Dear Editor,

We thank the editor and both reviewers for their generous assessments of StarBEAST2 and their constructive criticism. Since our first submission, our colleague Remco Bouckaert has helped us further improve the performance of StarBEAST2 by optimising its core routines and is now a co-author. In order to reflect the performance of the current version of StarBEAST2, we have rerun all analyses and compared StarBEAST2 directly with *BEAST. In this revision we analyse more loci (up to 100) with StarBEAST2, and now also include accuracy and performance results when using a strict clock.

While preparing our new results, we discovered that the choice of sequence type to be used with concatenation – either phased haplotypes, or unphased sequences with ambiguity codes for heterozygous sites – was partly responsible for the poor performance of concatenation when estimating per-branch substitution rates. The use of phased haplotypes improved estimates of substitution rates, although StarBEAST2 estimates were still more accurate. The concatenation performance and accuracy results in main text figures are now based on phased haplotypes for a fairer comparison. The unphased results are referred to in the main text and presented in Supplementary Material.

For each specific editor or reviewer comment (in blue text), we give our response below (in black text).

Sincerely,

The authors

Editors’ comments to the author:
I have now received two reviews of your article and, unfortunately, the reviews were mixed. Despite the mixed recommendations, both reviewers had significant concerns with the current version of your paper. While the reviewers agreed that the updated approach to implementing the relaxed clock is an improvement, I tend to agree with Reviewer 2 that you need to demonstrate this improvement through more extensive simulation and larger empirical data sets. I also agree with the convergence issues comment and would like to see your improved method demonstrated on a state-of-the-art data set, not a minimal data set. Given these extensive critiques, I'm forced to recommend rejection of your paper at this time, but would encourage you to resubmit if you think you can accommodate these concerns. The paper clearly fits into the framework of the journal and is very well written.

Regarding larger empirical data sets, we have now tested StarBEAST2 on the state of the art data set from Giarla and Esselstyn (2015), a study based on ultra-conserved element (UCE) sequences from Philippine shrews of the genus *Crocidura*. We show that StarBEAST2 can be easily used with subsets of 100 loci from this data set. This is in contrast to the original paper, which used subsets of 50 loci with *BEAST.

Regarding more extensive simulations, we have re-re-analysed the empirical *Pseudacris* data set, this time using all species (21 instead of 19) and all loci (26 instead of 22). The new simulation data set also uses 21 species, and its birth/death rates and locus lengths more closely match the empirical data set than before. In contrast to our original submission which analysed simulated subsets of 22 loci, we now analyse subsets of 26 and 52 loci using StarBEAST2.

The computational performance analysis of clock models, new operators and population size integration is now performed on the simulated data set in addition to *Pseudacris*.

Reviewer 1 comments to the Author
This paper describes a new version of *BEAST, entitled StarBEAST2, which offers an improvement in computational performance and branch-rate estimation. This is achieved by the inclusion of new MCMC proposals and the application of a relaxed clock to the species tree. StarBEAST2 presents a useful development and is likely to be used widely by the research community. I have a few mostly minor comments on the paper. Generally, the paper can be shortened and clarified in a few sections.

Page 2 line 53 column 1. There is no mention of binning-based approaches here, such as those developed by Tandy Warnow's group. These present a computationally quick hybrid between concatenation and multispecies coalescent methods. It would be useful to mention them here, as well as the criticisms of "concatalesence" methods raised by Gatesy and Springer (2013).

We have added the following paragraphs to the introduction:

"Another issue with summary methods is when the assumption of no recombination within loci is frequently violated because overly long loci are used (Gatesy and Springer, 2013). To resolve larger and deeper species trees using summary methods, longer and more informative loci may be required to infer more accurate gene trees. However the larger and deeper a tree, the more recombination events will have occurred. The use of longer loci and the higher incidence of recombination will both increase the risk of recombination occurring within loci, which has been dubbed the "recombination ratchet" (Springer and Gatesy, 2016).

As an alternative to increasing locus length, fully Bayesian MSC methods like *BEAST infer more accurate gene trees by sharing information between loci through the species tree (Szöllősi et al., 2015). In this way very accurate species trees can be estimated using only weakly informative loci, which may not be possible using MP-EST (Xu and Yang, 2016). Summary methods can use gene trees estimated by *BEAST for more accurate results and to avoid the recombination ratchet (Zimmermann et al., 2014), but the resulting species trees still cannot be used for molecular dating."

Page 3 line 14 column. The authors should clarify what they mean by "true divergence time" – the point at which gene flow ceases (reproductive isolation).

We have rephrased this as "Because some time is required for genes to coalesce~~nce~~ looking backwards from a speciation event, the expected molecular distance between two species is greater than if coalescent events occured simultaneously when gene flow ceases."

Page 3 lines 20-23 column 1. "Such overestimation of divergence times can result in dramatic inflation of estimated top branch lengths". This is a circular statement.

We agree with the reviewer and have removed this statement.

Pages 4 and 5. Most of the details of the new MCMC proposal can probably be moved to Supplementary Material, but the authors may wish to keep these in the main text.

As the new coordinated proposals provide much of the improvement in performance of StarBEAST2, and dramatically improve the performance of species tree relaxed clocks, we prefer to keep them in the main text for the benefit of other methods developers.

Page 6 line 44 column 2. "gene branch through all containing species branches" should be rephrased for clarity.

We have rephrased this as “along the entire length of a gene tree branch.”

Page 6 line 50 column 2. The example given here, of an early estimate of the rate of hominoid mitochondrial evolution, seems an unusual choice. Perhaps a more recent and prominent example would be useful here.

We now use a rate of 10^{-3} per site per million years, within the range observed for mammalian nuclear genomes from Phillips *et al.* (2009).

Page 7. The headings in Table 1 need explanation, perhaps in the form of footnotes.

Footnotes to explain the headings have been added to Table 1.

Page 9. The headings in Table 2 need explanation, perhaps in the form of footnotes.

Footnotes to explain the headings have been added to Table 2.

Page 10 lines 26 onwards, column 2. This problem would probably be less important when analysing large trees (with many branches).

We have added the following paragraph to our discussion of SPILS:

“Mendes and Hahn (2016) demonstrated that SPILS causes systematic bias when estimating branch lengths, and we show that this translates into systematic bias when estimating per-branch substitution rates. Because the bias is caused by incomplete lineage sorting which is a function of population sizes and branch lengths, there is no reason to expect that large trees with varying population sizes and branch lengths would be any less biased.”

Page 12 line 25 column 2. “Species tree relaxed clocks were slower than the gene tree relaxed clocks”. This statement is confusing because it could easily be misinterpreted as referring to the rate of evolution.

We have rephrased this as “The ESS per hour convergence of species tree relaxed clocks was lower than for gene tree relaxed clocks.”

Pages 14-17. The Discussion section does not add much to the paper, and mostly repeats what was already reported in the Results. I suggest merging the two sections and removing any redundancy.

We agree with the reviewer, and have merged the two sections.

Page 15. The axes are missing labels in Figure 6.

The missing labels have been found.

Page 16 line 23 onwards column 2. Much of this section is repetitive and should be deleted.

Relevant parts of the discussion have been merged with the results section.

Page 17 line 37 column 2. Clarify that only 1 haplotype was included for each species.

This has been rephrased as “We fixed the number of species at 5 and the number of haplotypes per species at 1.”

Page 17 line 39 column 2. Give the units for the birth and death rates.

We have clarified that both rates are per substitution.

Page 17 line 43 column 2. “Two gene trees were sampled within each species tree with mean clock rates of 0.5 and 2”. So this would generate a 2-locus data set? Perhaps I missed the description of the analysis of this data set.

We now clarify that this is a 2-locus data set.

Page 17 line 47 column 2. Define UCLN AND UCED. Why were these discretized into 100 bins?

The following paragraph has been added to New Approaches:

“The new species tree relaxed clock model is available in StarBEAST2. Branch rate models that can be used with a species tree relaxed clock currently include the well-established uncorrelated log-normal (UCLN) and uncorrelated exponential (UCED) models (Drummond *et al.*, 2006), as well as the newer random local clock model (Drummond and Suchard, 2010).”

The UCLN and UCED models as described in Drummond *et al.* and as implemented in BEAST 2 both use discretized distributions, and we somewhat arbitrarily used 100 bins to test mathematical correctness. For computational performance and accuracy analyses, we ~~now~~ use the default setting in BEAST which is to set the number of bins equal to the number of estimated rates.

Page 18 line 31 column 2. What is the rationale for sharing the kappa and gamma shape parameters across all loci? This seems very unconventional to me.

We now estimate kappa and alpha separately for each HKY+G locus across all analyses. For the four *Pseudacris* GTR+G loci we estimate the transition rates and alpha separately for each locus.

Page 18 line 40 column 2. Most readers will find it easier to interpret numbers in base 10 than in base 2.

We now include whole numbers for chain lengths and sampling frequencies.

Reviewer 2 comments to the Author

This paper describes some modifications to the implementation of species tree inference in the program *Beast. These are in two main areas: (1) the addition of a coordinated proposal of changes to the species tree and gene trees that is essentially similar to the *Beast proposal recently published by Jones (2016); (2) a modification of the way in which *Beast implements a relaxed clock -- rather than allowing independent rates on each branch of each gene tree as is currently done in *Beast (which makes little biological sense) the authors assign substitution rates to populations on the species tree and the gene tree branch rates are approximated as weighted average of the rates for populations that the gene tree branches span. This seems like a definite improvement.

The major weaknesses of the paper are that the tree proposals in part 1 have very little originality, and mainly implement existing methods. Moreover, they seem to have little effect in improving the performance of the program. The example data set they analyze has only 19 species and 22 loci. Similarly, the very limited simulation study that is done (96 replicates) only looks at the performance of *Beast2 with 22 loci and 19 species (36 sequences), mimicking the empirical dataset. More information is needed regarding the limitations of the program in terms of the number of species and loci. Any program that cannot analyze at least 50 loci in a reasonable time is not an improvement over existing methods. There are many published examples for which the previous version of *Beast fails to converge with even 50 loci (e.g., Giarla and Esselstyn, 2015). The authors should at least show that the new version works for these previous failures. Methods such as BPP have recently been used to successfully analyze datasets with over 900 loci so the authors are setting the bar far too low in their simulations. I would like to see evidence of convergence with at least 100 loci.

We now show that StarBEAST2 can be used with subsets of 100 loci from the empirical data set of Giarla and Esselstyne (2015). These StarBEAST2 runs usually converge within 200 hours, which we believe to be a reasonable time-frame. We also believe that the 13.1x to 13.8x improvement in computational performance over *BEAST on empirical data sets represents a ~~definite~~ improvement.

Another weakness of the paper is that the description of the modified coordinated proposal is inadequate to understand the details of the algorithm. The authors should provide a detailed figure identifying all the branches, bounds, etc, described in the text.

We now more formally describe the CoordinatedUniform proposal, and use the same nomenclature as Jones (2016), to enable this operator to be fully understood by other methods developers. We have also added an example of the operation with figures to Supplementary Materials.