 1,

Thanks for all of your comments. I go through and address each point below, but your close reading has helped immensely in identifying unclear areas. I really appreciate both your enthusiasm, and highly constructive points of critique throughout the manuscript. Of particular help was your suggestion to add a set of simulations that explores both branch length priors. I performed this, and the results were

useful both in aiding a better understanding of the method, and helping to explain some of the patterns observed in the empirical analyses.

Thanks again!
caroline

Comments to the Author
Reviewer #1:

This paper presents an exciting and under-explored idea: inferring the placement of fossil taxa using morphometric data. The author presents a model describing the evolution of morphometric data along a phylogeny and provides new software for inferring the phylogenetic position of extinct taxa using this information. The performance is assessed using simulations and empirical data. There has been a significant lack of work on the use of morphological data in model-based phylogenetics, which may be especially valuable for fossil species lacking molecular data, and this work has the potential to make an important contribution to this topic. However, to maximise the potential and impact of this study the author must (1) substantially increase the clarity of the manuscript, (2) provide additional detail and justification for parameter choices, (3) include more extensive model validation and (4) provide a more substantiative discussion, including details of potential caveats and the role of this research in the broader context of existing tools.

I have incorporated the helpful suggestions from yourself and the other reviewers to help address point #1 here. I have also expanded more on the parameter choices used to simulate and analyze data throughout the methods section. To address points 2 and 3 together, I have also added another set of analyses on simulated data per your suggestion, to explore the effects of using different branch length priors. Finally, I have attempted to address point 4 by extending the discussion, including sections that explicitly explore anticipated challenges and shortcomings, as well as the potential to incorporate the framework into existing computationals tools.

All sections of the manuscript could be substantially improved with further editing. In particular, the methods and results sections are difficult to follow and lack sufficient detail for the reader to fully assess the performance of the new method. Additional simulations do not seem necessary for the present study, however, the use of additional statistics to explore the output would benefit the reader. The author could also consider eliminating some of the analyses presented in this study. For example, the analysis used in the simulations does not match the analysis applied to the empirical data, making the empirical results and the impact of different priors especially difficult to interpret. Some additional figures would also improve the accessibility of the methods. Below are detailed comments, intended to help the author address these issues.

I respond to each of the specific manifestations of these concerns below.

Provided the author approaches these concerns, this paper would make an excellent contribution to Evolution.

General comments
In its current form, the methods section will not be accessible to those unfamiliar with phylogenetic comparative methods (including BM, OU models etc). This may include many of the researchers who

are most likely to benefit from or be interested in the new model.

I have tried to give more clarity in this area, and pointed more precisely to relevant citations where appropriate.

It would be useful to get a better sense of what the data actually is. For instance, on line 185 the data is referred to as a "character alignment" but up to this stage the reader doesn't gain a clear sense of how the term "character alignment" could apply to morphometric data. It would be valuable to address this since the use of morphometric data in phylogenetic analysis is fairly novel. A description of the data could be provided at the beginning of the methods section and could be accompanied by a simple diagram illustrating the link between morphometric landmarks to the input actually used by the program. In general, the author could be more mindful of the terminology used to describe the data throughout the text. For example, empirical dataset 1 has "51 continuous measurements gatheredfrom pollen and seed specimens" (line 271) and empirical dataset 2 has "33 3D landmark coordinates representing...crania" (line 282). Are these different types of continuous trait data? In what format do they go into the program and how do the data points relate to the "character alignment"? It is also not totally clear whether the user inputs a fixed topology or if this is just one possible option?

I tried to clarify this issue in the first couple of paragraphs of the ‘Empirical analyses’ section. I also added a clearer description of the data format in the ‘Software’ section.

Insufficient information is provided that would allow someone to reproduce the simulations. For example, what were the parameters used to simulate the data? Needs to be clearer how "clean" versus "dirty" dataset were generated (page 13). This information may also be important for evaluating the results.

I attempted to provide more information here in the ‘simulations’ section. This section now also links to a Github repository that contains all of the matrices used and the R commands used to generate them.

The distinction between Bayesian and likelihood implementations is a bit confusing, and this is compounded by the poor correspondence between the simulations and the empirical analysis. It appears that the author applied the likelihood approach to the analysis of simulated data and the Bayesian approach to the empirical data. In addition, it is not possible to fully assess the role of different priors (applied to the analysis of empirical data), since different priors are not explored using simulations. The impact of the scaling approach applied to empirical dataset 2 (page 15) also isn't explored using simulations.

I agree that this comes off as a bit confusing. I’ve removed the references to likelihood implementations, as the distinction is mostly just technical and philosophical, and the method itself is similar to the framework used by standard Bayesian phylogenetics packages (MrBayes, etc.). But I was essentially trying to state that one can run MCMC without priors as a ‘likelihood’ approach. This is done occasionally in the statistical literature, but there is probably no reason to even mention it here.

As for the priors, since the choice of branch length prior did not significantly alter the results of the empirical datasets, I elected to simply use the most standard (exponential) prior when analyzing the simulated data. Generally, the literature finds the Dirichlet prior to improve estimation of branch lengths, but the general theoretical expectation is that topology reconstruction is fairly consistent across priors, except when predictable sources of bias are present.

Overall the results and discussion are very short, meaning the reader doesn't gain a sufficiently deep insight into the performance of the new method. Insufficient model validation is presented. The description of the simulation results would benefit from the inclusion of a broader range of statistics. The use of the distance between true and reconstructed node is limiting. For example, "1.85 to ~3 nodes away from the correct placement" (line 320) -depending on the tree, "3 nodes away" could be interpreted as being very bad. Applying additional statistics would provide a more comprehensive evaluation of the results, e.g. Robinson-Fould distance, clade correspondence and branch lengths could also be used to compare the estimated trees to the truth.

The reason that I don't use RF distances is that the majority of the topology is fixed in the comparisons, since this is presented in the manuscript as a fossil placement method rather than a fully fledged phylogenetic reconstruction method. The approach could be extended to perform that function, but I do not implement or explore this here. So as a result, the metrics that you suggest seemed to me to be less appropriate than the phylogenetic distance measure that I used in the manuscript.

My analysis of the simulated data was formulated to resemble the procedures and statistics used by Berger and Stamatakis (2010) (cited in the manuscript), who also reported only the distance measure used in my paper to compare sets of simulated fossil placements.

I definitely agree that '3 nodes away' could be very poor. This might be made better by the fact that this tree is relatively large (42 taxa), but still would be pretty far (although in the same neighbourhood, I suppose). I thought this issue begged more attention in the manuscript, and so I've added more information here.

The performance of different priors should also be addressed, otherwise the use of different priors in the empirical analysis is not that meaningful.
As I wrote above and indicate in the manuscript, the use of different branch length priors does not significantly alter the results. Since the simulations were designed to simply test whether my implementation performs in a reasonable way, I opted to keep things simple by only analyzing the data under the more standard choice of branch length prior, which are not expected to significantly alter topological reconstructions in clean datasets.

Finally, this section would also benefit from a comparison of the different filtering modes in terms of sensitivity and specificity using the datasets containing "clean" and "dirty" traits.

I present this information in tables 1 and 2, and have edited the text to more clearly refer readers there.

The final section of the manuscript could include more discussion of the new method in a broader context. For example, what is wrong with the model available in RevBayes (briefly mentioned, line 102)? Could the new model be incorporated easily into this software package or not? What about separating the rate and time parameters? This would be really exciting, as it would enable morphometric data be used in the estimation of divergence times (see also the comment below regarding this issue on page 7). Some details about potential caveats of the method would be valuable to include in the discussion. In general, the author should consider the point that they have achieved something really exciting, but the current version of the manuscript doesn't do this work justice.
These are all great points. I have tried to clarify in the introduction that RevBayes does not implement the model and method that I described in this manuscript. It may be possible (although I am not sure) to achieve something similar (without the data filtering aspect) using the syntax of the Rev language, but it would perform inefficiently and probably less useful to other researchers who may or may not want

to learn the Rev language. I have updated the manuscript to indicate that this approach could be incorporated fairly straightforwardly into an existing package.

I also totally agree that I did not give enough space to exploring the caveats of the method in the first version of the manuscript. I have added a new section before 'Comparison to other approaches' that explores potential issues in more detail.

Specific comments

Page 7, the justification for not separating rate and time needs to be improved or corrected. The author writes, "it is possible to simultaneously estimate divergence times and topology while analysing continuous traits, this can cause additional error" - it is not at all clear why this would be the case.

I attempted to clarify this somewhat in the manuscript, but the point was just to communicate that researchers may not want to use continuous traits to estimate dates in a morphological clock framework-- which as itself (as far as I know) not been explored or tested anywhere in the literature. I also tried to clarify in parts of the methods and the discussion that it *may* be possible in principle to do this, but that is a separate issue that I've explicitly tried to avoid in this manuscript.

Arguably, the joint estimation of topology and divergence times would be better, given that uncertaintyin the placement of fossils would be reflected in the divergence estimates. It is okay that the author didn't separate these parameters in this study but its not fair to say this approach should be preferred.

I think it is difficult to know at this time whether estimating dates from a continuous trait morphological clock would be preferrable than using the traits to estimate fossil positions and distance-based branch lengths. I wrote in the manuscript that I could envision applying the placements and estimates of uncertainty as input for node-dating methods that accommodate uncertainty in calibration placement (Guindon 2018, cited in the manuscript), which would have the same effect of causing uncertainty in the placements to impact the divergence estimates, without needing to perform morphological clock dating. But I try to do the total-evidence approach more justice in the revision by stating in several places that the method could be explored in a total-evidence dating framework.

Page 8, derivation of the likelihood required (at least in the supplementary material). For example, why is there a 2\pi in equation 4 (perhaps it should be \sigma?) It could help to include a figure indicated how the terms (i.e. v_internal) relate to the tree. This may be confusing to those unfamiliar with the PCMs. (page 8)

I understand that the likelihood needs to be derived, but it is described fully in three different citations, and so I hope that the brief description that I give in the paper give a sufficient background to communicate the gist to readers. I did edit the section a bit for clarity, and also added two additional citations that give full derivations of the likelihood and pruning algorithm.

Line 166-168, more justification is required for the empirical Bayesian approach. Informing the priors using your data is widely considered inappropriate. Include more detail or eliminate this statement from the methods, since it isn't discussed/used again in the paper?
This is a philosophical issue-- there are many well respected empirical Bayesians in the statistical literature. I've tried replacing this sentence to explain more precisely what is meant here. I also am

happy to remove this sentence in the final version if necessary, but it seemed worth noting, since using the empirical scale of the data to inform the scale of the prior should be a fairly innocuous practice.

Line 170-176, since this paper describes a novel method and a new piece of software, it might be good to include more information about the MH proposal ratios (at least in the supplementary material).
I agree that more clarity was needed here for the proposal ratios. Since I use ‘stock’ MCMC moves for the branch lengths and topology, I cite the relevant pages in Ziheng Yang’s manual for the full derivations, in order to keep the manuscript more compact.

Line 187-188, "Felsenstein gives a more detailed explanation of the algorithm", again, it might be useful to include more of this information in the paper.
I give a general overview of the algorithm as this has been described in detail elsewhere and the algorithm is not central to the method. Because this aspect is a reimplementation of previously published work, citation seems the most appropriate. I am happy to add more information if there is a specific issue that seems unclear.

Line 193-195 - "topologies achieved from this procedure are restricted to the construction of starting trees, while branch length inform the specification of the branch length priors" - it is not clear what is meant by this. It implies the data is being used to inform the prior, which is generally considered inappropriate (see above comment about line 166-168).

I’ve answered this above, but inappropriate by whom? Empirical Bayesian approaches are established area of model-based statistical inference. But the purpose here is just to move the compound Dirichlet prior into the correct scale, so that the distribution at least is within generally the correct order of magnitude.

Line 200, "traits are less likely to be excluded a priori on the basis of perceived unreliability" - isn't this a problem with discrete character matrices that was brought up in the intro (line 64-66)?

The difference here is that many researchers assembling character matrices from discrete characters introduce their own implicit biases when choosing which characters to include. The filtering approach that I adapt from Berger and Stamatakis is designed to identify reliable sites quantitatively using a model-- the reference scaffolding tree.

Page 10, Data filtering: It is not totally clear why the data needs to be filtered, since Bayesian methods should (in theory) be able to filter through the noise, provided an appropriate model is applied to the data. Excluding data potentially reintroduces bias, which the author seeks to eliminate (also related to the comment directly above).

In cases where certain data points are egregiously and obviously misleading, it would be helpful to ‘clean up’ the dataset. In a way, the filtering procedure is essentially a prior that assumes that the scaffolding tree comprised of extant taxa is good. In this case, traits that disagree with this guide can be de-emphasized due to their unreliability. And I am not sure in what way Bayesian methods themselves ‘filter through the noise’ intrinsically.

To address this concern, I have also added a paragraph on page 16 that frames the data filtering procedure a bit more clearly.

Page 11, line 206, How were the random trees generated?

I've attempted to clarify this oversight.

Page 12, It is not clear why the integer approach would be different to the floating point approach and result in poor mixing. This suggests there may be a bug and requires more investigation (or removing from the paper).

I have clarified this a bit based upon the comments reviewer #3 (above), who offered a good explanation for this phenomenon.

Line 247, this paragraph includes a discussion of the author's previous work, which seems to be very relevant background to the paper - this seems like it would be better placed in the introduction.

I agree. I added a paragraph in the introduction giving more background on continuous traits in general, and also setting up a discussion later on some of the caveats to geometric morphometric data.

Page 14, line 274-280. It is a bit unclear why only the run with the highest likelihood was retained, instead of the more common practice of combining all the runs for a better estimate of the posterior. It is also unclear whether this section also applies to the canids dataset.

Retaining the best run and combining all of them are two valid approaches to summarizing MCMC results. Keeping the best one is perhaps more similar to running MCMCMC, since the multiple runs are just being used to scope out the posterior surface and find the highest peak. Although I recognize that it is common in the literature to combine all runs, it is also common practice to retain the best run. The primary reason that I included this information in the manuscript is to highlight the necessity of performing multiple runs (regardless of the summarization approach used), since getting stuck in local optima is a very common occurrence with standard MCMC.

Page 16, Software: this section should be moved earlier and it should be made clearer that this package contains the implementation of the method and not just post-processing code. The inputs of the package should also be stated clearly, at the moment the fact that the extant topology is fixed (?) by the user is unclear.
Moved to the beginning of the methods, and tried to clarify these issues. I also added a figure (fig. 1) that should make the use of a fixed extant tree clearer.

Line 423, "reduces common sources of bias in morphological datasets", such as? Be more explicit here. Does the new approach introduce other biases?
Added a clause to the sentence to clarify the meaning. I also add a few passages throughout the text to incorporate the suggestion of reviewer #2 regarding the potential for the filtering approach to introduce error in cases where there is underlying gene tree discordance that is forcibly resolved by the scaffolding approach.

The reader would benefit from a better overall structure. For example, the empirical results are separated by a discussion on the general shortcoming of fossil taxa and future applications of the method (page 20, 21).
I have moved the paragraph in question to the 'Caveats' section below so that the empirical results are stated with less interruption.

Minor comments
Page 1, abstract, the description of the study and what is Page 14, line 274-280. It is a bit unclear why only the run with the highest likelihood was retained, instead of the more common practice of combining all the runs for a better estimate of the posterior. It is also unclear whether this section also applies to the canids dataset.achieved in the paper is too vague.
See my response above. I also clarified that the general description of the MCMC procedure applies to both datasets.

Line 4: "many fields", what would these be beyond the field of evolutionary biology? consider revising.
revised

Page 1, introduction, the structure of this section is a bit messy.
I have reorganized and added background in several places.

line 22, typo, character EVOLUTION
I can't find this typo.

line 28, "above", typo
I also can't find this

line 47, "subtle benefits" - too vague
removed

Line 50, "cited above", this is an odd form of citation.
What is a better alternative? The paper is cited above-- perhaps just remove the "cited above" here?

Line 52-53, this statement is incorrect, (i.e. overall the molecular data will make the likelihood smaller), perhaps you mean the overall influence of the molecular data will be greater?
Good catch-- I was just thinking about how the log likelihood is of a higher magnitude, which increases the weight of larger datasets. I've reworded.

Line 57-67, It is confusing that different issues relating to molecular data are described in terms of "layers" or "levels", since they are not related in hierarchical manner.
I agree upon rereading. I have reworded this to avoid the confusion.

Line 64, filtered = generated? since non-informative characters are not really scored in the first place, they can not really be filtered. Maybe you mean something else?
Rewrote this section to increase clarity

Line 64-66, this sentence appears to be referring to ascertainment bias. If so, it would be good to clarify this is what is meant and include a reference.
That is not what I meant-- rewrote to hopefully improve clarity.
Line 65, "At another level, the discrete character matrices most commonly employed in phylogenetics can often be difficult to adequately model" - is this really a problem with the data? Rather, doesn't this reflect our failure to develop adequate models? Better models could incorporate the issues described previously in this paragraph.

Possibly. But people such as Goloboff et al (2017) have specific concerns regarding the use of Markov models in morphological phylogenetics. I don't necessarily fully agree with those workers, but I do think that there are biological reasons to be skeptical regarding the match between substitution models and discrete trait evolution, which often occurs according to unidirectional shifts (eg., present → absent) instead of alternating back and forth (eg., A → G → A). Also, practically speaking, although some workers have proposed improved models, these are very young and mostly fairly incremental. On the other hand, it is fairly easy to justify the use of Gaussian stochastic processes (such as BM) with many continuous data, whose evolution and distribution across taxa does frequently appear roughly Gaussian.

Goloboff, Pablo A., Ambrosio Torres, and J. Salvador Arias. "Weighted parsimony outperforms other methods of phylogenetic inference under models appropriate for morphology." *Cladistics* (2017).

Line 102, what do you mean "through scripting"? This could mean suitable models are not available in RevBayes OR the user has to modify the input files (which they would have to do for any analysis using this program). This should be clarified.
Reworded for clarity

Line 109, "citation", typo
I am sorry-- I don't see this typo. Perhaps I edited it out before getting to this comment.

Line 139, "error in fossil placements", what do you mean by this? statement redundant with "fossil placements"?
Changed from 'error' to 'confidence' in the manuscript.

Line 149, why mention the VCV method at all?
Because it is by and large the standard approach in comparative methods and is well understood by most folks who deal with any of these methods. Since it is mathematically equivalent to the pruning calculation, it is worth mentioning that these are the same to give a frame of reference to readers approaching the paper from a PCM background.

Line 163, "uncharted territory" - this is a bit too colloquial
I think that this comment is stylistic, and generally prefer a more conversational tone.

Line 187, "likelihoods converge" - on what?
'Convergence' refers in numerical opimization to a point where values stop changing substantially, or an equilibrium state is reached.

Line 193, "messy datasets" - too vague
rewritten

Line 234, "fossil" - clarify what is meant by this. Perhaps using a figure?
I've meade a new figure to explain the procedure better (fig. 1).

Line 236, "Extinct lineages were retained", how do these lineages relate to the five "fossil" samples? A diagram of a simulated datasets may be useful.
Clarified by removing this clause. The idea was just that I kept tips that went 'extinct' prior to the end of the simulation. The five 'fossils' used for reinsertion are a subset of those.

Line 328, "noise may be structured to display" – typo?
Where is the typo? I have edited this sentence, so perhaps I caught it before reaching this comment?

Line 357, 358 - "group confidently" - This sentence is too vague, also intuitively it seems Ampelocissus wildei should cluster with the rest of the Ampelocissus clade, which it doesn't - is this correct or not ?
Removed ‘confidently’. I just meant that their positions are stable. As for Ampelocissus wildei, the original study (Chen 2009, cited in the manuscript) found a variety of placements, depending on the parsimony weighting scheme used, and how characters were coded (that study discretized the continuous traits). Also in that study, the fossil taxa frequently are placed outside of their respective extant genera. So it’s ‘true’ placement is not known with certainty. However, the taxon gets placed here across all runs under this method, and has comparatively high posterior support. So it is difficult to say whether or not it is correct.

Line 366, Figure legend, "The placements are depicted only in the subclade containing all 6 fossils" – this could be clearer, meaning the real tree analysed is bigger than that displayed in the figure?
Yes, that is the meaning. I clarified this and point to the full newick, which is available online.

Line 369, "The placement also changes across different runs" - this information is redundant when you also provide node support values.
Good point. I edited this paragraph.

Line 371, "the Ampelocissus parvisemina placement shows a 0.2 posterior probability, which decreases to 0.058 under the dirichlet prior (Fig. S1)" - Figure S1 actually shows 0.62 support for Ampelocissus parvisemina ?
This is true. I am not sure how I got that before… I have edited this paragraph to fix this discrepancy. I’ve also switched out the tree inferred under the exponential prior and that resulting fro the Dirichlet, since the other empirical test uses the Dirichlet, and the new simulations suggest that the Dirichlet is more reliable (and results in higher support in this particular case).

Line 382, "the confidence and uncertainty surrounding a reconstructed set of possible placements" - too vague, what do you mean?
I’ve rewritten this sentence for clarity.

Line 386, what is meant by "may be used to construct non-arbitrary, true Bayesian priors on node ages". Be more explicit. Also remove the word "true" from this context, since we never really know the true parameter values.
I’ve added a clause for clarity at the end of this sentence. Basically, I am trying to get at the difference between arbitrarily translating an expert’s assessment of a fossil placement into a probability distribution, and using probabilities in fossil placements achieved through statistical means.

Line 393, "used by" - do you really mean "used"? or do you mean "generated". This is an important distinction in this context. It could be valuable in terms of assessing the results to know how these trees were generated in the first place (i.e. in the methods description).
I mean ‘used’. Jones et al., use a fixed topology assembled in part from a previous parsimony analysis of discrete characters, and in part by attempting to summarize the consensus view of the field qualitatively into a tree.

Line 397, "grouping erroneously with Vulpes" - how do you know this placement is erroneous?

Clarified. This placement would just place the taxon into a pretty disparate family from the one that it is known to belong to.

Line 431, "would resemble the analysis of genomic rearrangements" - in what way?
I removed this clause-- I am not sure that there is any real benefit in devoting manuscript space to sufficiently explain the analogy.

Reviewer: 2

Dear Reviewer 2,

Thank you for your thoughtful comments. I really appreciate both your critical feedback and the nice words that you had for my manuscript. I have incorporated many of your suggestions in the revision. I respond to each point that you raise below.

Thanks again!
Caroline

Comments to the Author
Integrating extant and fossil taxa into singular phylogenetic hypotheses is important to evolutionary biology, because it allows us to better understand how old various groups are, what their ancestral traits are, and where ancient lineages lived. Primarily due to convention and dataset availability, the phylogenetic placement of fossil taxa has largely relied discrete morphological variation. Many traits that are are naturally continuous are discretized into bins (e.g. height in cm as "short" or "tall") so they are compatible with discrete methods. Most expect this creates unpredictable problems (though it'd be nice to see a study characterizing to what degree).

Placing fossils with continuous-valued morphometrics using continuous-valued models is a direct solution to a direct problem. The submitted paper proposes that we should do just that. The paper's approach assumes a molecular phylogeny as input, then using that as a scaffold, it fits the phylogenetic position of fossil taxa to it under a Brownian model of trait evolution. The result is a tree of extant and fossil taxa with branch lengths measured in expected units of Brownian variance per unit time. This topology may then be used to assign fossil calibrations to the relevant node or nodes.

Overall, I think this paper is a step in the right direction. The analysis is quite straightforward and competently executed. The writing is clear for the most part. As it is, the paper did not stir my biological curiosities very much. That could be fine since it's a methods paper. But I am also concerned that it might have limited appeal to the average Evolution reader. Perhaps it's better published in Syst Biol or MEE. Of course, I defer to the editors to make that decision.

Besides the line-by-line comments below, here are some major points that came to mind:

1. The argument for why we should apply the method could be made far more engaging by performing

a (approximate) discrete vs (exact) continuous fossil placement analysis on the empirical datasets.

I had considered this, but thought ultimately that including this would overly complicate the manuscript. Several previous papers explore the continuous vs discrete issue explicitly (Goloboff 2006, *Cladistics*, Parins-Fukuchi 2018, "Continuous traits can improve morphological phylogenetics", *Syst. Biol.*), and so I would prefer in this paper to build upon this previous work rather than using this paper as a platform to continue building this case. In addition, there exist so many methods for discretizing continuous characters, that I fear that the study would quickly become very cluttered by needing to qualify the potentially different results that could occur using different methods.

Perhaps more importantly, it was not my aim in this manuscript to make the case for the use of continuous rather than discrete characters for the purpose stated. Ultimately, I feel that continuous and discrete datasets are both (and will continue to be) important in placing fossil taxa. As is clear from my empirical analysis, while it is possible to place fossils using existing geometric morphometric data, it is not always straightforward, and can require some finessing to make work. It is also on my agenda to implement likelihood calculations for at least binary discrete characters so that combined discrete-continuous datasets can be analyzed.

Still, I definitely agree that this issue is important, and that it deserved more examination in the text. I have devoted some exploration of this in both the intro and the discussion, along with more citations of previous work in the area.

2. The motivation for the three-stage analysis is not entirely clear, especially where a total-evidence approach would work. Why not estimate molecular evolution, morphological evolution, and phylogeny jointly?

This would depend on the dataset being analyzed, I suppose. It would certainly be possible to estimate everything jointly, but that particular inferential paradigm can create problems with computational scalability and make it difficult to tease apart potential sources of error or uncertainty. When analyses are broken into discrete steps, it can become more straightforward to identify sources of uncertainty stemming from error and low information at each step, while avoiding the potential of creating unpredictable interactions between variables that may result in a complicated or misleading joint posterior surface that is difficult to reconstruct using MCMC. So I struggle to see the benefit in employing a total-evidence approach in this case. More importantly, although I recognize that many researchers prefer the total evidence framework, the scaffolding approach remains valid, and so my method can be seen as a complement to future work that may incorporate the models and framework that I describe here in a total evidence paradigm. If this was desired, it would be possible moving forward to incorporate my approach into a total-evidence framework if that is desired. This was done, for example, with the Fossilized Birth-Death model, which started as a prior in a standalone dating method, but is now more commonly used in a total-evidence framework in BEAST 2 and RevBayes.

I do provide some additional justification of the scaffolding approach around lines 40 and 178, and also suggest later in the discussion that the approach could additionally be incorporated into a total-evidence framework if desired.

3. Using morphometric coordinates in a comparative analysis can pose very real challenges that the author does not seem to have addressed. In particular, how does the method separate species variation caused by evolution from variation caused by rotation? Felsenstein has made many aware of this issue over the past ~10 years, though no solution is immediately available -- http://blogs.uw.edu/fhleqg/.

Dean Adams has also done important work in this area. Readers should be given fair warning that phylogenetic analyses require morphometric data that's handled with care!

I think this is a great point, and one that I agree deserves more space in the manuscript. I've added discussion of the potential challenges posed by geometric morphometric data in both the intro and discussion sections, especially around line 530. As best as I know, most authors (Catalano et al. 2010, Smith and Hendricks 2013, etc) have just handled this by making sure the coordinates are properly aligned using Procrustes transposition, which is what I do here. But I recognize that this procedure can be imperfect and frequently encounters errors.

4. Traits exist on different scales and evolve atdifferent rates. The simple site-iid Brownian model of trait evolution has difficulties with these real world conditions. Placing traits on a common but arbitrary scale might enable the method to converge, but it's not ideal. Allowing for site-rate heterogeneity would be a fairly easy to address this issue -- see Reference [1].

I completely agree. I am currently developing a method to accommodate site-rate heterogeneity in these data, but since the method stands as-is as a distinct departure from existing approaches, I aim to propose the Brownian model and inferential method here on its own without the added complication of adding all the extensions that would ultimately improve things further down the line.

5. The author is probably aware that the "dirty trait" scenario is not simply due to methodological error. Gene tree-species tree incongruence caused by deep coalescence, hybridization, gene duplication and loss, and other biological processes that result in trait trees that mismatch the assumed species tree. Reference [2] is new and interesting work on this topic. Noise from model misspecification or hemiplasy will induce subtler effects that may be difficult to filter. Worth considering.

I am aware of reference [2], and agree that the issue is extremely interesting when dealing with morphological phylogenetics. I could imagine a method down the line that seeks to model and/or accommodate such discordance (say, by using multiple reference topologies to filter morphological signal, perhaps). Still, I was aiming here to give a usable starting point to start cleaning up continuous morphological data to draw reasonable inferences regarding fossil placement, and so thiis issue extends somewhat beyond the scope, at least as far as implementation is concerned. I have, however, added some brief coverage of this issue in the manuscript (around line 41), since I agree that it is relevant and important in the context of filtering for signal.

Comments

11: "though" to "through"
fixed

22: "character" to "characters"
reworded

28: Berger and Stamatakis has not yet been introduced
fixed

39 "result" to "results" (to match "suggest" on the next line)

fixed

53-56: From my quick reading of B&S (2010), they assume (but don't show) that the imbalanced numbers of characters make it difficult to place fossil taxa. The total-evidence approaches that appeared shortly after 2010 show that fossil taxa can in fact be placed, probably because there is no molecular-morphological conflict for fossil taxa that have no molecules. So why is the scaffolding approach necessary here?

I address this a bit in point 2 above, but I am not certain why one would need to justify the scaffolding approach. There are different sets of questions that might better addressed using total-evidence vs. scaffolding approaches, and there is of course no reason why the method that I describe here couldn't be latter adapted into a total-evidence framework if one desired. I agree that the recent surge in total-evidence methods has provided ample evidence that fossil taxa can be placed in a total evidence framework, but use of a scaffolding approach can yield a separate set of benefits. For instance, if one wanted to place fossils to serve as node calibrations (ie., didn't want to use morphological tip dating), fossil placement approaches could be a quantitative alternative to expert opinion of fossil positions. This might be especially useful when combined with dating approaches that accommodate uncertainty in fossil calibrations (like Guindon 2018, below), by giving quantitative estimations of posterior confidence in fossil placements.

Guindon, Stéphane. "Accounting for calibration uncertainty: Bayesian molecular dating as a "doubly intractable" problem." *Systematic biology* (2018): syy003.

59: Before focusing on systematic error, the author should note that misleading signal can also originate from important biological processes, e.g. convergent evolution, hemiplasy.

Added a sentence here

64-66: Lewis (2001) is cited in the next sentence, but is also relevant here, since the raised this issue of ascertainment bias for variable characters and how to correct for it in a likelihood framework.

I agree that it could be relevant, but since Lewis provides a widely-used correction, I didn't feel that it was necessary to mention it here.

66-76: The author should be careful about using a simple symmetric Mk model to represent state-of-the-art methods for discrete morphological evolution. Especially so because their study uses a simple iid Brownian model. For example, Harrison and Larsson (2014) apply a discrete-gamma model for among-(morphological) site rate variation, like what's used in molecular analyses. Wright et al (2015) apply discrete-beta method that allows binary characters to vary gain/loss rate ratios across sites.

This is true, but methods sections generally indicate the use of symmetric Mk models (sometimes with/without Gamma, it seems). So even if the symmetric Mk isn't the 'state-of-the-art', it is still by far the most widely deployed and utilized model. I have added a qualifying statement at the end of the paragraph nevertheless.

95-96: I would consider both BM and Mk to be analogously simple models. Both suffer from similar inadequacies, yet Mk is strongly criticized where BM is not.

I agree that BM is a very simple model, but it does offer many properties that Mk does not. For instance, Mk (like JC) assumes that the stationary frequencies of character states are equal, whereas there is not really an analagous restriction in continuous traits and phylogenetic BM. Unlike Mk, BM also benefits from the justification provided by the central limit theorem-- although character state changes may conform better to non-Gaussian distributions over certain lineages/time spans, over deeper timescales, these all collapse into a standard Gaussian distribution. Of course, there may be other violations such as directional change, but their effects and prevalence are not well understood. Quantitative trait evolution might also proceed according to stasis and sudden jumps, ala more complex Levy processes, but the identifiablility between BM and more complicated models across a tree when branch lengths are expressed in unit variance are not clear. For instance, if jumps are frequent and normally distributed, this would be difficult to differentiate from BM in a modelling framework. So while I agree that both BM and Mk are very simple, the specifics underlying their simplicity differ. BM is more flexible by its' nature, while Mk's simplicity is very restricting and results in clear and egregious violations.

102-104: What specific challenges do PHYLIP and RevBayes fail to address?

I added a couple sentences in the manuscript here, but PHYLIP has an overly simplistic tree searcher that frequently yields poor results, and doesn't offer the option to later extend such features such as Gamma site-rate heterogeneity, as is discussed above. It is probably possible to specify a similar phylogenetic BM model to the one I use in the paper in RevBayes (perhaps by using the Rev syntax), but this would be restricted to a total-evidence approach (ie., not a scaffolding approach like I describe), and wouldn't have the trait filtering step. I am also not sure if RevBayes models specified through scripting are in any way computationally efficient, whereas *cophymaru* is fairly efficient (although by no means optimized extensively for performance). So they are just altogether different approaches from the implementation that I describe in the manuscript.

109: "citation?" needs to be fixed

Ack, I forgot to add that before submission. It is there now.

Equation 3: Equation 1 shows that sigma is defined per usual, so Var(X_t) should equal t*sigma^2

yes, thanks for the catch!

135: "requirement that phylogenies be scaled to absolute time" -- the time scaling is arbitrary, but they're generally scaled in units proportional to time (not necessarily absolute time)

This is true. I've restated.

135-138: It is not obvious to me what the "additional error" is.

I have clarified this somewhat. Essentially, I am hesitant to perform morphological clock dating using continuous traits, at least within the scope of this paper. This is another potential can of worms that I feel merits its own dedicated study (perhaps multiple).

138-140: It would be good to cite an example of the case in mind.

Done!

141: I am growing concerned that sigma*t is either defined imprecisely or that the author is using the wrong equation.

I'd defined it imprecisely. You are of course right, as above, that it should be sigma^2*t. Thanks for catching again!

165,166: Capitalize "dirichlet" throughout

Done

202-204: Filtering for concordant characters will inflate certainty in fossil placement. We do not expect all character histories to generate a VCV structure that matches the species tree for legitimate biological reasons.

The method essentially involves specifying a prior that assumes that the molecular scaffold is reasonable. Gene tree conflict notwithstanding, this is generally a reasonable assumption and practice in systematics, where for large groups, it is still fairly common to reconstruct species trees from only a small number of standard chloroplast/mt gene regions.

As for the biological processes that might cause discordance in certain characters, the filtering approach is designed to identify those traits that show convergence/homoplasy. And so in the Bayesian context, the posterior certainty would be accurate (and thus, not necessarly 'inflated' in a pejorative sense), since the prior imposed by the filtering procedure explicitly defines homoplasy relative to the guide tree. The same scenario would apply for other processes. I have also given some more exposition of this issue in the paragraph after this (around line 287).

206: Randomly generated how?

Clarified.

209, 210, and Equation 11: The summed variable should read \delta_{nj} ??

yes, I had missed the j. it is fixed.

213: How does Equation 7, which gives the contrast log likelihoods, produce integer values?

I am confused. It doesn't – it just gives the joint log likelihood of all of the nodes at the end of the pruning procedure. Did I miss somewhere where I say that it does? (if so, I really need to fix it!)

220: So are the weightings computed against MLE fitted reference and random trees, then those fixed weights are passed into MCMC? Not clear to me, unfortunately.

Ah, yes, the weights are stored and passed into the MCMC to be used for all subsequent likelihood computations. The weightings are computed using only the input guide tree composed of extant taxa. I added a sentence clarifying this (L. 271).

Equations 12-13 and 14-17: Just a stylistic note, but these sets of equations can be simplified with

indicator functions or equation case statements.

I'm not really sure what you mean, but thanks for the note. I will try to simplify.

233-235: The simulated dataset is fairly narrow. Stochastic error in the total tree (and the sampled fossils) must influence the simulation results. Why not use a new tree for each of the 100 replicates?

This is a possibility, but I used one tree in order to remove one possible source of variation. The sampled fossils were chosen to span the full range of timescale represented by the tree, so that the test was run on both deep and shallow nodes. There were also long and short branch lengths represented. Using a new tree for each replicate might cover a broader range of stochasticity, but the result would only shift up or down in absolute scale, and it would be less straightforward to figure out which timescales/branch lengths/etc were represented across all the replicates.

247-251: Reconstruction accuracy is potentially different from phylogenetic placement. And if covariance is not an issue, then simulating data with covariance and recovering the desired placement using an iid model will make a more convincing case to biologists who are concerned about ignoring trait covariance.

I cover this issue in a previous published manuscript of my own (that I cite) and so do not examine the topic in the manuscript. And ot be clear, I don't mean to suggest that covariance can uniformly ignored. This can be an issue in molecular phylogenetics also, for instance, when species tree inference is misguided by long spans of sites that exhibit the same misleading signal due to some legitimate biological process. But my only intention here and in the previous article is to suggest that the problem may not necessarily be more intrusive in continuous traits than discrete morpological and molecular characters.

For the empirical datasets, neglecting evolutionary covariance in datasets could potentially inflate confidence in topology. We wouldn't our model to register one hundred perfectly covarying traits as one hundred independent cases in support of a particular tree. Reference [3] discusses this in terms of effective sequence length.

Agreed, as I state in the previous point. This is, as you are certainly aware, a potential issue in any phylogenetic study. Anyway, I appreciate your bringing up these points, as the issue can be a bit challenging to write about. I have edited the manuscript in this area to attempt to clarify my meaning here a bit (starting around line 317).

287: Perhaps use "transposed" instead of "transformed" -- I first thought "transformed referred to what's described in lines 289-291.

changed, thanks!

286-288: It might help to note that the likelihood surface probably isn't entirely flat, it's just that the widths of its peaks narrow as the variance shrinks which causes MCMC difficulties mixing.

Fair point. Noted in the manuscript (L. 377).

289-293: All traits evolve under a single rate of Brownian motion and a single (implied) timescale. By rescaling the traits to some arbitrary variance, it's the likelihood contribution of each site will be also be

rescaled by some stochastic quantity partly determined by however the traits happened to evolve. Schraiber et al (2015) introduced a site-rate variation model for continuous trait models, analogous to the +Gamma for discrete traits. This could help mitigate the scale issue.

This is true, and is probably ultimately the most sensible solution. I am currently implementing a site-rates extension to the model to compensate for this, but don't feel fully confident in the specifics of the implementation yet, and so haven't written about it here. I have, however, noted in the manuscript that a site-rate model will ultimately be important for this approach moving forward.

328: Sentence is cut off -- "However, in reality, noise may be structured to display"

Not sure how I missed that. It is fixed.

386-387: "non-arbitrary, true Bayesian priors" does not mean much to me in this case. is the author referring to empirical node calibrations? If that's the case, what is the non-arbitrary, true distribution of a divergence time with respect to a node's age?

I have rewritten this passage in the text, but the idea is just that support values could be used as an analytical means to quantify the uncertainty accommodated by approaches such as Guindon (2018), as an alternative to attempting to quantify conflicting 'expert opinions'.

407: The exponential branch prior is known to have a strong effect on posterior tree length. I would assume the rate-1 exponential priors push the tree length to be much longer than that under the compound Dirichlet -- this allows traits excess opportunities for variation, which could systematically degrade node support.
Interesting idea! I've added more comparision and discussion of the different support and results achieved under each prior. I have also added a set of simulations under the Dirichlet prior that gives a more firm basis through which to interpret these differences in the empirical results.

Reviewer: 3

Dear Reviewer 3,

Thanks for reviewing my paper and for the comments. I appreciated especially your insightful comment marked at line 213. I address each point below.

Thanks again!
Caroline

Comments to the Author
This paper describes a novel method for placing fossil taxa on a reference phylogenetic tree using continuous morphological data. It reminds me of pplacer, but for continuous characters. Overall this is a very interesting method that I think will be useful to people in various biological fields. I just have a few minor comments and edits.

Line 171: Remove "or confidence"

fixed

Line 222: Remove “is”

I think I found where you mean? From this sentence: “instead highlighting signal taken to be more reliable”. Is that correct?

Line 213: It seems to me that the integer-weighted likelihood is just equal to 100 * the float-weighted likelihood. Therefore, with the integer weighting, in essence you are raising the likelihood to the 100th power, or copying your data 100 times. This essentially renders the likelihood as a point mass and turns the MCMC into a very inefficient maximum likelihood algorithm, whereby proposals with lower likelihoods are always rejected, and proposals with higher likelihoods are always accepted. On the other hand, the floating-point weight scheme is equivalent to the unweighted scheme if all characters are sufficiently concordant with the reference tree.

This is an interesting way to think about this. Thanks. I have ‘borrowed’ your explanation in the revised paper. If you are willing to share your identity, I would be happy to cite you through a ‘pers. comm.’ or other means.

It is interesting to me that all weighting schemes are outperformed by the unweighted method when there is no “dirty” data. Actually, this is what I would expect. When the assumed underlying model is correct (there is only a single tree topology generating the data), I would expect any data filtering to actually introduce a bias in the tree estimation process. However, when the assumption of a single tree is violated, your unweighted model might be outperformed by an alternative scheme.

Defintely. This is particularly the case for the ‘float’ scheme, which can result in different weights to each site, even when all the data are clean, due to stochastic error.