Review and Assessment of Performance of Genomic Inference Methods based on the Sequentially Markovian Coalescent

Stephan Schiffels based on reviews by 3 anonymous reviewers

A recommendation of:

Open Access

Published:

12 November 2020

The human genome not only encodes for biological functions and for what makes us human, it also encodes the population history of our ancestors. Changes in past population sizes, for example, affect the distribution of times to the most recent common ancestor (tMRCA) of genomic segments, which in turn can be inferred by sophisticated modelling along the genome.

A key framework for such modelling of local tMRCA tracts along genomes is the Sequentially Markovian Coalescent (SMC) (McVean and Cardin 2005, Marjoram and Wall 2006) . The problem that the SMC solves is that the mosaic of local tMRCAs along the genome is unknown, both in their actual ages and in their positions along the genome. The SMC allows to effectively sum across all possibilities and handle the uncertainty probabilistically. Several important tools for inferring the demographic history of a population have been developed built on top of the SMC, including PSMC (Li and Durbin 2011), diCal (Sheehan et al 2013), MSMC (Schiffels and Durbin 2014), SMC++ (Terhorst et al 2017), eSMC (Sellinger et al. 2020) and others.

In this paper, Sellinger, Abu Awad and Tellier (2020) review these SMC-based methods and provide a coherent simulation design to comparatively assess their strengths and weaknesses in a variety of demographic scenarios (Sellinger, Abu Awad and Tellier 2020). In addition, they used these simulations to test how breaking various key assumptions in SMC methods affects estimates, such as constant recombination rates, or absence of false positive SNP calls.

As a result of this assessment, the authors not only provide practical guidance for researchers who want to use these methods, but also insights into how these methods work. For example, the paper carefully separates sources of error in these methods by

observing what they call "Best-case convergence" of each method if the data behaves perfectly and separating that from how the method applies with actual data. This approach provides a deeper insight into the methods than what we could learn from application to genomic data alone.

In the age of genomics, computational tools and their development are key for researchers in this field. All the more important is it to provide the community with overviews, reviews and independent assessments of such tools. This is particularly important as sometimes the development of new methods lacks primary visibility due to relevant testing material being pushed to Supplementary Sections in papers due to space constraints. As SMC-based methods have become so widely used tools in genomics, I think the detailed assessment by Sellinger et al. (2020) is timely and relevant.

In conclusion, I recommend this paper because it bridges from a mere review of the different methods to an in-depth assessment of performance, thereby addressing both beginners in the field who just seek an initial overview, as well as experienced researchers who are interested in theoretical boundaries and assumptions of the different methods.

References

[1] Li, H., and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. Nature, 475(7357), 493-496. doi: https://doi.org/10.1038/nature10231

[2] Marjoram, P., and Wall, J. D. (2006). Fast"" coalescent"" simulation. BMC genetics, 7(1), 16. doi: https://doi.org/10.1186/1471-2156-7-16

[3] McVean, G. A., and Cardin, N. J. (2005). Approximating the coalescent with recombination. Philosophical Transactions of the Royal Society B: Biological Sciences, 360(1459), 1387-1393. doi: https://doi.org/10.1098/rstb.2005.1673

[4] Schiffels, S., and Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. Nature genetics, 46(8), 919-925. doi: https://doi.org/10.1038/ng.3015

[5] Sellinger, T. P. P., Awad, D. A., Moest, M., and Tellier, A. (2020). Inference of past demography, dormancy and self-fertilization rates from whole genome sequence data. PLoS Genetics, 16(4), e1008698. doi: https://doi.org/10.1371/journal.pgen.1008698

[6] Sellinger, T. P. P., Awad, D. A. and Tellier, A. (2020) Limits and Convergence properties of the Sequentially Markovian Coalescent. bioRxiv, 2020.07.23.217091, ver. 3 peer-reviewed and recommended by PCI Evolutionary Biology. doi: https://doi.org/10.1101/2020.07.23.217091

[7] Sheehan, S., Harris, K., and Song, Y. S. (2013). Estimating variable effective population sizes from multiple genomes: a sequentially Markov conditional sampling distribution approach. Genetics, 194(3), 647-662. doi: https://doi.org/10.1534/genetics.112.149096

[8] Terhorst, J., Kamm, J. A., and Song, Y. S. (2017). Robust and scalable inference of population history from hundreds of unphased whole genomes. Nature genetics, 49(2), 303-309. doi: https://doi.org/10.1038/ng.3748
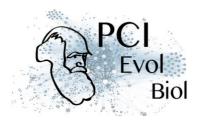
Reviewed by anonymous reviewer, 2020-11-02 01:14

I am satisfied with the authors' revisions.

Revision round #1

2020-08-25

This preprint by Sellinger et al. describes several analyses around the Sequentially Markovian Coalescent, a methodological framework used heavily in the field of demographic inference from genomic data.

The preprint has now been read by three anonymous reviewers. I have also read the paper carefully, and I agree with the reviewers' generally positive assessment. As reviewer #3 noted, while some of these results are probably already scattered around in the literature (also in Supplements), a systematically conducted and concisely summarised analysis of these various important caveats for SMC methods is still missing. So I definitely think this will be a useful and relevant contribution.

As you can see, all three reviewers have some comments for improving clarity, and possibly expanding the study a bit. I personally find two suggestions for adding analysis to be particularly worth considering: First, reviewer #1 proposed to add a constant population size scenario as a "basic" model to supplement the more complex demographic scenarios you currently have. Second, reviewer #3 suggests to add error quantification in small tables in all analyses using the mean square error.

I'm in principle happy to recommend this paper after a revision addressing the raised points by the reviewers. Please give good reasons if you believe some suggestions should not be followed.

Thanks again for submitting this interesting paper and I look forward to receiving the revised version.

Additional requirements of the managing board:
As indicated in the 'How does it work?' section and in the code of conduct, please make sure (**if appropriate**) that:
-Data are available to readers, either in the text or through an open data repository such as Zenodo (free), Dryad or some other institutional repository. Data must be reusable, thus metadata or accompanying text must carefully describe the data.
-Details on quantitative analyses (e.g., data treatment and statistical scripts in R, bioinformatic pipeline scripts, etc.) and details concerning simulations (scripts, codes) are available to readers in the text, as appendices, or through an open data repository, such as Zenodo, Dryad or some other institutional repository. The scripts or codes must be carefully described so that they can be reused.
-Details on experimental procedures are available to readers in the text or as appendices.
-Authors have no financial conflict of interest relating to the article. The article must contain a "Conflict of interest disclosure" paragraph before the reference section containing this sentence: "The authors of this preprint declare that they have no financial conflict of interest with the content of this article." If appropriate, this disclosure may be completed by a sentence indicating that some of the authors are PCI recommenders: "XXX is one of the PCI XXX recommenders."

Preprint DOI: https://doi.org/10.1101/2020.07.23.217091

Reviewed by anonymous reviewer, 2020-08-18 03:09

Download the review (PDF file)

Reviewed by anonymous reviewer, 2020-08-17 16:00

Download the review (PDF file)

Reviewed by anonymous reviewer, 2020-08-25 02:50

The authors conducted a simulation study of the strengths and weaknesses of some demographic inference packages based on the sequentially Markov coalescent, under various data and parameter regimes. SMC

methods are now widely used, in an increasingly diverse array of settings, and it is important to understand what causes them to succeed and fail. Although several of the conclusions reached here are scattered about in the literature, this is a more systematic attempt to organize them into a coherent set of recommendations for practitioners. So, it seems like a useful contribution.

I don't have any major concerns or objections, but I think the paper could be improved a bit, and perhaps expanded in a few related directions.

Major comments

- Error quantification: The performance of a statistical estimator is generally measured in terms of mean-squared error. The results shown in Figures 1-7 are qualitatively useful for building intuition about how each of the scenarios affects inference, but it is impossible to quantify the difference in performance between (or even within) different figures. Consequently, the discussion is entirely qualitative. Each figure should have an accompanying table with the MSE for the corresponding methods and scenarios, and those could be used to argue more rigorously about the strengths and weaknesses of various methods.
- Regularization: in several of the scenarios analyzed, the results seem like they could be improved by adding a penalty term. SMC++ supports regularization natively, and it could be easily added to the authors' eSMC package, but regularization is not really explored in the paper except briefly in Table 2. A thorough study of how regularization affects demographic inference, both in terms of what form of regularization to use as well as how to tune the hyperparameters, is currently missing from the literature to the best of my knowledge (but see the recent preprint from Kelley Harris' lab on their method mushi). I realize that one could easily write a whole other paper on this, and am not advocating for major additions along these lines. Still, another subsection or two on this topic would be very useful in applications.
- There are various other confounders that could be taken into consideration. I think ascertainment bias in particular would be interesting to look at. The ASMC paper delved into this a bit, but there is more that could be done. How badly does ascertainment bias (potentially in a related population) and SNP sparsity affect PSMC? This could have important practical consequences since a lot of fields still rely on microarrays. It could be incorporated into the present paper by running msprime on a large sample size and only keeping SNPs above a certain MAF threshold.
- I don't quite get the focus on estimating rho/theta (though I understand its effects on inference). I tend to think of this as a nuisance parameter when running SMC methods. It is not reasonable to assume that rho is constant over the whole chromosome anyways, nor should we expect this to be a good estimate of the chromosome-wide average rho when the true underlying rates are heterogeneous.

Minor comments

- This sawtooth demography is slightly different from one in the original MSMC paper. The final nadir at 10^1 generations occurs too recently, and the population size should be constant with Ne=14312 from 33 generations ago to present. This isn't such a big deal since the model was pulled from thin air in the first place, but since the community has coalesced around this model as a benchmark, it's better if the paper used the same version of it as everyone else. The stdpopsim package can simulate directly from this demography in a few lines of code.
- Sections 2.1.4 and 3.1 / Fig. 1: The titles give a somewhat misleading impression. A theoretical convergence result would be very nice, but that's not what is offered. I would prefer to call this something like "Best-case convergence".

- 381: "SMC++ seems especially sensitive" -- I don't see this reflected in Figure 4. If anything, it looks less sensitive than the other three methods. This makes sense to me since for high values of rho, the frequency spectrum is a better estimator of demography than methods which use only linkage information.
- 565f: "Could be used in more complex scenarios". Recent theoretical work (see the article 'How Many Subpopulations Is Too Many? Exponential Lower Bounds for Inferring Population Histories' by Kim et al, JCB 2020) strongly suggests that this is not possible. The IICR is not useful for recovering complex demographic histories.
- Various spelling or grammar errors:
  - 35: ecologist
  - 44: estimations/interpretations
  - 47: state-of-the-art
  - 50: well-known, Pairewise
  - 101: simulates
  - 149: discretized
  - 287: I think it should be "whose dynamics"

Author's reply:

Dear Recommender,

Please find attached our revised manuscript entitled "Limits and Convergence properties of the Sequentially Markovian Coalescent" by Thibaut Sellinger, Diala Abu Awad and Aurélien Tellier, which we would like to be considered for recommendation in PCI Evolutionary Biology.

First, we would like to thank you for giving us the opportunity to resubmit this manuscript and your positive comments. We would also like to thank all reviewers for appreciating the importance of our work and for their useful comments.

We paid close attention to answering all the reviewers' comments and have modified the manuscript accordingly. We believe that we have improved its readability, as we rewrote some sections of the manuscript that the reviewers felt were unclear. We have also included new Supplementary Figures (seven in total) corresponding to the requested analyses and six Supplementary Tables, containing the mean square error of past demographic inferences of all the figures in the manuscript. These measures have helped us, and hopefully will help the readers, to better understand our results. However we found that the MSE alone cannot precisely measure the performances of the methods (see the reply to the reviewers' comments for more detail).
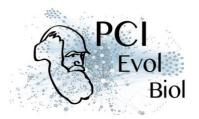
In addition to what reviewers requested we made some additional corrections. First, we realised that the time window of the theoretical convergence analyses (now called best-case convergence) was ill-defined by a factor 2. All analyses were therefore run again to fix this. Secondly, there were minor errors in the msprime command lines when simulating data for SMC++, which required that we re-simulate all data using msprime and have re-run all analysis of SMC++. Slightly different results are observed for Figure 4 and for Supplementary Figure 14 compared to the first version of the manuscript, but all other SMC++ results are identical. We noticed that the section concerning transposable elements was confusing, and thus rewrote the section while adding two supplementary Figures (35 and 36). We hope our motivations and our results now appear in a clearer way. Lastly, we fixed a plotting issue in Supplementary Figures 15 and 22.

We hope that this revised version fulfils the criteria for recommendation in PCI,

Many thanks in advance,

Yours sincerely,

On behalf of the authors, Thibaut Sellinger.

[Download author's reply (PDF file)](#)