Dear Prof. Dutheil and Reviewers,

In the following part, we have crafted replies to all the comments mentioned by yourself and the anonymous reviewers. We hope to have adequately answered the concerns or dealt with the suggestions for improvements. We appreciate the extensive time taken and detailed engagement with our work.

Kind regards,

Kevin Korfmann

Round #1

by Julien Yann Dutheil, 03 Oct 2023 18:25
Manuscript: **https://doi.org/10.1101/2022.09.28.508873** version 3

Revision needed

This manuscript by Korfmann and collaborators reports extensive developments of new genomic inference methods based on the beta-coalescent. This work extends classic models based on the Kingman coalescent, possibly bringing such approaches to a broader range of organisms, notably microbes. The manuscript represents a significant methodological advance, which comes in three ways:

1. A new inference model, extending the multiple sequentially Markov coalescent approach (MSMC) to account for multiple mergers.
2. A new graph neural network approach that can learn coalescence parameters from ancestral recombination graphs.
3. Approaches based on the newly introduced models to infer regions under selection along the genome.

The two reviewers highlight the innovative aspects of the work and its great application potential. However, both indicate that the presentation of the model and results should be improved. They provide detailed comments and suggestions that, I believe, will be useful to the authors to improve their manuscript.

Reply: We thank the recommender and the reviewers for acknowledging the novelty (and depth) of our approaches, and for their very thorough reviews and insightful suggestions. We truly appreciate the time spent on reading and understanding our work, and we use the suggestions to improve the manuscript.

I further highlight below some points that I think the authors should address:

1) The extensive mathematical developments require that their detailed exposure be provided as supplementary material. This exposure is, however, incomplete in several places, and some critical information is missing in the main text:

It is indicated that SMbetaC can be run on ARGs instead of sequences (e.g. l211). How does the inference work in such a case? That is, what are the hidden states? I could not find a description of this approach in the supplementary text, which only describes the standard model where genealogies are the hidden states.

Reply: We agree that this point was not well explained in the former version of the manuscript. The inference on ARG is performed as in Sellinger et al. 2021 (https://doi.org/10.1111/1755-0998.13416). As described in the manuscript, two different input formats can be given to the SMBetaC :

- 1 Sequence data through a file containing all segregating sites and their positions on the genome

-2 Tree sequence data (i.e. coalescent time/genealogies along the genome)

In both cases the underlying HMM is exactly the same and so are the hidden states. The hidden states are the coalescent time of the first coalescent event and the index of which individuals are coalescing (as now described in the appendix).

When ARG (or Tree sequence) is used as input, the tree sequence is directly processed (i.e. read through) to obtain the sequence of hidden states which are normally inferred from the sequence data through a forward-backward algorithm.

We improved the model description in the manuscript and appendix to clarify these points (line 144-151 in the main text and section 1.3 in appendix S1), We now state :

“All approaches can either use the ARG or sequence data as input. Giving ARG as input for MSMC and MSMC2 is enabled by a re-implementation included in the R package eSMC2 previously published in \cite{Sellinger_2021}. It is important to mention that there are no theoretical differences in the models whether sequence data or ARG is inputted (see \cite{Sellinger_2021} and Supplementary Text S1 for details). The difference is that in one case the hidden states are inferred from sequence data with a forward-backward algorithm and in the later the sequence of hidden states are directly built from reading the inputted ARG (skipping the forward-backward algorithm). ”

How are selection scans performed? Is the alpha parameter allowed to vary along the genome? How can it be inferred in a "local" manner?

Reply : No “true selection scans” are performed in our study. With this analysis, we sought to pinpoint how selection has a local effect on the genealogy potentially leading to multiple merger coalescent events. To model this phenomenon we allow alpha to vary along the genome through the use of a sliding window (10kbp in our study) for alpha (see figure 7 & 8). The alpha parameter is then inferred using a Baum-Welch algorithm restricted on the data from the window (as it would be when inferred globally). Our results, therefore show the local inference capability of alpha, with subsequent wider implications for future development of selection scan methods based on an underlying beta coalescence model.

I agree with reviewer 2 that a detailed assessment of the math in the supplementary material would require much time. However, a relatively simple and efficient check can be made for such complex models: simulating data under the exact inference model (that is, under the SMbetaC model). The maximum likelihood theorem stipulates that the parameter inference should be unbiased under such conditions. Simulating data under the "real" process, as the authors perform, is of greater practical importance. Still, simulations under the inference model offer insurance that the model is correctly implemented, and I encourage the authors to verify this.

Reply : We thank the recommender for this idea. Unfortunately due to time constraint and as our manuscript is already very dense, we have not performed such simulations under the inference model. The main issue is that these simulations are not implemented in msprime and we could not easily tweak msprime to obtain the SMbetaC output. We argue nonetheless that the SMbetaC model is locally (so for a small genomic region) very similar to the Beta coalescent with recombination, so our test of the performance on msprime-generated data should locally not be much biased (as demonstrated by the accuracy of our method). We keep the idea of this test to be implemented in our next SMC study.

2) I did not understand why the author looked at the "classic" LD, and as pointed out by reviewer 2, the discussion on the Markovian hypothesis is unclear, if not inaccurate. First, the Markovian assumption is also violated under the Kingman coalescent; this is not specific to the beta-coalescent. Furthermore, while the SMC captures some kind of LD (so-called topological LD), how it relates to the more classic notion of LD based on haplotype frequencies is not straightforward. As the manuscript is already dense, I suggest removing this part and focusing on the topological LD (the transition matrix).

Reply : We apologize for the confusion and that our choice in the presented graphic did not support our argumentation. However we disagree with reviewer 2 on the importance of the LD in our study (because it is this specific LD shape under the Beta Coalescence that mainly explains the better performance of the GNN when compared to the SMC).

To improve clarity, we now present a new plot (figure 3) that better highlights the differences between the LD under the Kingman coalescent and the Beta coalescent. As you can see on the plot, on average under the Kingman coalescence the LD is generally monotonic and decreasing with distance. However under the Beta coalescence, the LD can be “Saw-tooth” shaped. And those local spikes in LD under the Beta-coalescent are what we meant by “violation of the Markovian approximation”. At the same time multiple merger coalescent events can affect multiple sections of the genome, which is unlikely under the Kingman coalescent. This effect cannot be captured by a Markov Process (as depicted in Figure S3).

But we agree with the reviewer and recommender that the hypothesis is always violated even under Kigman. Thus we rephrase this section of the manuscript for accuracy and now state that “LD monotonously decreases in average with distance under the Kingman coalescent suggesting the hypothesis of Markovian change in genealogy to be a fair approximation of the genealogical process in that case \cite{Wilton_2015}."

Minor:

l29: I am not sure how common knowledge the "survivorship types" are. Maybe a reference could be added?

Reply: We added a citation to highlight the classic ecological concept of survivorship curves in Line 25: "[...](so-called type I and II survivorship in ecology, see survivorship curves \textit{e.g.} by \citep{demetrius_adaptive_1978})[...]"

l50: Haploid organisms: could a few sentences be added to indicate the main differences with a diploid model? It is discussed in the "Discussion" part, but I feel some information for the non-expert would be helpful here.

Reply: We added Line 52: "In the polyploid case, where each parent contributes multiple genomes, the SMC formulations of putative intra and inter individual coalescence events would need to be carefully modelled, since this effect would lead to smaller coalescence probabilities and a shifted predictable demographic time-window into the past."

l162 "All SMC approaches used in this manuscript are found in the R package eSMC2.": as I understand this sentence, the authors have reimplemented the MSMC model. Is that so? l263 (also l283), the authors say that they used MSMC and MSMC2. If the authors do not mean the original software, they should state it clearly. In such a case, they should also indicate how the implementation differs from the original in terms of parametrisation, estimation procedure, etc.

Reply: We indeed used reimplementations of MSMC and MSMC2 algorithms previously described in Sellinger et al. 2021. Therefore, we now redirect the readers to that manuscript for more details. However, briefly, the main differences in the implementations are the possibility to receive ARGs as input and in some default parameters (e.g. window size and/or number of hidden states). The implementations we use tend to be computationally slower than the original ones when sequence data is inputted. The optimization algorithm is also different as in the original implementation of MSMC and MSMC2 the parameters cannot be bounded.

The manuscript was modified accordingly :

"Providing the ARG as input for MSMC and MSMC2 is enabled by a re-implementation included in the R package eSMC2 \cite{Sellinger_2021}. It is important to mention that there are no theoretical differences in the models whether sequence data or ARG is inputted (see \cite{Sellinger_2021} and Supplementary Text S1 for details). The only difference is that in one case the hidden states are inferred from sequence data with a forward-backward algorithm and in the later the sequence of hidden states are directly built from the inputted ARG."

Fig1: I agree with reviewer 1 that Figure 1 is not informative. It isn't easy to guess what the various graphs, dots, squares and curves represent.

Reply: We acknowledge that the figure by itself is not enough to understand the complete methodology and purely serves to provide an intuition behind how the genealogies are compressed into our demography estimates. We modified the text itself to directly cite the Figure from which this Figure has been inspired by (Fig. 1 of https://arxiv.org/pdf/1806.08804.pdf) and adjusted the text to make it more readable.

We also notice that despite the prevalence of neural networks in the field of population genetics since 2016, it continues to be a challenge to find the threshold between how much information needs to be explained and how much information can be regarded as given. For instance, some open questions on the need to explain terminologies like convolution, embedding, and dense networks remain. However, information can be found in the cited deep learning review article.

To conclude, modifying the figure itself would break the direct association between the figure of the original paper and we hope that our modification of the text and adding the citation sufficiently clarifies our message.

l171: number OF coalescence trees. As THE batch size is fixed.

Reply: Fixed. Thank you for spotting it.

l212: can msprime simulate selection? How exactly?

Reply: msprime can simulate an approximative version of a selective sweep (see https://tskit.dev/msprime/docs/latest/api.html#msprime.SweepGenicSelection).

For example:

```
model_kingman = [msprime.SweepGenicSelection(position=(L/2),
start_frequency=(1/Ne),end_frequency=0.99,s=s,dt=(1/(40*Ne))),
msprime.StandardCoalescent()]
```

And then adding these objects to the model parameter of the model parameter of the msprime.sim_ancestry function.

We modified the text to mention this functionality.

l291: I agree with reviewer 2's comment and suggestion on the scaling. Furthermore, some references should be added.

Reply: We fixed the scaling issue and every plot is now using the same scaling.

Fig2: I did not get why PSMC is mentioned here (and, unless I am mistaken, only here).

Reply: Thank you for spotting this typos. We meant PSMC' , which is MSMC but on sample size 2 (i.e. 2 haploid sequences). MSMC2 stems from PSMC' while PSMC' stems from the SMC' unlike the original PSMC from 2011 stemming from the SMC. We added PSMC' to strongly state that only 2 sequences are simultaneously analyzed (i.e. a PSMC’ type of algorithm such as in eSMC or MSMC2).

l328: It does not seem to me that the GNNcoal approach exhibits "high accuracy" in the case where alpha = 1.3

Reply: We agree that for alpha=1.3 the results cannot be described as highly accurate. We modified the section accordingly: “[...] and high accuracy from 1.9 to 1.5 with a noticeable drop in accuracy for 1.3. The latter results can be caused by the ever increasing sparsity towards the lower spectrum of the β-coalescent.“

Furthermore, we added Supplementary Figure S18 and S19 to show the decrease in the number of genealogies and ancestral nodes mentioned in the introduction.

FigS4-S7: the figure titles should state the demographic scenario (currently, all figures share the same title). Furthermore, in the case of population expansion/collapse, the population size change falls out of the resolution of the inference model so that it only infers constant population sizes in several cases. For alpha = 1.7 and 1.9, a more ancient size shift should be considered (Figures S5-7).

Reply: Thank you for pointing out the title issue. We now replaced it by adding the part “when population undergoes “*insert scenario*” to each title. We chose to analyze each data set under the same scenario and unfortunately it is impossible to find one fitting for all alpha cases. However, we believe that sufficient analyses have been conducted to conclude on the performance of our methods.

Fig4: what are the light grey lines?

Reply: The light lines show individual replicates of the SMβC method. We added this explanation to the legend.

Fig5: I think this figure might be easier to read (notably to compare the panels) if the y-axis represented (relative) errors ((estimated value - true value)/(true value)

Reply: We believe the main purpose of this Figure is to show the overall difference in prediction capability of the methods, highlighting the large variance of the SMBC method. Here, we do want to keep the absolute errors, since we want the reader to see the absolute deviation in alpha. For comparing exact values (and thus indirectly relative values) we provided Supplementary Table S1. We added “For exact values and standard deviations of the respective experiment see Supplementary Table S1.” to the legend of the respective Figure.

l363: it seems that the wrong figures are mentioned here.

Reply: We believe these are the correct Figures to be references here as they represent the inference under Kingman coalescence.

l463: In practice, we will never get the true ARG, so this does not constitute an advantage of the GNNcoal. Maybe this should be rephrased as a perspective, like "as ARG inference method improve, GNN models will offer a promising alternative to..."

Reply: It is true that we won’t get the true ARG, but we may be able to use a larger number of (inferred) coalescent times due to larger sample sizes. This is what we meant by stating to “leverage information on a whole genealogy”. Therefore scaling the GNN with sample size constitutes an advantage over SMBC as deriving the mathematics quickly becomes infeasible. The sentence was modified based on the suggestion by the recommender.

Reviewer 1: Comments on “Simultaneous Inference of Past Demography and Selection from the Ancestral Recombination Graph under the Beta Coalescent”

Main

In this manuscript, Korfmann, Sellinger et al. present two novel methods to model and study the genetic ancestries of species in which a single individual can produce a large number of offspring, which is biologically plausible but violates the assumption of the standard coalescent. Both methods are based on the β-coalescent, which models genetic ancestries with multiple merger events, but they tackle the problem with two different approaches: while the first method, SMβC, extends the sequentially Markovian coalescent (SMC), the second method, GNNcoal, is a graph neural network (GNN) trained on genealogies simulated under the β-coalescent. The authors first tested the performance of these methods by inferring various demography scenarios and the multiple merger parameter values using the true genealogies simulated under the β-coalescent model, then using mutations reflecting more realistic application. Second, the authors investigated whether GNNcoal trained with simulations under various scenarios can distinguish different factors underlying multiple merger events, namely skewed offspring distribution and selection. Finally, they examined whether the two methods can be used to identify the target of selection along the genome while simultaneously inferring demographic history. Overall, both methods, especially GNNcoal, performs well if the true genealogies are known and the multiple merger parameter is not too extreme. While SMβC is more robust to inferred genelogies and use of observed mutations, the performance of GNNcoal depends on the accuracy of genealogies.

I have two main comments.

1. The authors report promising performance of GNNcoal to distinguish Kingman coalescent without selection, Kingman coalescent with selection, and β coalescent. However, only the results based on true genealogies are reported. As this type of analysis has a huge potential to help decide downstream analyses (methods based on Kingman coalescent or multiple meregr coalescent) in practice where true genealogies are not available, results of the same analysis using inferred genealogies will be very informative to empirical biologists. This is related to comments on L395-407 and Fig. 6.

Reply: Thank you for this comment. We agree that having prior information, such as knowing the underlying model that most closely captures the data, would be of great help. Following your suggestion, we extended the analysis to training on inferred genealogies and re-classifying the individual models. An important insight here was that, when we want to distinguish selection from multiple merger, it proves important to train on inferred trees as opposed to training on exact simulated trees. As the uncertainty in the tree inference looks similar to multiple merger, the model needs to be exposed to this uncertainty during training to be able to distinguish it, which in fact proved to be successful (see Fig 6, C). Notably, not bias the model with different timescales from kingman or beta coalescent we either normalized the simulations for the exact training or we trained only on trees obtained by tsinfer (no dating), which means training only on topologies for subfigure C).

2. The authors discuss long range effects of multiple merger events, but such effects are not directly shown. They should consider investigating the true genealogies to discuss actual multiple merger events and their effects on LD. This is related to individual comments to L304-307, L308-322, L481-483.

Reply: As mentioned above in the reply to the recommender, we understand that our analysis of LD needed to be improved. Hence we improve Figure 3 to better show the long range effects (i.e. potential local increase of LD over long distances) of multiple merger. We hope that the new figure, legend and explanation in the text do clarify and demonstrate the effect of multiple merger events on LD.

Besides the scientific comments, the presentation/communication in the manuscript should be improved for better accessibility to general evolutionary biologists. Examples include:

1. Fig. 1 could be improved to deliver important message to evolutionary biologists, who are not necessarily familiar with machine learning, and include model description of SMβC in addition to GNNcoal;

Reply: We generally agree with this comment. However, we believe that SMC has been in the center of population genetic communities attention for inference purposes for many years, leading review articles and countless papers applying the method. Further, we would like to redirect the reader to the extensive supplementary material of SMBC.

2. In some places it is unclear whether true genealogies or inferred genealogies were used (e.g. L408-410, L422-423);

Reply: Thank you for this comment. We clarify at various places by replacing “data” with “exact genealogies”.

3. Some supplementary figures contain only one demography scenario for the entire analysis (e.g. Figs S8, S14).

Reply: We agree with the point raised. However, ARGweaver analyses proved to be computationally too expensive (S14). For instance, completing the ARG inference for the 4 scenarios provided took about one month to complete. For this reason and to avoid unnecessary but costly computational analyses, we decided to run only the most difficult sawtooth scenario. As the analyses using ARGweaver are not central to our study, we aim here to make the case of the feasibility of graph neural networks as future research direction in the field.

Similarly, for figureS8, we believe to have shown, with many demography examples, that this early GNN model is capable of inferring demography, and we believe that with future iterations inference can be improved. While it would be interesting to explore the lower data limits in more depth, we prefer to leave this endeavor for future studies.

Individual comments are listed below.
Comments on text
• General: Supplementary figures and tables are referred incorrectly.
– e.g. “Table 1 in S1” should read “Table S1”.
Reply: We adjusted it to be consistent.

• General: Many paragraphs start from methods without stating the objective/hypothesis/expectation (e.g. L379; L387; L396; L408; L422).
Reply: We now clarify these instances. For 397, we added “[...] with the objective of evaluating the performance on ARG reconstructed data from ARGweaver [...]”. For 387 we write: “We then evaluate SMβC on simulated sequence data to compare the necessity of reconstructing the ARG for the SMC method [...]”. For 396 we added: “[...] to assess our methods' capacity to distinguish between them.” For 422: “[...] to test our methods' simultaneous inference capabilities”. For 408, we append a sentence stating the hypothesis (see answer below)

• Abstract: “we are able to distinguish skewed offspring distribution from selection while simultaneously inferring the past variation of population size”
– In Figure 8, demography and selection are inferred but skewed offspring distribution is not explicitly reported.
Reply: The abstract refers to results from Figures 6 and 8.

• L23-25: Please provide references for different types of survivorship.
Reply: We added a reference in Line 25.

• L220-222: It would be helpful if the authors mentioned what the four values of α mean in terms of the genealogies (under the simplest demography model) by giving some numbers. I can tell that genealogies under α = 1.3 have more multiple merger events than those under 1.7, but I cannot imagine how common such events are under scenarios with these α values.
Reply: That’s a great question. Our methods essentially have to deal with an ever increasing sparsity gradient. And not only does the number of genealogies decrease drastically, but the multiple merger signatures become more and more prevalent. To answer the question directly, we now provide two outputs. As a way to provide an intuition on the multiple merger events, we added Figure S18 and S19 to show the decrease in available genealogies and ancestors with various values of alpha. We hope these help to to derive some insight into the little amount of data that is left for inference under the Beta coalescence for various alpha

values. We also computed from msprime the number of merger of size 1 to 10 (for sample size n=10) taken from 1,000 simulations (the graph is below and now added as Figure in the Appendix (S20)).

• L232-233: Please state why mutation and recombination rates were set differently across scenarios with different α.
Reply: In 231 we write: ". In order to obtain a sufficient number of SNPs for inference".

• L235-236: The authors might consider giving the two GNNs different names to avoid confusion by readers. (Or, are they considered the same GNN?)
Reply: They are considered to be the same GNN, with only one change in the last layer to output 60 parameters for the demography, or 1 parameter for the alpha value.

• L260&262: I am not familiar to these range notations. Does the two notations (using "-" first and "," second between two values) mean something different? Could it simply be something like 1.75 ≤ α < 2.00?
Reply: These range notations are identical, and we modified them according to the suggestion.

• L288-290: "due to the scaling discrepancy between the Kingman and β-coalescent". This is based not on Fig. 2 but on Fig S1.
Reply: The uncorrected scaling issue can be seen in Fig. 2, as demonstrated by the sawtooth scenario.

• L292-294: Was the scaling done upon the MSMC2 output, or was it done by modifying the MSMC2 algorithm?
Reply: We added "after the inference" to make it clear that scaling was done upon the output.

• L300: "Linkage" should read "Linkage disequilibrium"
Reply: We changed it l313

• L301-302: Please give some number representing higher LD. I cannot tell this well by visually comparing panels of Fig. 3 alone.
Reply: Figure 3 has been redone. We hope it is now visually clearer.

• L302-303: To show "a higher variance in LD", the window size should be consistent across Fig. 3 A-C.
Reply: We modified the window to contain on average the same amount of SNPs across the chosen scenarios. We also believe the new Figure 3 is now much better in representing the variance of LD.

• L304-307: I suggest the authors that they check whether "the long range effect of strong multiple merger events" really exists directly in the true genealogies.

Reply: We replotted the LD decay in presence of multiple merger in Figure 3. We hope the observed variability is informative enough to reflect the long range effect of multiple merger events (see also reply to the recommender above).

- L308-322– In this paragraph, the objective and conclusion are not consistent. The objective seems to concern the biological effect of multiple merger events on LD, while the end of the paragraph focuses on inference using SMC.

Reply: We are sorry for the confusion, though of course both topics are linked. As its name suggests, SMC approaches rely on the Markovian assumption to model and approximate the genealogy distribution along the genome. Decreasing LD suggests that when going along the genome different (and previously unobserved) genealogies should be observed (which explain the high accuracy of the SMC approximation in such cases). We clarify the writing there.

– I speculate that Figs. S2 and S3 are meant to discuss the effect of multiple merger events on LD, but they are not communicated well enough.

Reply: We apologize for the confusion. Figure 3 describes the consequences of multiple merger events on LD. Supplementary Figure 2 and 3 show how multiple merger events affect the observed transition matrix of SMC approaches. The transition matrix has been defined in the manuscript. Note that the transition matrix is what is used to infer model parameters, see Sellinger et al 2021 for more details.

Under the Kingman coalescent (Figure S2), no structural differences are observed between the expected transition (the model one) and the observed transition matrix. Hence under the Kingman coalescent the SMC is a relatively accurate process to model the genealogy distribution along the genome.

However, in Supplementary Figure 3 we observe structural differences between the two transition matrices. The observed dark blue lines are an over-representation of transition events at specific time points corresponding to multiple merger events. This over representation of transition events suggests the incapability of the SMC to correctly model genealogies along the genomes in presence of strong multiple merger events.

This results from the property of multiple merger events (i.e. the same coalescent events) to affect all sites on the genomes (explaining the LD spiked in Figure 3 and why the coalescent process cannot be markovian along the genome), and thus explaining the recurrent transition to hidden state when multiple merger events occurred (additionally explaining the high variation in inferred population size parameter). This is what is meant by “violation of the Markovian assumption”.

Yet, the probability under kingman for a coalescent event to affect all positions of the genome is negligible (explaining why the SMC approximates quite well the genealogy distribution under the Kingman coalescent).

We reshaped this section for clarity line 348 onwards : “However, under the β-coalescent (with $\alpha=1.3$) we observe significant differences between observed and predicted transition events at times points where multiple merger events occur (Figure

S3). More precisely we observed transitions at specific time points (corresponding to multiple merger events) occurring much more frequently than what is predicted by the model (dark blue lines). This plot thus shows that multiple merger events do not affect the genealogy at every time point (as modeled by the SMβC) and that multiple merger events are over represented in the distribution of transitions events due to the long range effects of multiple merger events (\textit{i.e.} many positions of the genome contain the same information, an effect which cannot be captured by the SMC approximation). This plot thus unveils the discrepancy between the expectation from the SMC (\textit{i.e.} approximating the distribution of genealogies along the genome by a Markov chain) and the actual effect of multiple merger events on the genealogy distribution along the genome."

– In addition, I think studying the transition matrix which deals with two neighbouring genealogies is not enough because 1, it does not directly show multiple merger events,

Reply: We are not sure we fully agree with the reviewer on that point, as we now described in the appendix dedicated to the SMBC.

and 2, it does not show correlation between "coalescent trees which are located at different places in the genome, and expected to be unlinked from one another" (L89-91). This is related to my comment on L304-307.

Reply: All SMC methods result from the Markovian assumption which is made. Hence the transition matrix is by definition the distribution of the genealogy at a position on the genome given the genealogy at the previous position (i.e. between 2 neighboring sites). At present, studying anything else in addition to this matrix would be beyond the scope of the SMC and this study.

• L331-332:
– "both approaches seem to recover fairly well the true α value (Figure 4..)":
Figure 4 does not show inferred α.
Reply: Thank you for seeing this. It should have been Figure 5 that was mentioned here.

– I suspect Figure 5 shows it but it is not referred to in the main text.
Reply: Thanks for pointing this issue out. We have fixed it. We indeed meant figure 5.

– Is Figure 5 the exact same as Table S1? If so then Table S1 is not necessary.
Reply: Indeed, Figure 5 shows the same values. In Table S1 we also provide other additional information and results.

• L333-334: Which figure/panel is explained here?
Reply: Figure 4. We added it.

• L340-343:
– If I understand correctly, the operation of increasing mutations and recombination rates by 50 folds is equivalent to using a 50x larger genome. Please make it clearer, or the purpose is unclear.

Reply: In the presence of multiple merger events, branch lengths (in generations) are on average smaller when compared to Kingman. In fact they can be so small that no mutation or recombination events may occur. As our (and most other) approaches rely on mutations and recombination for inferences we increased their respective rates to improve their chances of occurring, not to create a larger genome. We chose to not increase population size due to the non-linear effect of alpha on the Ne. We now modify the sentence accordingly to clarify this point by adding: “Since branch lengths (in generations) are on average smaller in the presence of multiple merger when compared to a Kingman coalescent, we choose to increase the rates as opposed to increasing the genome lengths, which does not affect the branch lengths (but increases the number of genealogies).”

– Please make clear whether the inferences were based on true genealogies or mutations. If the former, why are mutations necessary?

Reply: Yes the inferences are based on the ARG. As described in the appendix dedicated to SMBetaC, the hidden states are in units of coalescent time. Hence the branch lengths (in generations) need to be transformed into coalescent times which require a mutation rate and segregating sites. Although this could be theoretically calculated from the branch length in generations and the mutation rate alone, it is easier to directly use the outputted mutations.

• L345-346: I assume that this statement is based on comparison between Table S2 and Table S1 (i.e. Figure 5, correct?), but by looking at the tables I cannot tell if Table S2 has better accuracy than Table S1. Please plot Table S2 so they are visually comparable.

Reply: We apologize for the difficulty to read the two tables. We understand it can be confusing to compare both cases from the values. Yet we believe adding the suggested plot would not increase clarity or readability. Hence to facilitate comparison between the two data tables we added further description to the main text line 394 :
“Overall our results show that SMBetaC can recover alpha with higher accuracy when more data is available. To be more precise when M=3, the overall average inferred alpha values improve from 1.6, 1.53 and 1.42 (Table S1) to 1.64 , 1.49 and 1.36 (for data simulated respectively under alpha = 1.7,\alpha = 1.5 and \alpha = 1.3). Yet When M=4 a gain in accuracy is only observed for alpha = 1.5 and alpha = 1.3. Indeed, the overall average inferred alpha values changed from 1.60, 1.54 and 1.47 (Table S1) to 1.58 , 1.47 and 1.39 (for data simulated respectively under alpha = 1.7,alpha = 1.5 and alpha = 1.3)”

• L353: “Figures S4 to Figure S7”. Figure S8?

Reply: The comment refers to the SMBC part and was therefore correct.

• L367-369: Please elaborate what exactly is meant by “scaling discrepancy between the Beta and Kingman coalescence” in introduction.

Reply: We added the math of the scaling discrepancy now directly into the introduction to help resolve any vagueness of the formulation.

• L378-381: Please state the purpose/objective clearer. What “latter” refers to is not clear.
Reply: We apologize for the confusion and rewrote the section (line 402-403)

• L395-407: This result is very cool. I wonder whether the GNNcoal classifier performs as well using inferred ARGs (also mentioned in “Main”).
Reply: We added new results based on inferred ARGs now. The classifier works when trained on inferred trees, using only topologies alone in order to not introduce any bias based on the different Ne scalings..

• L408-410:
– The hypothesis and expectation should be clearly stated. It is difficult to tell whether the demonstrated result fits the expectation under the hypothesis.
Reply: We amended the part: “Since strong selection can lead to multiple merge coalescent or rapid and successive coalescent event (as the beneficial alleles spreads very quickly in the population) \cite{Durrett_2005,Bisschop_2021,Sackman_2019}, we investigate if our approaches can model and recover the effect of selection.”

– Are the data used in the analysis true genealogies or mutations/inferred genealogies?
Reply: Data are inferred genealogies from simulated sequence data.

• L412-413: “Both approaches”. To me only SMβC seems to recover smaller α at the target locus of selection in Figure 7.
Reply: We agree with this observation. However,gnncoal works well for NeS=1000 but the dotted line covers the spike. We rewrote this section.

• L415-420: Please state the objective/question first. It is difficult to understand why this paragraph is here due to lack of this information.
Reply: We rewrote the paragraph (lines 433-443)

• L422-423:
– Please describe the objective.
Reply: We added the objective, see above.

– Was the simulation under Kingman or beta coalescent?
Reply: Figure 8 was simulated under the Kingman coalescent. We clarified the sentence by adding “Kingman coalescent”.

– Are the data true genealogies or mutations/inferred genealogies?
Reply: The data refers to true genealogies unless otherwise indicated.

• L423-426: What does “only up to a scaling constant” mean?
Reply: There is a known scaling discrepancy between the effective population size of the Kingman coalescent and the Beta coalescent. This scaling discrepancy makes itself noticeable during demographic inferences, see Figure 2 and S1.

• L444: “α > 1.3”. I would say α ≥ 1.7 based on Figures S15 and S16.

Reply: We modified it to α ≥ 1.5 which is a more accurate description.

- L445: “a larger amount of data is necessary”. Which result is this statement based on?

Reply: The inherent difficulty of attempting inference on the Beta coalescent arises due to the increasing sparsity with decreasing alpha values. Not only is more sequence required to have a sufficient number of trees, but also the number of ancestors decreases per tree due to multiple merger signals. We modified the sentence: “However, for lower values of α, a larger amount of data is necessary for any inference, specifically in the form of a high effective population size (correspondingly adequate mutation and recombination rates) and sufficient sequence length, which becomes nearly impossible when α tends to one.”

- L460: cf: XSMC (https://www.biorxiv.org/content/10.1101/2020.09.21.307355v1) as an SMC with continuous state space.

Reply: Thank you, we added the suggested citation.

- L470-471: This was not directly shown in Results. This requires analysis of true genealogies. This is related to my comments on L308-322.

Reply: This comment relates to the underlying mathematical formulation of beta coalescent. We believe that reproofing the theoretical underpinnings by simulations falls outside of the scope of this study.

- L471-476: “high variance”. Based on the results, I would expect inference of constantly lower effective population size as the effect of multiple merger events on SMC, instead of higher variance. Please elaborate why higher variance in inferred demography is expected.

Reply: Your expectation of constantly lower population sizes is accurate and we show that issue when running our Kingman experiments (see Figure 2). In this study we aimed to build models to be able to (inherently) account for expected lower population size (due to multiple merger events) and to correct for it. Notably, the GNN is set up to automatically account for the smaller effective size , because it was trained on random demographics under random alpha values. The notation of high variance comes from the amount and strength of multiple merger signals in each genealogy. If we were to plot the the average number of ancestors as a function of the alpha value for the first n genealogies, we would almost only have 9 ancestors (2*sample-size - 1 - sample_size) for a sample size of 10 and alpha 1.9. For alpha 1.1, we would have anything from 1 to 9 ancestors and correspondingly large variance affecting our inferences (see also the larger variance in the number of multiple merger seen in the new Figure S20 when decreasing alpha values)

- L481-483: “recurrent occurrences of the same multiple merger events at different locations on the genome”. Existence of such ancestral nodes in the true genealogies should be shown in Results. This is related to my comments on L308-322 and L470-471.

Reply: We agree that, after development of the algorithm, it would be interesting to delve deeper into multiple merger theory. For instance, we could identify a setting where GNNcoal performs well (has a low loss) or performs poorly (has high loss) to learn something new about the phenomena of multiple mergers and the ability to detect them. However, deeply studying the statistics of the beta coalescent to describe its behavior goes beyond the aims of this study. With the new LD figures, we already show the non-monotonous decrease of LD and seemingly chaotic correlation of sites which are not immediate neighbors.

- L502-505: Please consider restructuring the sentences for clarity.

Reply: Thank you, we modified it to “Improving the performance of GNN$coal$ on sequence data requires more efficient and accurate ARG inference methods, such as to incorporate inferred (non-exact) genealogies into the training, thereby by accounting for inference errors.”

- L506-537: In this paragraph the authors focus on GNNcoal, but it is difficult to tell until the end. Please make it clear that GNNcoal is focally discussed in this paragraph in the beginning.

Reply: The purpose of that paragraph is to introduce the current methodological limits and it is therefore general and not specific to GNNcoal, as we refer to “both approaches”. We introduced a paragraph break to make our logic clearer, i.e. where the general part stops and the GNNcoal part starts.

- L510: “linkage” should read “linkage disequilibrium”

Reply: We added the “disequilibrium” part.

- L525-526: This sentence should be in line 523.

Reply: Thank you, we reordered the sentences following your suggestion.

- L533-534: “selective processes favor coding regions”. Selection can act on regulation such as cis-regulatory (non-coding) regions.

Reply: We added the regulatory part: “[...] favor coding regions or regulatory potentially non-coding regions [...]”

- L538: “new state-of-the-art”. Redundant, so either “new” or “state-of-the-art”.

Reply: We removed the “new” part.

- Some references are incorrect. For example, ref 32 is not in Molecular Biology and Evolution but in Genetics.

Reply: Thank you for seeing this. We corrected these now.

Comments on figures and tables

Figure 1

As a biologist I could not understand this figure. If this manuscript is meant to target evolutionary biologists, this should be better communicated.

Reply: We addressed this point earlier, please see above:

We acknowledge that the figure by itself is not enough to understand the complete methodology and purely serves to provide an intuition behind how the genealogies are compressed into our demography estimates. We modified the text itself to directly cite the

Figure from which this Figure has been inspired by (Fig. 1 of https://arxiv.org/pdf/1806.08804.pdf) and adjusted the text to make it more readable. [...].

We also now refer to recent review papers on deep learning for population genetic inferences to help readers navigate through the specific terminology and current applications.

In this figure the authors focus on explaining GNNcoal, but having a model diagram for SMβC would be helpful.
Reply: For SMC figures, we refer the interested reader either to the original papers https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4116295/ or to reviews like: https://onlinelibrary.wiley.com/doi/full/10.1002/ece3.5888 or book articles: https://link.springer.com/protocol/10.1007/978-1-0716-0199-0_7

Figure 2
Please refer to my comment on L301-302.
Reply: We changed the Figure now, to have approximately the number of mutations in each Figure. Also, we highlighted the noisiness of the multiple merger at lower alpha ranges by showing an additional single replicate. The axes are also on the same scale now, making the figure more visually accessible.

Figure 4
How many sequences were used in SMβC? According to the legend it is 10 but in the figure it is 3.
Reply: We choose either random permutations of 3 or 4 from a total of 10 sequences and take the mean of the result for the SMBC method, while the GNNcoal directly uses 10. We add the bracket “(mean of random permutations of 3 or 4)”.

Figure 5
This figure is not referred to in the main text. As a suggestion, the authors might consider focusing on one demography scenario (leaving results for other demography models in supplementary) and showing the results for both using true genealogies and observed mutations/inferred genealogies in this figure . Label of x-axis is missing.
Reply: Not referencing Figure 5 was an error, thank you for seeing that. However the x-axis is visible under C and D and refers to the whole Figure 5. We prefer to leave the inferred results in the appendix, as we are working on methodological improvements for data application.

Figure 6
Please refer to my comment on L395-407.
Reply: We added new results in the form of training on inferred sequences on a tiny demonstration dataset using only topologies.

Figure 7
As commented on L408-410, are the results based on true genealogies, or mutations/inferred genealogies?

Reply: We added “from exact genealogies” to be clear, that we didn’t use inferred genealogies.

According to the legend 20 sequences were used for SMβC but according to the figure 3 were used.
Reply: We apologize for the confusion but the legend is correct. 20 sequences were indeed used for the analysis. However, 20 sequences cannot be simultaneously analyzed. So our SMBetaC analysis only 3 sequences simultaneously (i.e. M=3) but SMbetaC analyzes every triplets possible. The meaning of M is described in line 161 of the “SMC-based method” section.

The results for SMβC are nice. But for Nes ≥ 100, I wonder if the multiple merger events due to selection may be effectively represented as burst(s) of coalescence even with methods based on the Kingman coalescent.

Reply: Yes indeed, that is why although data is simulated under Kingman a small alpha value is inferred around the location of the loci under selection.

Figure 8
As commented on L422-423, are the results based on true genealogies, or mutations/inferred genealogies?
Reply: The results are based on true genealogies unless otherwise indicated. However, to be precise we added the “true genealogies” when needed and in the introduction.

Figure S1
Might it be worth including these in Figure 2?
Please put the equations for the correction in Materials and Methods for clarity.
Reply: We put the equations now directly in the introduction as they are quite fundamental to the paper.

Figures S2, S3

As commented on L308-322, please explain how to read the figures and what to expect under what scenario.

Reply: We answered this comment above. For complementary information we invite the reviewers to read Sellinger et al 2021 concerning the properties of SMC methods, as we would essentially simply rephrase what is written and explained there.

The numbers written besides the colour scale are not explained.
Figure S8

Reply: We apologize for the confusion. The variable M is defined earlier in the method section of the manuscript : “Our new approach, SMBetaC, is a theoretical extension of the MSMC algorithm, simultaneously analyzing multiple haploid sequences and focusing on the first coalescence event of a sample size 3 or 4 (this parameter is named M throughout the manuscript). We define as M the number of lineages simultaneously modeled by approaches.”

The results of down sampling in GNNcoal only under the sawtooth scenario are shown. Please also present the results for other scenarios. Figure S14
The results under the sawtooth scenario are shown. Please also present the results for other demography scenarios.

Reply: After having provided many demography results, we believe to have sufficiently proven the current methods capabilities. Our research work does not aim at finalizing GNNcoal at this current stage (and for this paper), but providing iterative improvements of the method. Therefore, making the method more general and extending its capabilities is the current focus of our attention rather than extending our appendices. We hope the reviewer understands this choice.

Figure S17
Please clearly state that they are based on neutral simulation.
Reply: In line 482, we wrote “[...] (assuming neutrality) [...]”.

Tables 1, 2, S1-S4
The data should be plotted as in Figure 5 for better accessibility.
Reply: We believe that replotting the data would not increase accessibility as plotting values can be in this case somehow misleading thus not as accurate as the table containing the values.

Reviewed by anonymous reviewer 2, 28 Sep 2023 20:49

This preprint describes two new methods to estimate evolutionary parameters (coalescence rates and a parameter alpha describing multiple merger rates) from sequence data. The methods address the impressively hard problem of demographic inference in the presence of multiple-merger coalescent dynamics which is certainly novel. While I must admit that I could not go through the two supplementary texts in the necessary detail to fully review it (they are too extensive for me and a review of them is simply beyond my time budget), I see no reason to doubt the authors' expertise and suggest to then rely on community review after publication.

I have a number of comments on the main article and supplementary Figures which hopefully help improving the clarity of the paper or possibly point to some gaps in the story that need to be filled before recommendation:

1) L 248ff: With the GNN method, I did not understand why the smoothing of the inferred demography from the GNN happens after the inference. It appears to me as if regularization should be built in right into the inference method. For example, why not infer B-splines, or if that is too hard, put penalties on large jumps between the piecewise constant rates in the model?

Reply: The smoothing does not happen after the inferences. Both methods infer the population in a stepwise manner. However the simulated demographics are smoothed to make them more realistic and not vary erratically. We modified the word “outputted” to “simulated” and added “before inference” to clarify that point.

2) Figure 2: I am quite confused about the "scaling discrepancy between the Kingman and beta-coalescent" (L 289f), as seen in the figure. In the figure, it looks like the notion of "population size" in the beta-coalescent is something that is between 2 and 3 orders of magnitude below what is called a "population size" in the Kingman-coalescent. Surely this cannot then mean the same concept? I don't know beta-coalescent theory well, but I suppose whatever is described there cannot be interpreted as a "population size" in the same sense as in the Kingman coalescent.

Reply : The confusion comes from our presentations of the results (the population size under the Beta Coalescent does make some sense to some extent). In practice, to recover the population size, we use the observed number of segregating sites and the known mutation rate. Under the Wright-Fisher Model (and the Kingman coalescent), the following formula stand (for a haploid population under constant population size) : Theta = 2 Ne μ L (Theta : expected Number of segregating sites between two individual, Ne population size, μ the mutation rate per bp per generation and L the sequence length).

However, under the Beta coalescent this formula does not stand. Which means that under the same model parameter the Beta and the Kingman coalescent will lead to a different expected number of segregating sites (mainly due to differences in ARG topology distribution, see the new Figure S20). This difference in segregating site numbers leads to a scaling discrepancy between the two models if not accounted for.

Maybe I am overlooking something, but I think if this is really just some artifact in the definitions of rates, they should simply always be shown in their "corrected version" in the main text. At the very minimum, I suggest to replace Figure 2 by Supplementary Figure S1. But even better would be a good explanation, or perhaps general synchonisation, of the 100-1000fold difference in the concept of "population size" between the two models.

Reply : The concept of population size is similar under both model (i.e. number of individuals given a chance to produce offsprings for the next generation). As an illustration, self-fertilization in diploid populations can lead to a smaller expected number of segregating sites when compared to outcrossing population (i.e. Wright-Fisher model), yet population size makes sense whether the population is selfing or outcrossing. We clarify the message in the introduction by introducing the different definitions of effective size for Kingman and MMC models, so that it is easier for the reader to follow our arguments.

Changing the figures would not support the main message of our study. We believe that it is important to stress that when estimating the population size, we do it under a certain model assumption. When this assumption is wrong, so are their estimations. Furthermore, the inference of the population size under the beta coalescence is equivalent to the inference of the alpha parameter, as performed with our methods.

3) Figure 3 and text describing it: I think the authors made a confusing choice for Figure 3 to show different x-axis scales. The three plots all look the same, but have different scales, so the difference is hard to see. I suggest to use the same scale, so the reader can appreciate the difference.

Reply : We adjusted the scale, so it is the same on each plot now. Furthermore, we updated/corrected the Figure, so that there is an equal number of mutations in each window, when comparing across the different models (see reply to reviewer 1).

4) Related to point above in Figure 3: I don't quite understand whether the shown LD decay for lower values of alpha is really qualitatively different from the Kingman-coalescent. I believe the authors when they say that multiple mergers lead to long-range effects, but on the Figures, it doesn't look qualitatively different, it just looks quantitatively different. Where can I see the "qualitative"? Why does a longer LD decay necessarily demonstrate "violation of the Markovian hypothesis"? I think this both needs to be explained better, and it needs to be shown more convincingly.

Reply : As mentioned above, our initial plot did not capture the variance and shape of the observed LD under the beta coalescent. Thus we replotted this figure to better show the effect of the Beta coalescent on the LD (also in reply to reviewer 1).

5) Fig S2 and S3: The authors show these residual matrices of the observed vs. theoretical transition matrices. This is in principle nice, but after all leaves me a bit puzzled about what I'm supposed to see. The authors point out the fact that S2 looks more random, while S3 looks more structured, but I don't get why the seeming randomness in S2 should be interpreted such that the matrix is "well approximated" (L 313), nor do I get why the patterns in S3 should be interpreted such that there are "significant differences" between observed and predicted" (L 316f). It seems to me that whether or not the residuals are structured or not is somewhat of an orthogonal question to question whether the differences are significant or not. In particular, they live on the same color scales.

Reply: We apologize that we did not provide sufficient explanation on how to interpret those results. But we reshaped this section of the manuscript to further develop the meaning of those results.

6) By the way, the tick marks in the color legends of Figures S2 and S3 have an error as far as I can see. The topmost tick marks should be "[10^{-7}, 1[", and not "[10^{-10}, 1[", right?

Reply:The reviewer is correct, it should be "[10^{-7}, 1[". We fixed it.

7) Fig S6 and S7: Why did you choose the timing of the expansion or contraction of the population size to be so recent? It seems that for most chosen alpha values, the inference is far away from the "interesting" time period.

Reply: Indeed, the scenario could have been chosen differently. However, it is difficult to find suitable scenarios for all possible parameter settings. Yet, throughout this study, we provided many demography inferences (especially the saw tooth scenario covers a wide range of contractions and expansions) and therefore hope to have sufficient evidence of the capacity of both GNN and SMBCs methods.

8) L 347ff: I was confused by the text here, which I understood in such a way that the GNN was run on a downsampled dataset to sample size of three, then pointing at Futures S4-S7. But in those figures, the figure legend indicates that the full 10 haplotypes were used. Is this just a typo, or did I misunderstand something?

Reply: We agree that referencing these Figures was confusing. Only Figure S8 displays the GNNs evaluation on a sample size of 3. We modified the sentence to: “Results for sample size ten are displayed in Figure S4 to S7 and downsampled results with sample size three of GNN$coal$, which appear to be similar, are displayed in Figure S8, demonstrating that the GNNs can better leverage information from the ARG in presence of multiple merger events.”

10) L 384f: "In contrast, SMbC produces better inferences of alpha ..." -> better than what?

Reply: “[...] when compared to the GNN” has been appended for clarity.

11) Figure S1, caption: The math seems a bit garbled to me, with single brackets as superscripts and such.

Reply: We ungarbled the math and introduce some bits now in the introduction, so the concept of the scaling between Kingman and Beta coalescent hopefully becomes clearer.