Dear Recommender,

Please find attached our revised manuscript entitled "Simultaneous Inference of Past Demography and Selection from the Ancestral Recombination Graph under the Beta Coalescent" by, Kevin Korfmann, Thibaut Sellinger, Fabian Freund, Matteo Fumagalli and Aurélien Tellier.

First we would like to both reviewers for reading and appreciating the revised version of the manuscript. We understood that the main and last remaining problem was the scaling discrepancy between the kingman coalescent and the implementation of the beta coalescent in msprime.

We found ourselves puzzled at first as the implementation of msprime was beyond the scope of our study but we however understood the concern and modified the manuscript to improve the clarity behind the implementation of the Beta coalescent in msprime.

Many thanks in advance,

On behalf of the authors,

Kevin Korfmann & Thibaut Sellinger

Revision round #2

Decision for round #2 : ***Revision needed***

Revision needed

Dear authors,

I have received feedback from the two reviewers, and you will see that they are generally satisfied with your revision. There is one remaining point from reviewer 2 (why a beta-coalescent with alpha=2 does not exactly converge to a standard coalescent) that needs further clarification. As also pointed out by reviewer 1, the beta-coalescent might not be as widespread knowledge as more classical models; therefore, I believe it is important to make its presentation as clear as possible. If you would be able to address this last point, I would then recommend the manuscript for PCi Evol Biol.

Best regards,

Julien Dutheil.

Reviewer 1 :

The preprint has improved significantly from the previous version in the presentation and communication. I would like to acknowledge the authors for addressing all major and most specific comments, evident in the main text and the inclusion of new figures. Regarding one of the major comments, it was exciting to see the newly added results of GNNcoal trained with true or inferred genealogies. While I maintain my concern about the reader-friendliness of presenting some data in tables instead of in figures, I recognize that this is not a critique of the scientific content. This matter, therefore, can be appropriately discussed in the correspondence between the authors and the recommender. Finally, I would like to congratulate the authors on the revised version of the preprint and I thank the recommender for inviting me to review this exciting manuscript.

Reply: We thank Reviewer 1 for their comments and their appreciation of the revised version of the manuscript.

Reviewer 2 :

“Most importantly, I do not understand, why the Beta-coalescent is not exactly transitioning to the Kingman coalescent for α = 2.”

Reply: We would like to thank the reviewer for the detailed inspection of the underlying model. The first point we would like to make is that the scaling indeed plays an important part as is evident in Figure 2 and recognized by the reviewer. This figure describes the evaluation of PSMC/MSMC on the msprime implementation of the Beta coalescent. Likewise both SMBC and GNNcoal have been evaluated on the msprime version of the Beta coalescent and any scalings introduced by msprime are either directly mathematically transferred in SMBC or learned implicitly through training on simulations by GNNcoal.

To find justification for the reason of implementing the scaling we reached out to Dr Jere Koskela, Reader in Statistics at Newcastle, who is involved with the implementation of the respective parts in msprime. In his reply he confirms that the Beta coalescent and its coalescent rates in the limit as alpha goes to 2 (and plugging it in the Delta-distribution), we obtain the Kingman coalescent. However, this view lacks any relation or notion of time scales, which is where the issue lies. The msprime paper indeed implements the Galton-Watson-process of the Schweinsberg, 2003 paper, which adds a notion of a time scale, whose scaling can be found either in the msprime documentation and literature mentioned below. Furthermore Dr Koskela, also highlights a discontinuous jump in timescale as alpha->2 and actually is equal to 2.

For completeness we attach the relevant part of Dr Koskelas’ kind reply in the following: “[...] If I just take a Beta-coalescent as an abstract mathematical object and send α to 2, I get the Kingman coalescent with no further caveats or complications (indeed, there is no notion of a timescale). If I specify that I'm working with the pre-limiting sequence of supercritical Galton-Watson-type population models in Schweinsberg's paper, then there is a notion of timescale and sending α to 2 affects it. Setting α >= 2 gives a timescale of C(α)N generations for a constant C(α) > 0 which depends on α but not N (Schweinsberg's Lemma 6), while 1 < α < 2 yields the timescale in the msprime BetaCoalescent documentation (Schweinsberg's Lemma 13). In fact, the 1 < α < 2 timescale collapses to zero as α -> 2 [...], so there is a discontinuous jump in the timescale from α -> 2 to α = 2, i.e. the limit and the timescale do not commute. [...]”

What does that mean for SMBC and GNNcoal?

As stated earlier, any time scales implemented in msprime are also inherited by our models (e.g. α is upper bounder by 1.99 in SMBC). Due to the discontinuity when moving from Beta coalescent to Kingman coalescent, studies are required to carefully evaluate the expected strength of the underlying sweepstakes of the model organism and choose the appropriate neutral model. This is especially crucial since the phase from α>1.9 up to 1.99 where *msprime* suffers from numerical instability issues, which are actively being addressed and improved currently by Dr Koskela and the authors of *msprime*. (see below reply to question 3).

Schweinsberg paper: https://www.sciencedirect.com/science/article/pii/S0304414903000280

Doesn’t this show that there is some scaling problem of the mutation rates in your simulation?

Reply: We checked our simulation scripts and our mutation rates are in line with how our msprime was designed. As explained above the discrepancy originates from the simulator implementation and not by our use of it. We have now introduced one sentence in the introduction (Line 61-67) and in the method section to clarify the issue with msprime (Line 256-261).

Minor points:

1.) The explicit formulas for the scaling-factor are incomplete: In the formula for the so-called "scaling constant" on Line 64, there appears a β, which has not been introduced or defined as a parameter.

Reply: We apologize for this confusion, this beta stands for the Beta function. We corrected the manuscript.

2.) The quotations after these formulas are unhelpful, at least to me. I took a look at all three papers (refs. 8, 55 and 56), and while I admit I didn't read them in all detail, I could not really find these formulas. Perhaps these formulas could be derived for the reader (with references) in a short Supplementary Chapter or a methods paragraph. They can then be taken out of the text in lines 62-64, actually, where they are a bit overwhelming I think

Reply: Once again we apologize for the confusion. The formulas can be explicitly found in the msprime documentation and we have therefore added the reference to the msprime manuscript where the beta coalescent was introduced (2022 in Genetics) as well as the article from Schweinsberg in 2003 where the events rates are derived. We also added a short description in our methods as well as the documentation of msprime in the data availability section to make it easier for the reader to find information specific to msprime.

3.) The authors' response about my critique of their figure 2 is partly convincing. I get that you want to make the point that indeed the population size inference gets wrong if the assumptions break down. But, coming back to my main point above, this point only comes across if you actually show that the discrepancy between expectation and fit actually vanishes for α → 2. I find it hard to believe that for α = 1.9, the violation of the Kingman-coalescent assumption is already so stark that the population size is mis-estimated by a factor 100, which is what I see in Figure 2a. To repeat myself: I think there is something wrong with that. What I would have expected from that figure is a fit which looks very good for, say, α = 1.99, perhaps marginally worse for α = 1.9, and then perhaps increasingly bad for lower values. Instead, what I see in your Figure 2 is a terrible fit in all four cases, with a discrepancy ranging from a factor 100 to 1000

Reply: We understand your point and hope to have addressed the scaling issue above.

However concerning the underlying point about scaling discrepancy due to biological factors (and not implementation) we agree with you. If the inferred alpha is greater than 1.9 (or even 1.8) we would simply assume the underlying model to be a Kingman coalescent and use eSMC2 (or msmc2). That is why the user is free to choose the scaling with SMBC (the Kingman coalescent one or the beta coalescent one resulting from the implementation of msprime), the output shape will not change, just its position on the y and x axis. The output from SMBC can also be scaled according to the user preference if they wish to introduce knowledge that SMBC does not have.

We now clarify the issue in the method part of the manuscript using the *msprime* manual as reference and beginning of the results part (Line 328-355).

Minor point: In Line 62 there is a typo, I think. It says Beta(2α,α), but I think it should be Beta(2 −α,α)

Reply: Thank you for spotting this issue. The minus was in the .tex but was not displayed in the pdf. We now fixed it.