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.

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.

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;
2. In some places it is unclear whether true genealogies or inferred genealogies were used (e.g. L408-410, L422-423);
3. Some supplementary figures contain only one demography scenario for the entire analysis (e.g. Figs S8, S14).

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".
- General: Many paragraphs start from methods without stating the objective/hypothesis/expectation (e.g. L379; L387; L396; L408; L422).
- 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.
- L23-25: Please provide references for different types of survivorship.
- 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 $\alpha = 1.3$ 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.
- L232-233: Please state why mutation and recombination rates were set differently across scenarios with different α.
- L235-236: The authors might consider giving the two GNNs different names to avoid confusion by readers. (Or, are they considered the same GNN?)
- 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 $1.75 \leq \alpha < 2.00$?
- L288-290: "due to the scaling discrepancy between the Kingman and β-coalescent". This is based not on Fig. 2 but on Fig S1.
- L292-294: Was the scaling done upon the MSMC2 output, or was it done by modifying the MSMC2 algorithm?

- L300: “Linkage” should read “Linkage disequilibrium”
- L301-302: Please give some number representing higher LD. I cannot tell this well by visually comparing panels of Fig. 3 alone.
- L302-303: To show “a higher variance in LD”, the window size should be consistent across Fig. 3 A-C.
- 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.
- 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.
    - 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.
    - 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, 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.
- L331-332:
    - “both approaches seem to recover fairly well the true α value (Figure 4..)”: Figure 4 does not show inferred α.
    - I suspect Figure 5 shows it but it is not referred to in the main text.
    - Is Figure 5 the exact same as Table S1? If so then Table S1 is not necessary.
- L333-334: Which figure/panel is explained here?
- 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.
    - Please make clear whether the inferences were based on true genealogies or mutations. If the former, why are mutations necessary?
- 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.
- L353: “Figures S4 to Figure S7”. Figure S8?
- L367-369: Please elaborate what exactly is meant by “scaling discrepancy between the Beta and Kingman coalescence” in introduction.
- L378-381: Please state the purpose/objective clearer. What “latter” refers to is not clear.
- L395-407: This result is very cool. I wonder whether the GNNcoal classifier performs as well using inferred ARGs (also mentioned in “Main”).
- 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.

  - Are the data used in the analysis true genealogies or mutations/inferred genealogies?
- L412-413: "Both approaches". To me only SMβC seems to recover smaller α at the target locus of selection in Figure 7.
- L415-420: Please state the objective/question first. It is difficult to understand why this paragraph is here due to lack of this information.
- L422-423:
  - Please describe the objective.
  - Was the simulation under Kingman or beta coalescent?
  - Are the data true genealogies or mutations/inferred genealogies?
- L423-426: What does "only up to a scaling constant" mean?
- L444: "$\alpha > 1.3$". I would say $\alpha \geq 1.7$ based on Figures S15 and S16.
- L445: "a larger amount of data is necessary". Which result is this statement based on?
- L460: cf: XSMC (https://www.biorxiv.org/content/10.1101/2020.09.21.307355v1) as an SMC with continuous state space.
- L470-471: This was not directly shown in Results. This requires analysis of true genealogies. This is related to my comments on L308-322.
- 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.
- 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.
- L502-505: Please consider restructuring the sentences for clarity.
- 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.
- L510: "linkage" should read "linkage disequilibrium"
- L525-526: This sentence should be in line 523.
- L533-534: "selective processes favor coding regions". Selection can act on regulation such as cis-regulatory (non-coding) regions.
- L538: "new state-of-the-art". Redundant, so either "new" or "state-of-the-art".
- Some references are incorrect. For example, ref 32 is not in Molecular Biology and Evolution but in Genetics.

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.

In this figure the authors focus on explaining GNNcoal, but having a model diagram for SMβC would be helpful.

Figure 2

Please refer to my comment on L301-302.

Figure 4

How many sequences were used in SMβC? According to the legend it is 10 but in the figure it is 3.

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.

Figure 6

Please refer to my comment on L395-407.

Figure 7

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

According to the legend 20 sequences were used for SMβC but according to the figure 3 were used.

The results for SMβC are nice. But for $N_e s \geq 100$, 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.

Figure 8

As commented on L422-423, are the results based on true genealogies, or mutations/inferred genealogies?

Figure S1

Might it be worth including these in Figure 2?

Please put the equations for the correction in Materials and Methods for clarity.

Figures S2, S3

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

The numbers written besides the colour scale are not explained.

Figure S8

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.

Figure S17

Please clearly state that they are based on neutral simulation.

Tables 1, 2, S1-S4

The data should be plotted as in Figure 5 for better accessibility.