# Referee: 2

comments on M.Smith manuscript for Biology Letters V2

11 Nov 2018

The second read of this manuscript is fine with some minor exceptions.

I am glad to hear that my previous edits addressed the referee’s previous issues.

The first two items of suppdata that appear in the main text (lines 252, 256) don't appear at all in the zipped archive available via the figshare.com link (maybe due to how the publisher makes suppdata available across multiple 3rd party data storage sites & not the author's fault). I'd still recommend adding a clearly written summary of all of the suppdata items to the main text with obvious links where one can get them.

I’ve revisited the SI presentation in order to make this clearer, and distinguished Data (and Analyses) from Supplementary Text.

Also as detailed in the 1st review get rid of the qualifier "arguably" on line 32 (see comment for line 158).

Removed

line 17: the author moves between common terms like "node" and R terms like "edges" (line 47). I think here they're meant to be the same and this is worth clarifying.

I now use the word “groups” in place of “nodes”, and replaced “edges” with the more precise “bipartition splits” where this is intended

line 36: strictly speaking you mean "no tree shows less conflict than a polytomy" rather than "... is more accurate". I understand accuracy to encompass both minimized conflict and maximized edges-in-common, and a polytomy succeeds only in the former.

I’ve made this modification. Congreve and Lamsdell define accuracy solely based on the number of nodes incorrectly resolved, and thus fail to reflect the “maximized edges-in-common” that the referee fairly argues ought to be encompassed by the term “accuracy”.

line 53: It's hard to discover many aspects of mesquite's functionality due to its awful GUI. Nonetheless, mesquite does have a "shared partitions" calculator (open 2 tree windows > analysis:tree > values for current tree > compare with another current tree > shared partitions) that represents edges (or partitions) in common. This agrees with the author's assessment that the two newick examples here share no partitions in common and might provide readers with another option to compare trees (and is robust to comparison trees with differing taxon samples). By the way, do your methods require the same taxon samples to compare trees? Using commands such as drop.tip, RF.dist or distinct.edges in R (and shared partitions in mesquite) can calculate reasonable similarity metrics for trees with varying taxon samples and it would be worth mentioning if this is true here. It's also possible to address the issue of RF rewarding polytomies by standardizing RF values by proportion resolved, i.e., (RF.dist/(Nnode(tree1)/length(tree1$tip.label)-1)).

I now provide more details about the pitfalls of normalization, and have provided a figure to emphasize the importance of selecting a suitable normalizing constant. The Quartet package contains the function SharedSplitStatus and SharedQuartetStatus that accommodate trees with different numbers of tips, which I’d discuss if space allowed; that said, the trees analysed in this study all shared the same taxon samples, so did not require this treatment.

line80: Probably for the datasets examined here the TNT search commands are fine, but do the "level" and "chklevel" TNT commands perform branch swapping on optimal trees?

These commands to not perform branch swapping, but check parameters during the search to ensure an adequate sampling of tree space.

Without something like "bbreak=tbr;" following the "xmult" command I'd be concerned about underestimating the number of MPTs. FYI I've gotten faster TNT results with a search like "hold 1000; xmult=replic 10 rss fuse 5 drift 10 ratchet 15; bbreak=tbr;"

I thank the referee for these suggestions and have re-run the analyses under the referee’s suggested settings.

line 140: what makes a taxon:character ratio "large"? It's not clear to me why you write "even" before "350 characters".

I’ve used ‘larger’ rather than ‘large’ to cast this distinction in relative terms, and removed the ‘even’. In my experience, morphological datasets in the literature typically have 1–2 characters per taxon.

line 158: As noted in the first review, I'd still recommend pointing out that another reason behind "Previous results that favour Bayesian methods over parsimony" is when authors simulate their trees "using branch lengths common for all characters", and when they misunderstand ultrametricity. Contra line 32 (as noted above) neither of these points is really "arguable", as explained in detail by Goloboff et al. 2018a, b.

I’ve removed the ‘arguably’ rider, and now reference both Goloboff et al. papers in the introduction. To emphasize the point, I also now mention it at the location suggested by the referee.

# Referee 1

The manuscript has improved and raises interesting points, but I still have a number of concerns. These major issues are detailed below and summarised here: (i) stating researchers have not previously considered resolution to any extent is not the case; (ii) the argument for not using bootstrapping is unconvincing; and (iii) the main conclusion/recommendation of the paper is unclear - is it (a) that implied weights parsimony trees have comparable accuracy (number of correct quartets/splits) as Bayesian trees, albeit with more incorrect quartets/splits in the implied weights trees? Or is that implied weights and Bayesian estimation show no significant differences when the implied weights trees are resolved to the same number of nodes shown on the Bayesian consensus tree? I understand these are not mutually exclusive points, but it is important as (a) implies researchers should accept incorrect nodes, whereas (b) implies at a certain point of resolution all methods are the same in terms of accuracy.

## Major points

A central premise of the paper is that previous simulation analyses have downplayed the role of precision (precision here meaning the number of nodes resolved on a phylogeny). This is not the case. Several papers have all analysed the impacts of support measures, and/or have assessed overall support values (O’Reilly et al. 2016; Puttick et al. 2017; Brown et al. 2017; Goloboff et al. 2017; O’Reilly et al. 2018; Puttick et al. 2018): these papers need to be cited and discussed more fully in the Introduction and Discussion. Indeed, the main conclusions of a group of papers has been that accuracy is improved but at the cost of precision in trees in some methods (e.g., Bayesian Mk) compared to others (e.g., parsimony). (O’Reilly et al. 2016; Puttick et al. 2017; Brown et al. 2017; O’Reilly et al. 2018; Goloboff et al. 2018; Puttick et al. 2018).

I now discuss this more fully in the introduction. What does seem to be the case is that, when it comes down to it, these authors measure the ‘performance’ of a method solely in terms of its ‘accuracy’ (i.e. RF distance); better resolution is not seen as a reason to favour one method over another.

The author correctly points out bootstrap and Bayesian MCMC posterior probabilities may not be directly comparable (i.e, 50% bootstrap is not the same 0.5 posterior probability). Indeed, this has been discussed in the literature (Erixon et al. 2003). These metrics do, however, provide a means for comparing the levels of split support for alternative methods; they do not need to mean exactly the same thing to be useful. These metrics just need to provide estimates of branch support that can be used to collapse poorly-support splits for each method, and then these trees can then be used to compare the relative efficacy of methods. The argument presented here is that the bootstrap is generally unsuitable, and that it can’t be directly compared to posterior probabilities. Yet, surely the second argument is also the case for Bremer supports. Additionally, there may be difficulties in interpreting bootstrap values, but the usefulness of the Decay Index/Bremer support has also been questioned (Wilkinson et al. 2000; DeBry 2001).

I have now explored a selection of resampling methods (jackknife and bootstrap with frequency and GC measures of support) in addition to the original Bremer analyses. Bootstrap support transpires to be more effective at assigning incorrect nodes low support values, so I have re-run all my analyses using this method (with gratitude to the referee for advocating what is indeed an improved approach). This has changed some small details of my results, and I’ve updated the text accordingly.

The main of conclusion/recommendation of the paper is unclear. Previous analyses have advocating collapsing poorly-supported nodes (Brown et al. 2017), and this has been done in the most recent simulation studies (O'Reilly et al. 2017; Goloboff et al. 2018; Puttick et al. 2018). So is the main conclusion of this study that implied weights parsimony can be as equally informative as the Bayesian Mk model if a certain number of nodes are collapsed on parsimony trees? Is this also applicable to empirical studies? If this is the case then the lack of significant differences between methods suggests implied weights parsimony will have the same accuracy and precision as Bayesian Mk model as long as you collapse parsimony nodes to the level of resolution found on the Bayesian consensus tree summarised from the posterior sample. This is an interesting observation, but it could be used to say researchers might as well use the Bayesian Mk model to start with. This is further supported by the observation that the Bayesian Mk model seems to perform better (more accuracy) than implied weights parsimony at a similar level of resolution in simulations with low proportion of characters and tips; the Bayesian Mk estimation will give you the best proportion of the number of correct to incorrect nodes compared to the implied weights parsimony trees (Figure 2b).

I’ve clarified my recommendations for practice; a further advantage of using the bootstrapping approach is that I can recommend a particular value below which nodes should be collapsed. Perhaps more importantly, I have provided best practice recommendations for future simulation studies, which should improve the interpretation of future analyses that employ more realistic models to simulate data.

A theme in the manuscript is the desire maximise the information in inferred trees, but it is not always the key aim of evolutionary analyses to get a fully bifurcating tree. Most analyses ask question as to whether a group is monophyletic, if a taxon falls in a crown clade, or which clade was the earliest to diverge, etc., Therefore a tree does not have to be fully resolved to be informative; a tiny proportion of resolved nodes may be enough to corroborate or disprove the evolutionary hypothesis being addressed.

I agree that a fully bifurcating tree is not the sole aim of evolutionary analysis. This would be to prefer resolution over accuracy. I argue that our aim is to generate the *most* informative trees possible. The more informative a tree, the more questions of the type posed in this comment the tree can (a) answer; (b) answer correctly. Information is not an all-or-nothing concept; it is a measurable quantity, and more is better!

## Specific points

## Abstract

Line 13. This is not always the case. Implied weights can perform well and better than alternatives (Goloboff et al. 2018), especially with high-quality data (Puttick et al. 2018).

This comment is true, but the abstract presents statements that have been made strongly in the literature (without claiming to be exhaustive); more detail of the subtleties is now provided in the introduction

Line 16. “useful” is a term that needs clarification. As discussed above, a tree certainly does not need to be fully resolved to be useful; it is highly dependent on the question being asked.

Rephrased.

## Introduction

Line 32. This seems a strange argument to have here. If this argument holds and it is indeed fatal, would that would also mean the current study is also deemed invalid?

I’ve removed the term ‘fatal’. The current study is more concerned with the interpretation of simulation studies, which is a fundamental consideration however data are generated.

Line 35. I do not agree that precision has been ignored. Rather the emphasis is on the ability to resolve as few incorrect nodes as possible is favoured. Situations have shown parsimony methods can resolve almost entirely correct nodes in some instances (see above).

This is now discussed in more length in the introduction. What is true is that “outperforms” is treated as “generates more accurate trees”; if considered at all, a method’s capacity for resolution does not factor in to authors’ recommendations for model choice.

Also, there are cases in which probabilistic methods can be more resolved than parsimony methods after taking branch support values into account (Puttick et al. 2018).

Puttick et al. treat a 0.95 bootstrap support as though it were equivalent to a 95% posterior probability, which it is not. As such, this comparison is not particularly instructive.

Lines 43-66. Dedicating 3 paragraphs in the Introduction to description of the quartet and symmetric difference metrics is too long as the paper is not primarily concerned with comparing tree difference metrics. This section could be moved to the Methods, or the reader could be referred to the Supplementary Text in which these methods are also discussed in detail.

I agree; I have rewritten this section for concision and moved it to the methods.

Line 53. The use of different brackets is confusing and traditional Newick format brackets would be easier to interpret. Alternatively, a figure from the Supplementary Text could be used in the main manuscript, with the current Figure 1 and 2 joined into a single figure.

The brackets have been replaced with regular parentheses.

Line 57. I do not follow why allowing only a certain number of values is necessarily a weakness of symmetric differences

I now spell this out more clearly in the text, and have emphasized that this imposes a limit on precision (to avoid repeating the term “resolution”, used in reference to trees).

Line 62. Quartets do not contain all information a tree could provide. For example, branch lengths are not included.

To avoid this ambiguity I have replaced ‘relationship information’ with ‘topological information’

Line 63. Does the use of ‘resolution’ here refer to the precision of metric values produced by Quartets compared to symmetric differences? If so, could a different word be used to avoid confusion with resolution meaning the number of nodes on a tree.

This is a great suggestion; I’ve used ‘precision’ instead. (I’ve also made it clearer that the quartet and partition measures are both symmetric difference measures, differing only in their unit of measurement.)

Line 71. Maybe clarify Gamma parameter to ‘rates sampled from a discretised Gamma distribution’ or similar

Implemented.

Line 70, 74. The use of the abbreviations ‘CL’ and ‘OR’ is not consistently applied in the rest of the manuscript. Please clarify

I now use these abbreviations throughout.

Line 76. The supplementary tree from Wright and Hillis (2014) contains 75 tips and 63 internal nodes, so is not fully bifurcating. This tree was presumably also used by O’Reilly et al. (2016), so to clarify, was the tree used in the manuscript here the same as tree (75 tips, 63 nodes) used by Wright and Hillis (2014) and subsequent papers, or was it a fully bifurcating tree sourced by the author? This is an important issue because Wright and Hillis tree contain polytomies (75 tips, 63 internal nodes), and so would have knock-on effects for the rest of the manuscript.

A single bifurcating tree (i.e. no polytomies) is used, following OR, as indicated in the text and references; the tree itself is provided in the supplementary material.

Line 108-118. It is not clear why some areas of the plot being impossible is problematic. As these areas will be the same for all plots for all tree estimation methods, they will be equivalent in like-for-like comparisons.

I have clarified that the issue is one of interpreting the plots; a ternary plot makes clear the full range of possible values, and a unit change corresponds to a unit distance in any direction on the plane.

Line 126. Should ‘increase’ be ‘decrease’ here? Please clarify

I’ve added additional text to clarify that replacing correct groups with polytomies means that a tree becomes less similar to a reference tree.

Line 128-130. It is possible most/all nodes could be incorrect on an equal weights or Bayesian Mk tree, so there could be a situation in which decreasing resolution on these trees would improve accuracy.

This is indeed possible, but it is not generally the case. I’ve added a reference to a SI figure that shows the effect of reducing resolution in each individual dataset.

Line 133. Surely it is a good thing to value more accuracy over imprecise resolution?

I’ve deleted this clause to avoid sowing confusion.

Line 136. Maybe replace ‘generated’ with ‘estimated’ or similar

Changed to “obtained”

Line 141, 148. Phrases here such as “equally useful” or “reasonably accurate” are ambiguous

I’ve replaced “equally useful” with the more precise “equally informative” (as the quartet divergence metric is a measure of a tree’s information content). I’ve linked ‘reasonably accurate’ to the figure, which depicts the quantitative difference in accuracy.

Line 154-157. This again assumes it is vital to have all nodes bifurcating to be evolutionary informative. There could be an example in which 1000s of sponges, cnidarians, ctenophores, and bilaterians are analysed. If each clade is monophyletic, only a handful of nodes need be resolved for the higher-level relationships, e.g., (sponges, (cnidarians, (ctenophores, bilaterians)), for the tree to be evolutionary informative. All other nodes in the tree could be unresolved.

I have added a phrase to reiterate that accuracy and resolution are complementary aspects of information. The referee’s example is a tree that contains some information. A tree that was more resolved, more accurate, or both, would contain more information; a tree that was less resolved or less accurate would contain less information. A perfectly resolved, perfectly accurate tree would be fully bifurcating and contain the maximal possible information about the ‘true’ tree, but it is not the case that a tree that contains less than perfect information contains no information at all.

Line 161. It is necessary to review the literature here in which previous studies have considered accuracy and precision (please see above).

To complement the detailed review in the introduction, I’ve added further discussion here too.

Line 163-164. More detail needs to be provided as to the further developments available to probabilistic models, such as tip dating, as well as topology tests that can be conducted using all methods (Templeton Test, Bayes factors, etc.).

I’ve added a reference to Bayes Factors, but in the interests of space cannot provide a full and detailed review of the other advantages of methods, which I expect will be familiar to practitioners in any case.

Supplementary Figures. There are a large number of Supplementary Figures – could this be cut down?

I’ve moved the Supplementary Figures to the Supplementary Data repositories. I’ve chosen to retain them (if in a less visible position, to avoid overwhelming the reader) in order to make my results as accessible, transparent and reproducible as possible. I feel that it is better to present these data rather than expecting the reader to generate the visualisations themselves. To orient the reader, I’ve introduced additional figures as summaries.

Supplementary Text. Page 4, type “maxximal”

Corrected.

## References

Brown JW, et al. 2017. Bayesian and likelihood phylogenetic reconstructions of morphological traits are not discordant when taking uncertainty into consideration: a comment on Puttick et al. Proc. R. Soc. B 284: 20170986.

DeBry RW. 2001. Improving interpretation of the decay index for DNA sequence data. Systematic Biology 50.5: 742-752.

Erixon P et al. 2003. Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. Systematic biology 52.5: 665-673.

Goloboff PA., et al. 2018. Weighted parsimony outperforms other methods of phylogenetic inference under models appropriate for morphology. Cladistics 34.4: 407-437.

O'Reilly JE, et al. 2016. Bayesian methods outperform parsimony but at the expense of precision in the estimation of phylogeny from discrete morphological data. Biology Letters 12.4:20160081.

O'Reilly JE, et al. 2018. Probabilistic methods surpass parsimony when assessing clade support in phylogenetic analyses of discrete morphological data. Palaeontology, 61:105-118.

Puttick MN, et al. 2017. Uncertain-tree: discriminating among competing approaches to the phylogenetic analysis of phenotype data. Proc. R. Soc. B 284.1846: 20162290.

Puttick MN, et al. 2018. Probabilistic methods outperform parsimony in the phylogenetic analysis of data simulated without a probabilistic model. Palaeontology (2018).

Wilkinson M, et al. 2000. A chain is no stronger than its weakest link: double decay analysis of phylogenetic hypotheses. Systematic Biology 49.4:754-776.

Wright, AM and Hillis DM. 2014. Bayesian analysis using a simple likelihood model outperforms parsimony for estimation of phylogeny from discrete morphological data. PLoS One 9.10:e109210.

# Editorial office comments to authors:

\*Please ensure that your data is cited in your reference list as described here https://royalsociety.org/journals/ethics-policies/data-sharing-mining/

Confirmed

\*Please confirm in your cover letter whether all the figures are your own or whether permission has been obtained for their use

Confirmed

\*If you have any images that can be used to promote your article on social media (should this be accepted) please upload them as a supplementary file

None are suitable

\*Please ensure that your supplementary files themselves include the title and authors of your main manuscript, and that you also provide an appropriate title and description for each of these on ScholarOne

Confirmed