Dear Dr Freckleton and Dr Grieves,

We would like to thank the reviewers for the positive feedback and the many useful comments on our manuscript (id: MEE-19-12-885) " Assessing the Impact of Continuous Traits on the Evolution of Discrete Traits: The Ancestral Condition Test ". We have taken the advice of the reviewers to revise the manuscript, we have evaluated our approach under additional conditions, and reworked parts of the manuscript to improve the clarity of our approach. We believe these changes have significantly improved the manuscript. We appreciate the careful and detailed comments that we have received and have responded to each of the reviewers’ comments. Our replies are given in *red italic* following each comment.

With kind regards,

Heath Blackmon

Reviewer: 1  
  
I'd like to see the method stress-tested a little bit more. I'm convinced the model works well under the scenarios presented, but I'd like to have a better sense of the boundaries of the model -- under what circumstances would it be likely to "break"? That is to say, I have no reason to expect the model to be flawed, but in the same manner that we try to appreciate the limitations of linear regression (e.g., see ?datasets::anscombe in R), it's helpful to know what to look for in a dataset.

*We agree that stress-testing is key to the continued progress of phylogenetic comparative methods. We have added a simulation scenario designed to create a biologically plausible but false signal (scenario XXX). Using this approach, we illustrate one case where our approach can be positively misled and we discuss possible solutions under these conditions.*

*Methods lines XXXX*

*Results lines XXXX*

*Discussion lines XXXX*

*Figure XXXXX*

Line 102: please expand on the rational for pruning continuous characters to only include those that have retained a particular binary state.

*We have changed the structure of our explanation to begin with the most general approach without pruning and now present the pruning approach as an option that can allow us to perform a more conservative test.*

*Lines XXXX*

Line 113: I might be reading this wrong, but is there a word missing near subtending?

Lines 117-122: Is there any risk of bias when you prune away observations? put another way, can you break the method by contriving a plausible situation in which pruning might be expected to cause problems?

*Because pruning is based on the observed discrete trait when ancestral state estimates are poor or uninformative we might expect a failure. To explore the potential for this we have added a simulation (scenario XXXX) where rates of transition are very high leading to many changes on most branches. Under these conditions we show that pruning leads to XXXXX.*

*Methods lines XXXX*

*Results lines XXXX*

*Discussion lines XXXX*

*Figure XXXXX*

Lines 129-131: Sentence fragment

Lines 195-203: Is there any precedent in the literature for branch scaling quartiles for simulations? It seems quite intuitive, but I'm unfamiliar with what sort of distribution would be expected.

*We now give a citation to another paper that has take a similar approach to create rate heterogeneity on a tree. In our case the distribution of trait values is driven by the underlying evolution of the continuous character and the strength of the scaling factor. With the most extreme scaling factor applied and a unidirectional model for the discrete trait we find that all tips transition to the derived state in the scaled subclade. In simulations with equal rates of transition in the discrete state and very high rates we find that tip states in the scaled subclade are fully randomized regardless of the ancestral state for that region of the tree.*

*Supplemental figure XXXX*

*Lines XXXX*

Lines 417-430: I'd like to see a bit more about data visualization (what potential pathologies should we look for) and a slightly longer explanation about the meta-analysis route

Figure 3: Yikes (regarding threshold model performance)!  
Table 1: I was initially confused at the table. I think it would be helpful to very briefly describe the other utilities of the package, even parenthetically at the end of the sentence on line 273.

General: I would say this is an optional suggestion, but it's always nice to have a bit of guidance on best practices for reporting results (i.e., when a non-theoretician empiricist uses this method in their work, any pointers on how should they convey results?)

*We agree this can be helpful and now include a short section on the typical way that these approaches might be reported.*

*Lines XXXX*

Reviewer: 2  
  
I only have a minor suggestion: the section "Ancestral Conditional Test" under "Methods"  is a bit verbose and the section would be better understood if it is presented in a slightly more quantitative way using notations.

*In response to various comments the methods section has been extensively reworked*

*Lines XXXXX*  
Reviewer: 3  
  
  
Line 47: change to ‘discrete traits’. Traits can also be multstate.

Line 50: This review is from 16 years ago. Please provide some more recent citations. For example, the methods cited later on this same paragraph have all being published after this review.

Line 53: I think these methods allow for working with multistates. Please change to “discrete” traits instead of “binary”

Lines 79 and 80: Maddison and Fitzjohn work points to an interesting issue that is more specific to phylogenetic trees than general false-positive rates. Fase-positive rates is the problem that directly affects the researcher using these methods, but the underlying cause described by Maddison and Fitzjohn is worth mentioning in my opinion. I suggest the authors mention the “pseudo-replication” problem.

Lines 88 to 89: One concern here is that as we increase the number of taxa we are also increasing the heterogeneity in the phylogeny. With empirical phylogenies, as we increase the number of species, we also increase the chance that we are sampling from a distinct regime of trait evolution and methods that can accommodate rate heterogeneity start to become necessary when phylogenies get larger and larger. I think that it is worth mentioning a caveat to this method, that we get more power as we increase the number of species but we also increase the chance that homogeneous models (such as the single BM and the MK model) will become inappropriate. A possible solutions is to extend the test to apply multiple rate BM models (maybe even something like BAMM) and a hidden-state Markov model (something like CorHMM).

*We agree this is an important issue and we have added a short discussion of it. While we have not extended the method to incorporate more complex models of trait evolution we have added a simulation scenario that incorporates a two optima OU model for the continuous trait to illustrate XXXXXX*

*Discussion of problem line XXXX*

*Additional scenario:*

*Methods XXXX*

*Results XXXX*

*Discussion*

Line 101 to 103: It is unclear what “estimate of the continuous trait” means. The example between lines 98 and 103 is vague. I was not able to understand which analyses is being done here and why is this a suitable solution.

Line 102 to 103: Under a Markov model of trait evolution, distinct from a parsimony approach, a lineage that seems to have retained a discrete state could have undergone multiple transitions instead. Unless the authors are assuming a unidiretional transition scenario. It is not clear.  
  
Methods section:  
  
Lines 117 to 124: This method should be described in a more general manner. I believe the authors are assuming that the unidirectional case is the most common and the bidirectional case is the special case. However, in the majority of comparative methods, a free Markov transition matrix is the general case and the constrained transition matrix (i.e., the unidirectional transition) is the special case. Please explicitly state which scenario applies to which part.

*We agree that our approach to describing the restricted case first made the explanation a bit unclear. To increase clarity we now present the bidirectional case first and then expand on this by showing how it can be restricted to a unidirectional case and why this may be a particularly powerful approach in certain biologically realistic scenarios.*

*Methods XXXXXX*

Step 1:  
  
The separation between the bidirectional case and the unidirectional case should be made more clear here too. Again, the more general case is the free MK transition matrix. The constrained model is the special case.  
  
Step 2:  
  
Stochastic character mapping can be conducted using either Bayesian MCMC or Maximum Likelihood estimates of the Markov transition matrix. Please describe which of the type of estimates should be used. The Bayesian MCMC is preferred since integrating over the uncertainty in parameter estimates is very important when focusing on ancestral state reconstruction. The stochastic mapping approach can be used to take into account the uncertainty in character histories by making forward simulations of the state transition along the branches. However, if one only uses a single Markov transition matrix to produce the stochastic simulations, then the approach is not taking into account the uncertainty in parameter estimates, only of character histories. Taking samples from a Bayesian MCMC can allow to do that.

*Cogitate on this and make a decision*  
Step 3:  
  
On lines 145 to 147: In the bidirectional case both type of transitions can happen along the same branch. If relative rates of transition are fast enough, then multiple transitions can happen along the same branch. From personal experience, this occurs very often with Markov models estimated from empirical datasets. How are these parsed? Are these branches reliable as a summary statistics?

*We have added a simulation scenario XXXX that explores the impact of higher rates of trait evolution in the discrete trait. When these rates are very high we begin to have many transitions on each branch which allows us to explore the impact of having multiple transitions. We find that XXXX*

*Methods XXXXX*

*Results XXXX*

*Discussion XXXX*

One lines 149 to 150: Repeating this process across stochastic mappings will take into account the uncertainty in the estimate of the ancestral state for the discrete traits only. However, on Step 2 the ancestral state estimation for the continuous trait was performed using Maximum Likelihood, meaning that there is a single point estimate for the ancestral state of the continuous trait for each of the nodes of the tree. This test is focusing on ancestral state reconstruction. Uncertainty in ancestral estimation for both traits is very important. One way to do this is to use samples from a posterior distribution for both the Markov model and the Brownian motion model. Then a posterior distribution of ancestral states for the continuous trait can be generated by sampling rates of trait evolution from the posterior.

*Investigate time cost of extending Brownian sample from MCMC – probably say we that it was a hard pass.*  
Lines 150 to 155: How is the “estimated ancestral condition” pooled when there are multiple transition along the branches? When transitions rates are high enough, stochastic mapping approaches can produce multiple transitions along the branches. It is unclear which “type” of node that would mean.

*If the daughter branch has one or more transition into the other character state then it will be classified as a “producing node”. We have clarified the description of this in line XXX. We have also added the simulation scenario XXX that lets us better understand the impact of these possible multiple transitions along a daughter branch.*Step 4:  
  
Lines 159 to 162: The null distribution is produced by sampling ‘n’ nodes from a pool of ‘N’ nodes with the same estimated state. Where ‘n’ is the number of producing nodes. The maximum number of independent samples can be computed as the combination of ‘N’ into sets of size ‘n’. As far as I understand, there could be less than 10^3 possible combinations. Meaning that if the user set a really high value for the number of simulations to produce a reliable null distribution, the produced null distribution might have many pseudo-replicated samples. Is this correct?

Huelsenbeck et al. (2003) describes the stochastic mapping and also a test of correlation between discrete states using the overlap of stochastic mapped histories. They generate a null distribution by simulating new independent datasets with rate value equal to the posterior values sampled during the Bayesian MCMC. One suggestion here is to follow a similar simulation approach. The authors could simulate continuous traits under the same rate estimate from the underlying BM model. Since the generated data is independent from the discrete trait, one can estimate a null distribution. You could go one step further and also simulated discrete states from the estimated transition matrix. The same estimation approach (Steps 1 to 3) are repeated with each of the simulated datasets (including parameter estimates) and a null distribution of the test statistic is produced. This approach has the advantage of not producing pseudo-replicate samples and not be limited by the number of nodes in the phylogeny.

*We agree this is a better approach to creating the null distribution and have changed* *our function to produce the null in this manner.*

*Methods XXXXX*

Line 300: “detect a relationship”  
  
  
  
Line 328: “similar performance when applied”  
  
Lines 328 to 334: Here the focus of the results is the empirical phylogeny. However, in my opinion, the most important part here is the empirical trait distribution. Not always the heterogeneity in diversification rates is reflected on the heterogeneity of trait evolution. For this test, verifying a case in which we have realistic trait variation seems to be much more relevant than a case of an empirical phylogeny. Maybe the authors could focus on the “empirical trait” or “empirical data” rather than “empirical phylogeny”.  
  
GENERAL COMMENTS  
  
Sample size:  
  
It might be better to use the more specific term “size of the phylogeny” or “number of species” than sample size. As Maddison and Fitzjohn (2014) discuss in their paper, sample size is also reflected by the number of independent transitions. Of course, under the same model of trait evolution, as we increase the size of the phylogeny we expect a larger number of state transitions. However, for the model and for the simulations, the size of the phylogeny might not correctly translate to sample size. Sample size for this test is likely a combination of number of extant species, number of transitions, and number of nodes estimated under each of the categories (see Step 3).  
  
Pruning approach and causality  
  
After reading the manuscript I am not sure how the pruning approach works. My first concern is that this approach generates a distinct dataset from the original one. After pruning the phylogeny, the nodes used to perform ancestral estimates, for example, are not the same because some of the species were removed. This can create some problems. For example, comparing test results with and without the pruning approach might not make much sense, since one was made with phylogeny and trait distribution A and the other with phylogeny and trait distribution B.  
  
Pruning lineages can also influence the rates of trait evolution estimated for the continuous trait. First, the size of the phylogeny changed, meaning that there is less information to estimate rates of trait change. Second, if there is a causal association between the discrete and the continuous trait, then the species were removed non-randomly which can bias the rate estimates (likely to lower values). Combined, these two effect will change the ancestral estimates at the internal nodes for the continuous traits. I am not sure how this can result in a “more conservative test”.  
  
The pruning approach means that only the ancestral state will be represented. So there are no “producing nodes”. Step 4 of the methods (lines 159 to 161) mention that the subset is equal to the number of producing nodes. How is the null distribution produced in this case?  
  
Finally, I am not sure I fully understand the argument that the pruning test is a significant step towards separating correlation from causation. This is a very promising and important topic of the paper. I strongly suggest the authors to dedicate a section of the paper to explain in more details how this new test, and the pruning approach, improve our understanding of causation. I need more information to understand how this works.  
  
Together with the section I suggested above, the authors could show how much the results are improved if the pruning approach is used. It is feasible to simulate under a scenario in which the derived state is associated with lower values for the continuous trait: 1) make a stochastic character mapping of the binary states, 2) use a OU model with two optima (a lower optima associated with the derived state). Then compare the performance of the test with and without the pruning approach.

*We agree that our description of the methodology and utility of the pruning approach could be improved. We have reworked our description of this approach in lines XXXXX. Furthermore, we have added a simulation scenario that help to better illustrate the utility of the approach.*

*Scenario XXXX is a case where the continuous trait evolves via a two optima OU model. In this scenario the optima are determined by the state of the discrete character. Thus this simulation has a causative relationship where the discrete trait impacts the continuous (the opposite of what the Ancestral condition test is designed to detect). What our results illustrate is that like many comparative methods the ancestral condition test will return a significant result (power XX-XXX). However, unlike other approaches we can investigate the directionality of this relationship by using the pruning option we have developed.*

Continuous ancestral trait estimates are only performed at the nodes:  
  
This test is using information from the nodes for the continuous traits and information from the branches for the discrete traits. However, continuous traits also evolve along the branches. For example, imagine node 1 and node 2. Node 2 is one of the descendants from node 1. The branch connecting node 1 and 2 is a producing branch. If the binary trait transition happened close to node 1, then the value of the continuous trait at that moment is likely to be similar to the estimated value for node 1. On the other hand, if this transition happened later, closer to node 2, then the value of the continuous trait at the moment of the discrete transition is likely to be closer to the value estimated for node 2 than the one for node 1. Depending on the ancestral estimate for node 2, the position of the shift along the producing branch could lead to distinct conclusions of the association between traits.  
  
It is possible to estimate the value for the continuous trait along branches. For example, one can add a “pseudo-node” at the point immediately before the binary state transition and use the same computations of a standard ancestral state estimate to estimate the value for the continuous trait at that point in time. This would make it possible to control the time interval between the ancestral state estimate for the continuous trait and the transition of the binary trait. In the current version of the test, it seems that the branch lengths are not taken into account, meaning that it does not matter if the binary state transition happened early in the producing branch or not.  
  
Because the model is a single rate BM, it is possible take a weighted average between the ancestral and descendant node values weighted by the distance between the point and each of the nodes. Of course, this will become more complex under other model of continuous trait evolution.

*We agree this is a useful extension that should increase power. We have taken the approach of calculating the weighted average giving us the continuous value at the point of transition in the discrete character. Lines XXX*  
Model testing between bidirectional and unidirectional transitions  
  
The authors mention that the unidirectional case is used when there are strong external evidence for this case. However, one can also test the fit of a free Markov transition matrix against a constrained transition matrix. Model testing is the most used approach to choose among distinct transition matrices. Only in specific scenarios we have enough information to confidently use a unidirectional case and not perform any type of model testing. I suggest the authors to include an advice for users to test a bidirectional vs. a unidirectional Markov model using model comparison approaches.

*We have made the suggested addition lines XXXX*