nature portfolio

Peer Review File

Scaling up Bayesian population phylogenomics through virtual dimension reduction

Corresponding Author: Professor Ziheng Yang

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors describe an interesting approach to achieve improved performance for Bayesian inference under the multispecies coalescent model, and are clearly experts in this matter. The manuscript is well structured and sufficiently easy to read through; I list my comments below.

Line 31 : can this sentence be better connected to the previous one, i.e. explicitly state which of (i) and / or (ii) you're tackling with your new method? If you are tackling (i) but not (ii) then you should clearly mention this in the abstract (I don't mean that in any bad way).
Line 36: this leaves doubt as to whether mixing efficiency is expressed as effective sample size, or as effective sample size per time unit (the latter is the one that matters); please be explicit about this in the abstract.
Line 42: please avoid using terms such as 'Bayesian MCMC', as there's nothing Bayesian about MCMC (in this case: simply remove 'Bayesian').
Line 63 (left column): this could do with a citation.
Line 63 (left column): this does not change the serial nature of MCMC however, and you should point that out to the readers. What you are doing is more work in a single iteration / generation, which can indeed be achieved through parallelization and / or more complex transition kernels. But neither break / avoid / bypass / change the serial nature of MCMC, and it would be good to point this out to the readers to avoid giving the wrong impression.
Line 91: 'referred to', do you mean 'which you will refer to in this manuscript'? If not – and it's a term that has seen use before – this needs a citation.
Line 91 (right column): please number your figures in the order they are referred to in the text.
Lines 130-137 (right column): I appreciate the authors pointing out and clarifying this!
Line 174 (right column): but those uniform priors were improper, right? Perhaps best to explicitly state this?
Lines 242-243 (left column): While the authors clearly point to supplementary text, I do think that a bit more information in the main text is justified, i.e. one or two sentences – or a statement between parentheses – of which data dimensions are needed for the approximation to be good would benefit this section.
Line 192 (right column): same question as before, these are essentially improper priors that were used previously? Please clarify in the text.
Line 295: 'Furthermore,'
Line 281: 'is influenced'
Line 332: my first concern here is using only 8 baobab species, but I hope to see larger data sets (in terms of number of species) in the subsequent examples
Line 336: 'with a posterior of 100%': what does this mean? Or what are you trying to say here? I consider this to be very vague statement.
Line 354: even fewer species in this example?
Line 362: and only three species in the final example? If the authors do not analyse data sets that contain more species than the ones they have analysed here, they will need to make corresponding changes in some of the statements in their manuscript. Don't get me wrong: yes, these are big data sets but 'only' in 1 dimension (the number of loci, but not the number of sequences). For example, lines 45-46 should be rephrased to something like: 'the massive number of loci in modern-day data sets'; similar changes should be made to the manuscript where necessary.
Further, I would like the authors to add a paragraph in the introduction of the manuscript that states that they explicitly focus on data sets with a limited number of species that contain a massive number of loci, and explain why that's often a key focus

or use case (rather than data sets that contain dozens of species).
Line 331 (right column): how many replicate analyses were run per MCMC algorithm tested?
Line 385 (left column): 'multi-core machines', I have not yet seen any machine specifications (processor, memory, number of cores, …) so far; it's important to indicate that in the main text, and to also mention if you used different machines with different core counts and how the performance scales with available cores.
Line 386: so the most impressive reduction for the data set with the most species? I would urge the authors to include a data set of at least a few dozen species!
Line 404 (right column): but you only have a few sequences per locus, right? As I mentioned before, I think the matter of how good the approximation is deserves more attention in the main text.
Line 664: 5h18m for A2, then certainly a data anlalysis with more species should be feasible (you don't necessarily have to compare computational performance with the slower methods, just show that it's now possible to tackle data sets that have more species)!

(Remarks on code availability)
I lacked the time to thoroughly test the software myself, apologies.

Reviewer #2

(Remarks to the Author)
This paper implements novel improvements to the ambitious and useful BPP package. Perhaps it would help to make clearer to general readers where this updated BPP sits in relation to other packages. It is difficult to have a feel for how significant this advance is. Having not followed the gene-flow side of the MSC for a while, I can see that there have been some recent developments. Papers by G. Jones are cited, and looking at those papers I have an impression that these have motivated the broader approach of analytically integrating out thetas and migration parameters through the use of conjugate priors (while recognising the earlier work by Hey and Nielsen in this area). Jones's factorisation in equation 3 seems crucial to the approach. It looks as though novelty is in scale, and the tricks outlined in ll 187-248. The examples are impressive in terms of loci and numbers of taxa.

159-160 So this is because the lineage no longer exists because of branching? (determined by \taus?).

206 Spell out SPR.

Although there is a lot of description of mixing and mixing efficiency generally throughout the text, there seems no advice on appropriate ESS (and whether this actually conforms to what one might expect from independent runs...). Also, judging from Fig 4c the mixing efficiencies for different parameters are at quite different scales (e.g. \tau \varphi in all algorithms), so jointly one might assume (unless effectively independent in the posterior) the \taus and \varphi are a limiting step here.

In fact if one uses the approaches described from l. 289 presumably ESS is not necessarily a useful criterion (and even in simple MCMC it can be misleading if burn-in not well chosen) and better would be comparison from independent runs? Something along these lines might be convincing in the Supp section.

It may be helpful to have a table with the main parameters defined. For example, having missed the definition of \varphi in the context of MSC-I (line 268) it took a lot of searching to find out what this parameter is.

It looks as though the stopping criteria in Table 1 were just from fixed-length runs. Is there a comparison of the quality of inferences? I assume in principle one could obtain posteriors for the \thetas in A1 from the description starting at l. 289

In fact generally the plethora of models in Table S1 lead to a bit of confusion, which is not really handled in the Discussion. So, 4c (also issues below) suggests mixing performance of A1 and A2 are similar, with Table 1 suggesting A2 is faster than A1 - and this is the figure highlighted in the paper. But what are the times of all the other variants? and which one should a researcher use? In the last two species B5 seems to have quite high efficiencies for some parameters.

Possibly more detail is needed to explain what is going on in Fig 4b. It took quite a bit of re-reading to work out that the individual point and interval estimates are for different models.

Similarly 4c is quite hard. Remind readers that the text at the top of each figure is saying what the authors expect to be observed. Perhaps some guidance on which model is in which axis (not always obvious).

Possibly in a more general journal Fig 5 could go in the Supp. It does not do much for me, unless explained better.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
Review of Flouri et al: Scaling up Bayesian inference…

Fredrik Ronquist 2025-10-22

This paper introduces a set of improvements in the efficiency of MCMC inference for the Multispecies Coalescent (MSC) model with migration or introgression/hybridization (MSC-M and MSC-I models, respectively, with the acronyms used in the paper). Together, these improvements represent a huge leap forward in inference efficiency on these biologically very important problems. Therefore, I am supportive of publication of this manuscript in a journal that would reach a wide authorship. Landmark computational papers tend to be underrepresented in such journals, which potentially means that there will be unnecessary delays in the relevant research community discovering and taking advantage of these developments.

In general, I think the paper is solid and well written but it is rather rich in technical detail and light on accessible explanations of the nature of the improvements introduced. I think it would enhance the impact of the paper considerably if this could be addressed by the authors. Most importantly, I think the authors should introduce the major tricks they use through simple examples that are explained in detail and visualized, so that readers who are not familiar with the details of MCMC inference can get an intuitive understanding of why the tricks are so effective compared to the state of the art. The current introduction of the trick based on conditional distributions (Left column, L91-96) is too short. I misunderstood it the first time I read it (see below), despite being fairly experienced with respect to MCMC algorithms. The crucial difference between sampling from the conditional of one parameter, and keeping that sample when sampling other parameters (conditional Gibbs sampling), and the idea of resampling from (or simply updating) the appropriate conditional distributions _while_ sampling other parameters, needs to be emphasized and explained. Similarly, I think the idea of approximate conditionals could be explained better, potentially with reference to Fig. 8 (which needs an improved caption). I could not find a single reference in the text to this figure. Finally, I think it would be helpful to add an illustration demonstrating the efficiency gain of sampling from the expectation of the conditionals rather than from samples of them.

To make room for this expansion, explaining the inference improvements in a more intuitive way to readers unfamiliar with MCMC techniques or statistical inference in general, I think it is possible to move more of the technical details into the online supplemental information, without losing any of the necessary rigor. Aside from these general comments, I do have some minor suggestions for improvement, as follows.

Left L91-96. “[VDROP] proposes certain parameters from their conditional distribution (achieving the same mixing efficiency as analytical integration)”. I first understood this incorrectly, so I think this needs rephrasing or more explanation. Consider, e.g., a bivariate normal with strong correlation structure. Updating each one of the two variables separately from their conditional posteriors, or one with an MH step and the other one separately from its conditional posterior will clearly not have the same mixing efficiency as integrating one or both of them out. What does have the same mixing efficiency is sampling one of them using some MH step, while also resampling the other one from its conditional posterior (or simply updating the conditional posterior of the other variable) given the proposed new state of the first variable. Please clarify this. See general comments for expansion on this and related issues.

Right L86. “differences of the” should probably be “differences compared to the”

Right L104 (appr). “\tau and \theta are measured in the expected number of mutations per site”. This is a very unusual measure of population size (\theta). How does this work? I assume this requires some notion of the sequence variation in the population?

Righ L111 (appr) “for two species” should probably be “for the two species”

Right L118-119. Maybe it is unnecessary to point out this but the history also includes the location of the migration events on the tree.

Right L141: “Let G_j be part of” should probably be “Let G_j be the part of”

Right L238-243: “To allow parallelization…” is a critical paragraph and the approach could be described better.

Left L249ff: “A1(I-IG-int)” and “A2 (I-IG-g)” seem overly complicated names for the two algorithms, and it is difficult to understand how they were derived. Please simplify the names, or explain what the acronyms mean. The explanation follows on Right L335ff but it needs to be here, at least in a simplified form and a reference to the more detailed explanation.

Left L264ff: “the mixing steps…”. Difficult to follow, please rewrite.

Right L281: “c in influenced” should be “c is influenced”

Right L288: “Consider for instance…” I find the example relatively uninformative with respect to the gains one might expect in the types of problems considered here. I suggest deleting this paragraph.

Left L334. “The BPP analysis of the data under the MSC model inferred the species tree…”. In Figure 2, it is not clear how

the species tree is inferred; it appears that only species split times are changed, not the species tree topology. Please clarify how the algorithm samples over species tree topologies.

Right L313ff: This paragraph is entirely focused on mixing efficiency per generation or iteration of the MCMC algorithm. I think it would be good to indicate already here that the clock time and the computational resources needed are also important factors to consider in comparing the performance of different algorithms (as is also done in the discussion of results later in this section).

Right L340: “it” should be “int”

Right L363ff: It is not quite clear what algorithms are identical to the ones used in previous versions of BPP, and how the new algorithms compare to them. Please clarify (Is A6 the same as that used in BPP previously, or not? Similarly, which B algorithm if any is identical to the previous BPP algorithm).

Left L475: “events” should be “event”

Fig. 1. I suggest using the standard conventions of graphical models rather than this ad-hoc format. If so, then \tau, \theta and \omega should not be surrounded by a square; in typical graphical models, a square with rounded corners or other features indicating replication is used for a template or plate of replicated iid variables, but these three random variables are not replicated. Instead, these three variables should be free circles (independent random variables), each pointing independently to G_i. Furthermore, G_i and X_i should be on the same plate, not on separate plates, and X_i should be shaded to indicated that it is observed / clamped. Finally, the squares with priors (Gamma-Dirichlet prior etc) should be replaced by the parameters of the gamma-dirichlet etc priors inside simple squares, as they are constant nodes in the model. The distributions themselves are typically not included in the graphical model.

Fig. 2. It would be better if this figure used a different drawing style from Fig. 1, so that it is clear that it is an MCMC algorithm illustration and not a graphical model. It is unclear how species tree topologies are sampled; please clarify. Also, it is unclear here and elsewhere in this text whether hybridization/introgression events are assumed a priori, or whether the MSC-I algorithms sample over hybridization/introgression events. The Results section seems to indicate that those events are specified in the prior and not inferred, and the fact that sampling over them is not considered in this figure points in the same direction, but there needs to be more clarity on this point both in the text and in the figure.

Figs 4, 6, 7: I think it would be better to break the panels in the c) part of these figures into one set for the MSC-M model and one for the MSC-I model. Alternatively, introduce a vertical line separating the estimates from each of the two models. These are different models and it is not surprising if the estimates are different. Also, please comment on the few cases where estimates do seem to differ between algorithms generating estimates from the same model. There seems to be a few of those cases in Fig. 7, for instance. Could they be due to some of the approximations used or are there other likely explanations?

Fig. 8. Please explain in the caption what this figure is meant to illustrate; it is not clear from the current caption. I am not convinced that the figure is needed, I could not find a single reference to it from the text. With a better caption, there might be reasons to keep it (see general comments).

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I have carefully reviewed the changes the authors have made to their manuscript, along with their rebuttal letter. In their responses, the authors refer a few times to their replies to the Editor but I was unable to locate these, which was quite unfortunate. I must hence leave it to the Editor to carefully check those responses; I support publication of this revised manuscript in Nature Communications.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
I have read through the authors' response letter and revised manuscript and I can see that my comments have been well addressed. I think it is a significant and methodologically sound contribution to the field.

(Remarks on code availability)

RESPONSE TO REVIEWERS' COMMENTS
Reviewer #1 (Remarks to the Author):

The authors describe an interesting approach to achieve improved performance for Bayesian inference under the multispecies coalescent model, and are clearly experts in this matter. The manuscript is well structured and sufficiently easy to read through; I list my comments below.

Line 31 : can this sentence be better connected to the previous one, i.e. explicitly state which of (i) and / or (ii) you're tackling with your new method? If you are tackling (i) but not (ii) then you should clearly mention this in the abstract (I don't mean that in any bad way).

We are making improvements in both (i) MCMC mixing and (ii) parallelisation, so we have added (p.1)
"To deal with both issues,…"

Line 36: this leaves doubt as to whether mixing efficiency is expressed as effective sample size, or as effective sample size per time unit (the latter is the one that matters); please be explicit about this in the abstract.

The algorithms discussed in our study have similar running times, so the use of ESS and ESS per minute to compare algorithms will be equivalent. The more advanced algorithms involve more effort in coding but the running time is similar. We have included the following text (p.6 left column):
"unless stated otherwise explicitly, the algorithms tested in this study have similar running time per iteration."

Line 42: please avoid using terms such as 'Bayesian MCMC', as there's nothing Bayesian about MCMC (in this case: simply remove 'Bayesian').

Changed as suggested.

Line 63 (left column): this could do with a citation.

We have added a few citations.

Line 63 (left column): this does not change the serial nature of MCMC however, and you should point that out to the readers. What you are doing is more work in a single iteration / generation, which can indeed be achieved through parallelization and / or more complex transition kernels. But neither break / avoid / bypass / change the serial nature of MCMC, and it would be good to point this out to the readers to avoid giving the wrong impression.

It is rare for a major high-dimensional MCMC algorithm to have only one proposal changing all dimensions. More commonly multiple proposals each changing a subset of the parameters are used in each MCMC iteration/generation. Typically all those proposals have to be done sequentially. By relying on conditional independence, we are doing some of those proposals in parallel. If we consider the full MCMC iteration, it is true that the next iteration has to wait until the current iteration is completed, but within each iteration, we are bypassing or changing the serial nature of MCMC.
In response to this comment, we have added the following text (p.2 left column):
"Most high-dimensional MCMC algorithms use multiple proposal steps, each updating a subset of parameters, while collectively the steps ensure that every dimension of the state space is updated in each MCMC iteration.

While the MCMC iterations are sequential, proposals updating subsets of parameters within each MCMC iteration can be applied in parallel by relying on conditional independence."

Line 91: 'referred to', do you mean 'which you will refer to in this manuscript'? If not – and it's a term that has seen use before – this needs a citation.

We have changed "referred to" to "which we refer to".

Line 91 (right column): please number your figures in the order they are referred to in the text.

Changed as suggested.

Lines 130-137 (right column): I appreciate the authors pointing out and clarifying this!

Line 174 (right column): but those uniform priors were improper, right? Perhaps best to explicitly state this?

Proper uniform priors have been used in different versions of the IMa program. These uniform priors appear to create problems in comparisons of models. We have inserted the following text (p.4 left column):
"using uniform priors with user-specified upper bounds; see Thawornwattana et al. (2026) for a discussion of issues arising from the use of uniform priors in the IMa3 program (Hey et al., 2018)."

Lines 242-243 (left column): While the authors clearly point to supplementary text, I do think that a bit more information in the main text is justified, i.e. one or two sentences – or a statement between parentheses – of which data dimensions are needed for the approximation to be good would benefit this section.

The idea of approximate conditionals appears to be new and is potentially useful in other major MCMC applications. We have now expanded our summary in the main text, mentioning figures S1 and S2 (p.4 right column). The data size here is the number of coalescent events and increases with the number of loci and the number of sequences. We have the following (p.4):
"We evaluated the approximation under different scenarios (e.g., with the prior being consistent or conflicting with the data) and found that the approximation is good even with only 2 or 10 events (fig. S1). Both the gamma and inverse-gamma approximations converge to the correct conditional when the amount of data (i.e., the total number of coalescent events over all loci in that population) approaches infinity, so that the approximation is more accurate in larger datasets with more loci and more sequences."

Line 192 (right column): same question as before, these are essentially improper priors that were used previously? Please clarify in the text.

As before, proper uniform priors have been used in the IMa program. We have inserted the words "with user-specified upper bounds".

Line 295: 'Furthermore,'

Line 281: 'is influenced'

Fixed.

Line 332: my first concern here is using only 8 baobab species, but I hope to see larger data sets (in terms of number of species) in the subsequent examples

While it is worthwhile to improve the algorithms to handle large phylogenies, we note that the biological problem of gene flow exists on small phylogenies with only two species. The number of loci is the sample size in the inference problem, and it is important to implement algorithms that can deal with thousands of loci so that inference of gene flow becomes meaningful. We have added the following text to emphasized the importance of handing lots of loci (p.2):

"Indeed under the MSC model (with and without gene flow), sequence data at different loci are independently and identically distributed (i.i.d.), so that the number of loci is the sample size in the inference problem (Rannala and Yang, 2003; Jiao et al., 2021). Among many factors that affect the information content in the data (such as the number of loci, the number of sequences sampled per species, the sequence length, and the mutation rate), the number of loci is often the most important (e.g., Huang et al., 2020; Thawornwattana et al., 2025, 2026). For example, to estimate the rates of gene flow reliably, thousands of loci are often needed and it is important to develop algorithms that can accommodate many loci even if for a relatively small phylogeny."

Line 336: 'with a posterior of 100%': what does this mean? Or what are you trying to say here? I consider this to be very vague statement.

Previous studies found only weak support for estimated phylogenies. Here we found 100% support. We have edited the text to offer some possible explanations for the discrepancy (p.7):
"Previous analyses of genetic data found considerable uncertainty in the species phylogeny (Karimi et al., 2020; Wan et al., 2024). We inferred the species phylogeny under the MSC model using bpp (Yang and Rannala, 2014; Rannala and Yang, 2017). The maximum a posteriori (MAP) tree (fig. 5a) has a posterior of ~ 100%, possibly because the full likelihood method uses the information in the data more efficiently than summary methods used in previous studies."

Line 354: even fewer species in this example?

Please see our major comment #2 to the Editor.

Line 362: and only three species in the final example? If the authors do not analyse data sets that contain more species than the ones they have analysed here, they will need to make corresponding changes in some of the statements in their manuscript. Don't get me wrong: yes, these are big data sets but 'only' in 1 dimension (the number of loci, but not the number of sequences). For example, lines 45-46 should be rephrased to something like: 'the massive number of loci in modern-day data sets'; similar changes should be made to the manuscript where necessary.

As far as we know, the three datasets analysed in our paper are beyond the power of competing software programs including *beast, phylonet/MCMC-seq, G-PhoCS, etc. Bpp has been used to analyze larger datasets than those used here (e.g., 17,000 loci, each with 12 sequences on a phylogeny for 6 species in Flouri et al. 2023, table 1). In this study our objective is to evaluate MCMC algorithms rather than testing the limits of bpp. From the viewpoint of statistical inference, the number of loci is the sample size, so it is important to accommodate many loci. We have added some text to emphasize this. Please see our major comment #1 to the Editor above.

Further, I would like the authors to add a paragraph in the introduction of the manuscript that states that they explicitly focus on data sets with a limited number of species that contain a massive number of loci, and explain why that's often a key focus or use case (rather than data sets that contain dozens of species).

We have followed this suggestion and added a paragraph in the Introduction.

Line 331 (right column): how many replicate analyses were run per MCMC algorithm tested?

We did two replicate runs. The reliability of the runs is also clear from the nearly identical results (posterior means and CIs) between algorithms applied to the same data under the same model and priors in figures 5, 6, 7.

Line 385 (left column): 'multi-core machines', I have not yet seen any machine specifications (processor, memory, number of cores, …) so far; it's important to indicate that in the main text, and to also mention if you used different machines with different core counts and how the performance scales with available cores.

We have now explicitly mentioned the server used in our tests, "a Lenovo ThinkSystem SR850 server with 4x Intel Xeon Gold 6154 18C processors" (p.7). We used the same computer in our tests.

Line 386: so the most impressive reduction for the data set with the most species? I would urge the authors to include a data set of at least a few dozen species!

We have already included three empirical datasets, with realistic sizes. As explained above, the number of species is not the most important factor to influence the utility of our algorithms. The number of species in the three empirical datasets is 8, 6, and 3. We have seen consistent improvements of the new algorithms in the three datasets, so we see no reason to believe that similar improvements are not possible when data of 20 or 50 species are analyzed. The objective of our study is to develop MCMC algorithms with improved mixing, rather than finding out the computational limits of the current version of bpp in terms of the number of species.

Line 404 (right column): but you only have a few sequences per locus, right? As I mentioned before, I think the matter of how good the approximation is deserves more attention in the main text.

We have now added extra information for the 3 empirical datasets, such as the number of sequences per locus and the sequence length. For the baobab dataset, there are 9 species, 344 loci, 28-38 sequences per locus, and 759-9574 sites per sequence. The dataset is not small.
Here the data size for the approximation is the number of coalescent events and the total lineage-pair coalescent waiting time. We have added some explanatory text (p.8):
"(i.e., datasets with more loci and/or more sequences per species per locus leading to more coalescent events)"

Line 664: 5h18m for A2, then certainly a data anlalysis with more species should be feasible (you don't necessarily have to compare computational performance with the slower methods, just show that it's now possible to tackle data sets that have more species)!

We estimate that for simple models it should be feasible to run bpp on a species tree with 100 species. However, we have already included three datasets, and the paper includes many highly technical sections, so that a fourth dataset analyzed in differently from the other three would take space to describe. As we explained above, we have now added text to explain the importance of being able to handle many loci. Please see our comment above.

Reviewer #1 (Remarks on code availability):

I lacked the time to thoroughly test the software myself, apologies.

Reviewer #2 (Remarks to the Author):

This paper implements novel improvements to the ambitious and useful BPP package. Perhaps it would help to make clearer to general readers where this updated BPP sits in relation to other packages. It is difficult to have a feel for how significant this advance is. Having not followed the gene-flow side of the MSC for a while, I can see that there have been some recent developments. Papers by G. Jones are cited, and looking at those papers I have an impression that these have motivated the broader approach of analytically integrating out thetas and migration parameters through the use of conjugate priors (while recognising the earlier work by Hey and Nielsen in this area). Jones's factorisation in equation 3 seems crucial to the approach. It looks as though novelty is in scale, and the tricks outlined in ll 187-248. The examples are impressive in terms of loci and numbers of taxa.

Our Discussion section has a summary of advancements made in this paper. The factorisation is mostly due to Hey and Nielsen (2017), and Jones (2019) is independent or duplicated work with no correct software implementation.

Regarding where bpp sits among competing packages, we have added some text on p.9 left column, starting with
"Currently bpp is the only program that appears to implement the MSC-M model correctly …"

159-160 So this is because the lineage no longer exists because of branching? (determined by \taus?).

Yes. We have added the following text to explain when the indicator $I_{sjk} = 1$ (p.3):
"(that is, if both species *s* and *j* exist during time segment *k* and are permitted to exchange migrants in the model)"

206 Spell out SPR.

Changed as suggested.

Although there is a lot of description of mixing and mixing efficiency generally throughout the text, there seems no advice on appropriate ESS (and whether this actually conforms to what one might expect from independent runs...). Also, judging from Fig 4c the mixing efficiencies for different parameters are at quite different scales (e.g. \tau \varphi in all algorithms), so jointly one might assume (unless effectively independent in the posterior) the \taus and \varphi are a limiting step here.

Check.

In fact if one uses the approaches described from l. 289 presumably ESS is not necessarily a useful criterion (and even in simple MCMC it can be misleading if burn-in not well chosen) and better would be comparison from independent runs? Something along these lines might be convincing in the Supp section.

Check.

It may be helpful to have a table with the main parameters defined. For example, having missed the definition of \varphi in the context of MSC-I (line 268) it took a lot of searching to find out what this parameter is.

We have redrawn figure 1 to include the MSC-I model as well as the MSC-M model. We define parameters in both models in the figure legend.

It looks as though the stopping criteria in Table 1 were just from fixed-length runs. Is there a comparison of the quality of inferences? I assume in principle one could obtain posteriors for the \thetas in A1 from the description starting at l. 289

Yes, our MCMC runs are of fixed length. The different algorithms produce the same posterior so the 'quality' of inference should be the same. Their differences concern the computational cost; it takes longer for an inefficient algorithm to produce estimates at the same precision than an algorithm with improved mixing.

In fact generally the plethora of models in Table S1 lead to a bit of confusion, which is not really handled in the Discussion. So, 4c (also issues below) suggests mixing performance of A1 and A2 are similar, with Table 1 suggesting A2 is faster than A1 - and this is the figure highlighted in the paper. But what are the times of all the other variants? and which one should a researcher use? In the last two species B5 seems to have quite high efficiencies for some parameters.

Possibly more detail is needed to explain what is going on in Fig 4b. It took quite a bit of re-reading to work out that the individual point and interval estimates are for different models.

Similarly 4c is quite hard. Remind readers that the text at the top of each figure is saying what the authors expect to be observed. Perhaps some guidance on which model is in which axis (not always obvious).

The old figure 4 is now figure 5. We have edited the figure legend to be more informative, and the main text is now expanded to have a detailed description of one set of results (for the baobabs of figure 5).

Possibly in a more general journal Fig 5 could go in the Supp. It does not do much for me, unless explained better.

The old figure 5 is now figure 8. This contrasts the computational cost of algorithms A1 and A2, which is a major result of our paper. We have edited the figure legend to explain why the two algorithms have very different CPU usage when run on multi-threaded computers and why their running times are so different.

Reviewer #3 (Remarks to the Author):
Review of Flouri et al: Scaling up Bayesian inference…
Fredrik Ronquist 2025-10-22

This paper introduces a set of improvements in the efficiency of MCMC inference for the Multispecies Coalescent (MSC) model with migration or introgression/hybridization (MSC-M and MSC-I models, respectively, with the acronyms used in the paper). Together, these improvements represent a huge leap forward in inference efficiency on these biologically very important problems. Therefore, I am supportive of publication of this manuscript in a journal that would reach a wide authorship. Landmark computational papers tend to be underrepresented in such journals, which potentially means that there will be unnecessary delays in the relevant research community discovering and taking advantage of these developments.

In general, I think the paper is solid and well written but it is rather rich in technical detail and light on accessible explanations of the nature of the improvements introduced. I think it would enhance the impact of the paper considerably if this could be addressed by the authors. Most importantly, I think the authors should introduce the major tricks they use through simple examples that are explained in detail and visualized, so that readers who are not familiar with the details of MCMC inference can get an intuitive understanding of why the tricks are so effective compared to the state of the art. The current introduction of the trick based on conditional distributions (Left column, L91-96) is too short. I misunderstood it the first time I read it (see below), despite being fairly experienced with respect to MCMC algorithms. The crucial difference between sampling from the conditional of one parameter, and keeping that sample when sampling other parameters (conditional Gibbs sampling), and the idea of resampling from (or simply updating) the appropriate conditional distributions _while_ sampling other parameters, needs to be emphasized and explained. Similarly, I think the idea of approximate conditionals could be explained better, potentially with reference to Fig. 8 (which needs an improved caption). I could not find a single reference in the text to this figure. Finally, I think it would be helpful to add an illustration demonstrating the efficiency gain of sampling from the expectation of the conditionals rather than from samples of them.

We thank the reviewer for the encouraging comments and for the summary of our results.
We have followed the reviewer's suggestion and added a simple example with the bivariate Gaussian target to illustrate the algorithms. Please see our major comment #1 to the Editor above.

To make room for this expansion, explaining the inference improvements in a more intuitive way to readers unfamiliar with MCMC techniques or statistical inference in general, I think it is possible to move more of the technical details into the online supplemental information, without losing any of the necessary rigor. Aside from these general comments, I do have some minor suggestions for improvement, as follows.

Left L91-96. "[VDROP] proposes certain parameters from their conditional distribution (achieving the same mixing efficiency as analytical integration)". I first understood this incorrectly, so I think this needs rephrasing or more explanation. Consider, e.g., a bivariate normal with strong correlation structure. Updating each one of the two variables separately from their conditional posteriors, or one with an MH step and the other one separately from its conditional posterior will clearly not have the same mixing efficiency as integrating one or both of them out. What does have the same mixing efficiency is sampling one of them using some MH step, while also resampling the other one from its conditional posterior (or simply updating the conditional posterior of the other variable) given the proposed new state of the first variable. Please clarify this. See

general comments for expansion on this and related issues.

All of the reviewer's interpretations here are correct. We have now followed the reviewer's suggestion and added a section titled "Simple example for a bivariate Gaussian target to illustrate the algorithms", as well as a new figure 4, to illustrate the algorithms (p.5-6). Please see our major comment #1 to the Editor above.

Right L86. "differences of the" should probably be "differences compared to the"

Changed as suggested.

Right L104 (appr). "\tau and \theta are measured in the expected number of mutations per site". This is a very unusual measure of population size (\theta). How does this work? I assume this requires some notion of the sequence variation in the population?

The average coalescent waiting time is 2N so that the average sequence distance or heterozygosity is $\theta = 4N\mu$. Viewing the mutation rate μ as a scalar, the heterozygosity θ is also known as the population size parameter. This is a fundamental parameter in population genetics models as it among other things determines the coalescent rate.

Righ L111 (appr) "for two species" should probably be "for the two species"

We think "for two species" should be ok.

Right L118-119. Maybe it is unnecessary to point out this but the history also includes the location of the migration events on the tree.

Yes, the migration history at the locus includes the number, directions and timings of migration events. We have now mentioned this.

Right L141: "Let G_j be part of" should probably be "Let G_j be the part of"

Changed as suggested.

Right L238-243: "To allow parallelization…" is a critical paragraph and the approach could be described better.

We hope that the newly added example of the bivariate Gaussian target should make all our algorithms easy to understand.

Left L249ff: "A1(I-IG-int)" and "A2 (I-IG-g)" seem overly complicated names for the two algorithms, and it is difficult to understand how they were derived. Please simplify the names, or explain what the acronyms mean. The explanation follows on Right L335ff but it needs to be here, at least in a simplified form and a reference to the more detailed explanation.

This is lost during editing.
We have made sure that the abbreviations are properly explained when they are used for the first time.

Left L264ff: "the mixing steps…". Difficult to follow, please rewrite.

We have changed this to (p.5 right column):
"Step 5 is a scaling or mixing move that multiplies all species split times on the species tree (τ) and coalescent and migration times on all gene trees by the same scale factor.".

Right L281: "c in influenced" should be "c is influenced"

Fixed.

Right L288: “Consider for instance…” I find the example relatively uninformative with respect to the gains one might expect in the types of problems considered here. I suggest deleting this paragraph.

We considered this suggestion and decided to keep the brief summary in the main text. Our theory not only proposes a better estimator (u versus y) but also allows calculation of its performance gain in the reduced variance. The example is partly used to illustrate the calculation.

Left L334. “The BPP analysis of the data under the MSC model inferred the species tree…”. In Figure 2, it is not clear how the species tree is inferred; it appears that only species split times are changed, not the species tree topology. Please clarify how the algorithm samples over species tree topologies.

We have edited the text to read as follows:
“We inferred the species phylogeny under the MSC model using bpp (Yang and Rannala, 2014; Rannala and Yang, 2017). The maximum a posteriori (MAP) tree (fig. 4a) has a posterior of ~ 100%, possibly because the full likelihood method uses the information in the data more efficiently than summary methods used in previous studies.”
The cited papers (Yang and Rannala, 2014; Rannala and Yang, 2017) describe algorithms that search in the space of species phylogenies.

Right L313ff: This paragraph is entirely focused on mixing efficiency per generation or iteration of the MCMC algorithm. I think it would be good to indicate already here that the clock time and the computational resources needed are also important factors to consider in comparing the performance of different algorithms (as is also done in the discussion of results later in this section).

We have inserted the following text (p.6 left column):
“…unless stated otherwise explicitly, the algorithms tested in this study have similar running time per iteration.”

Right L340: “it” should be “int”

Right L363ff: It is not quite clear what algorithms are identical to the ones used in previous versions of BPP, and how the new algorithms compare to them. Please clarify (Is A6 the same as that used in BPP previously, or not? Similarly, which B algorithm if any is identical to the previous BPP algorithm).

Yes, we now mention that A6 (and A3 and B6) are equivalent to algorithms in the previous version of bpp (ver4.7). In tables S3-S5 we calculate the performance gain relative to the old algorithms.

Left L475: “events” should be “event”

Fixed.

Fig. 1. I suggest using the standard conventions of graphical models rather than this ad-hoc format. If so, then \tau, \theta and \omega should not be surrounded by a square; in typical graphical models, a square with rounded corners or other features indicating replication is used for a template or plate of replicated iid variables, but these three random variables are not replicated. Instead, these three variables should be free circles (independent random variables), each pointing independently to G_i. Furthermore, G_i and X_i should be on the same plate, not on separate plates, and X_i should be shaded to indicated that it is observed / clamped. Finally, the squares with priors (Gamma-Dirichlet prior etc) should be replaced by the parameters of the gamma-dirichlet etc priors inside simple squares, as they are constant nodes in the model. The distributions themselves are typically not included in the graphical model.

We have redrawn figure 2 (the old figure 1) to conform with the DAG syntax.

Fig. 2. It would be better if this figure used a different drawing style from Fig. 1, so that it is clear that it is an MCMC algorithm illustration and not a graphical model. It is unclear how species tree topologies are sampled; please clarify. Also, it is unclear here and elsewhere in this text whether hybridization/introgression events are

assumed a priori, or whether the MSC-I algorithms sample over hybridization/introgression events. The Results section seems to indicate that those events are specified in the prior and not inferred, and the fact that sampling over them is not considered in this figure points in the same direction, but there needs to be more clarity on this point both in the text and in the figure.

We have redrawn figure 3 (the old figure 2) so that it does not look like a DAG diagram.
Yes, the models of gene flow (MSC-I and MSC-M) are implemented as fully specified models.
We have algorithms that change the species tree topology under MSC with no gene flow, but don't have cross-model algorithms that change the species phylogeny or gene-flow events under models of gene flow.
In response to this comment and to reviewer #2's comment "where this updated BPP sits in relation to other packages"), We have included a brief discussion of the strengths and weaknesses of bpp, starting with (p.9) "Currently bpp is the only program that appears to implement the MSC-M model correctly... However, only within-model algorithms for estimating parameters such as the rates of gene flow are implemented in bpp, while cross-model MCMC moves for searching in the space of gene-flow models are lacking…"

Figs 4, 6, 7: I think it would be better to break the panels in the c) part of these figures into one set for the MSC-M model and one for the MSC-I model. Alternatively, introduce a vertical line separating the estimates from each of the two models. These are different models and it is not surprising if the estimates are different. Also, please comment on the few cases where estimates do seem to differ between algorithms generating estimates from the same model. There seems to be a few of those cases in Fig. 7, for instance. Could they be due to some of the approximations used or are there other likely explanations?

We have followed the reviewer's suggestion to insert some space and a thin vertical line separating results for the two models in figures 5, 6, 7. We have also edited the legend to figure 5 explain things more properly. Some of the differences in parameter estimates are due to the use of different priors (gamma versus inverse-gamma for θs) under the same model.

Fig. 8. Please explain in the caption what this figure is meant to illustrate; it is not clear from the current caption. I am not convinced that the figure is needed, I could not find a single reference to it from the text. With a better caption, there might be reasons to keep it (see general comments).

We have now deleted this figure. We have included a brief summary of figure S1 (p.4 right column), which is an expanded version of the old figure 8.