Detailed response to comments:

Reviewer 1 :

Major comments and suggestions:

I don't have any major comments.

Minor comments and suggestions:

It could be very useful further explore the case of simulated flat demography.

We added supplementary figure 11 to cover this case.

Since variation in the recombination rate can cause spurious waves in the inference (and virtually all species have a heterogeneous recombination map), it would be nice to show if/when one can trust "small" wiggles in the Ne curve. If simulations show that it is not uncommon to see an artifactual > 2 fold expansion/contraction, then guidelines specific to these cases (eg, when can I reject the "null" model of constant Ne) would be appreciated by the community.

We believe it is hard to quantify the actual effect of a heterogeneous recombination map on demographic inferences since it might depend on the topology of the recombination map. However, since the release of iSMC one can at least test the hypothesis of a constant recombination rate along the genome. We added guidelines (lines 599-605) which suggest running iSMC in order to infer the recombination map and then simulating data under the estimated recombination map and under a constant recombination rate in order to estimate the biases caused by the heterogeneous recombination rate.

Section 2.1.6 could be written in a more clear way (e.g., link time windows to hidden states).

We rewrote the section for clarity.

Like 149: unless I am mistaken the number of hidden states is not necessarily proportional to the number of parameters in SMC++ since they use cubic splines interpolation.

We removed the reference to the number of parameters (line 152).

Lines 298 and 304: Be more specific. Which percentage of the genome is removed? Are all fragments of the same size? Which size?

We added a more detailed description, indicating the length and the proportion of TEs in the sequences (line 340,344).

Line 335: This is overall a very instructive paragraph, and it seems to me that the CV is computed over the entries of a single transition matrix, but are there not replicates for

each scenario? Could it be worth to explore the CV of particular entries over different replicates?

We rephrased the sentence describing Figure 2 (line 376-380). We did exactly what the reviewer described; We simulated transition matrices under a specific scenario, then, from the replicates, calculated the CV at each entry of the transition matrix (line XXX-XXX).

Line 432: Somewhat unclear writing, I don't understand why they would not be accounted for as missing data (eg, emission probability = 1 as in PSMC)?

We apologize for the confusion. We rewrote the whole section describing simulations (lines 341-347). In Figure 7 we assumed that TEs are not detected and thus cannot be accounted for (e.g. when using the reference genome of a sister species). We agree that if TEs are detected and correctly masked, emission probability is set to 1 as in PSMC',MSMC and MSMC2. We thus added supplementary figures 35 and 36 to measure the effect of correctly masked TEs on inferences.

Typos and other small suggestions:

Line 96: "hypothesis" -> "hypotheses" (supposedly this is plural)?
Line 101: "simulates" -> "simulated"
Line 129, 133, 134: "genealogy" -> "Ancestral Recombination Graph"?
Line 163: "does" -> "do"
All fixed

Lines 165 - 168: a bit of an over-simplified picture since this explanation does not take recombination rate (and past Ne) into account. It could seem to the reader that the inferred hidden state at a homozygous site is always the shortest genealogy whereas for a heterozygous site it would be the tallest genealogy (ie, it sounds like the hidden states are completely free to change from one site to the next).

We rewrote the section to avoid confusing the reader (line XXX-XXX).

Line 169: the TMRCA does not necessarily change after a recombination event in the SMC'

We rephrased it correctly (line 171).

Line 249: add "exponential" to be more precise?

Fixed (line 272)

Line 316: avoid the term "perfect fit".
Fixed (line 354)

Line 363: “find” sounds a bit confusing. Does it refer to the occurrence of hidden states (unknown truth) or the inference procedure?

It refers to the inference procedure. We rephrased it correctly (line 410).

Line 363: “which” -> “with”?
Fixed

Line 390: “MSMC is able to retrieve the correct recombination rate [...]”. I can see it’s closer to the real value, but it’s hardly “correct”.

Fixed (line 445)

Line 430: "The smaller the sequences that are removed, the more rho/theta is over-estimated.” -> “For a fixed amount of missing data, the smaller the sequences that are removed, the more rho/theta is over-estimated.” or something like that?

Fixed (line 497)

Lines 444, 448: avoid using “perfect”.

Fixed (line 514)

Line 457, 460: “genealogy” -> “genealogies”
Fixed

Lines 458 - 460: "Our results suggest that whole genome polymorphism data can be summarized in a transition matrix based on the SMC theory to estimate demographic history.” The term “demographic history” could potentially be interpreted in the broad sense (eg, including structure and migration), but this has only been shown for the panmitic case. Maybe rephrase?

Fixed (line 530)

Line 468: “genealogy” -> “genealogies”
Fixed

Line 472: “hidden states” -> “transitions”
Fixed

Line 492: This is an important sentence, would be important to clarify how missing data is handled in the Results section.

Fixed (line 565)

Reviewer 2 :

Comments:

1. PSMC' has been mentioned repetitively and tested for multiple factors in this paper, e.g. in Fig 1. As far as I understand, when you refer to PSMC' software you meant you used eSMC, or not? If yes, please be consistent throughout the text. Also it has unnecessary confusions, because Schiffels and Durbin 2016 paper calls the special case of two haplotypes of MSMC as PSMC'. If not, please introduce what software you meant when you test PSMC', perhaps add this information at the beginning of your introduction (Section 2.1.1).

We apologize for the confusion. eSMC is a reimplementation of PSMC'. We now indicated eSMC each time it is used. Note that for the theoretical convergence or what we now call the "best case convergence", eSMC is mathematically equivalent to PSMC' . It is therefore the best-case convergence of PSMC' (and eSMC) that is plotted.

2. Section 2.2. It is necessary to add more details on how the simulation was done. For example, currently it does not show readers that how you simulated data for MSMC, MSMC2, eSMC. Please be explicit on what data format you simulated for each SMC-based approach, corresponding simulator software you used, and the total size of simulated sequences.

We clarified how simulation were performed (section 2.2 , lines 265-281).

3. Line 329-220, 'MSMC shows better fits in recent times than PSMC' and ...'. Be aware that you use four sequences for MSMC, and one for PSMC' here? I am very confused here that why you don't test on MSMC2 which is conceptually more similar to PSMC compared to MSMC. Also to make fair comparison, you should be consistent on the number of sequences you used in the tests.

We understand the confusion, and have added clarifications. The results mentioned are in section 3.1 for the theoretical convergence results (defined in section 2.1.4), now called best-case converge. One must not think in terms of "number of sequences" or even in "sequence length", but in terms of "total number of transitions". Though it depends on the sequence length and number of sequences, the number of transitions is in no means identical for sequences of the same length or number, which is why we can use different numbers of sequences without this being problematic. Concerning MSMC2, it is mathematically similar to PSMC' and eSMC (the main difference being the time window) and would thus display similar results. Yet, we implemented the time window of MSMC2 in the R package eSMC2. This way users can specify whether to study the best case convergence of MSMC2 or PSMC'.

We understand your concerns about "fair comparison", however we chose to compare methods in the "ideal" conditions specified by the guidelines for using them. We find our choices to be more pertinent in this setting, as each method has been built for different

data sets and performs best under different conditions, and it is how well methods perform when given an ideal data-set that we are interested in comparing. Results for one-on-one comparisons, as you have suggest we carry out, for big data sets can be found in https://doi.org/10.1038/s41588-019-0484-x).

4. Paragraph from line 335-345 needs to be rephrased in a clearer way. You are discussing two factors affecting the coefficient of variation of the transition matrix: i) the length of sequence, ii) the amplitude of population size variation. Then you talk about there is a lack of coalescence events in some specific time windows when there is a large population size change. Please rephrase because it is difficult to follow currently.

We hope to have rephrased it in a clearer way (line 372-383).

5. Figure 3, why two sequences for eSMC, MSMC2, and four for MSMC, 20 for SMC++? Isn't it better to be consistent? Also why 20 sequences of 10MB, rather than 100MB?

We apologize for the typo here. We used 100 Mb to simulate data for SMC++ (cf appendix 1 for the python script we used to simulate data). As mentioned above, the main aspect of our study is not to compare methods given the same data set, since this has either already been done or cannot be done. MSMC cannot run (at least on a local machine) on sample sizes bigger than 10Mb and requires phased data. SMC++ is recommended to be used with data sets with more than 20 individuals (https://github.com/popgenmethods/smcpp) and does not require phased data. Thus both methods cannot be correctly compared using the same data sets. On the other hand, MSMC2 and eSMC preform well with small sample sizes. We here investigate what can be inferred given a typical reasonable set of data for each SMC method.

6. Line 367-368, 'shifting the window towards more recent time leads to poor demographic estimation'. Which SMC-based approach you are talking about here, or in general? Add this information here and in Table 1 caption.

We used eSMC to perform this. We added the information line 414.

7. Line 380, ' the lower θ , the better the fit of the inferred demography'. From fig4, it is true that orange and red lines look better. But are you surethat you could conclude this correlation from your results? I find oranges lines fit better than red lines in my view. I would rephrase this sentence.

Indeed, this is true only for SMC++ and eSMC. We rephrased this sentence (line 435).

8. Line 298, you conclude > 10% of spurious SNPs can lead to a strong over-estimation of population size in recent time. . Note that you mentioned spurious SNPs in line 266, but you didn't explain your criteria of spurious SNPs (i.e. quality filter threshold). Please add this information.

Spurious SNPs are linked to the quality of the sequencing data and subsequent filtering and mapping quality. As we simulate data, we define a spurious SNP as one we randomly artificially added into the simulated sequences (section 2.2.2). The underlying idea is that the more stringent the SNP filtering, the fewer spurious SNPs maintained in the sequence, but as consequences, true SNPs are missed (especially those with a low frequency, see Pfeifer Heredity 2017 https://www.nature.com/articles/hdy2016102). And since our results suggest a lower bias when SNPs are missing than when there are spurious SNPs (see Figure 5). Therefore, we recommend more stringent filtering to be applied (see line 563).

9. Line 403-405, could you rephrase this sentence? It is not straightforward to understand.

We hope it is now clearer (section 3.3.2, line 466-479).

10. Line 418, “eSMC(i.e. PSMC’)”. This redirects to my point 1.

As indicated above we clarified the use of eSMC/PSMC’.

11. Line 437-438, “The longer the masked parts are, the stronger the effect on the estimated demographic history”. I am not sure I understand this perfectly. As the masks are designed for Transposable Elements (TE), you are basically saying the longer TEs the genome has, the stronger the effect is? You mean the total length of TEs in the genome?

We meant the length of TEs and not their proportion in the genome. We corrected the text to make it clearer (line 506).

Minor points:

1) Method developed in Ref 63 is called MSMC-IM, rather than IS-MSMC. Please correct this in line 84, 110.

2) Line 120-121, the item θ “can be greatly influenced by life-history” statement could be more accurate, e.g. ‘by life-history in organisms with self-fertilization and dormancy’

3) Line 266, typo ‘surious’ -> ‘spurious’.
4) Fig. 2, as you are plotting a square matrix, better to have same labels on x-,y-axis, and the same color skeme for four subplots.

All Fixed

5) As a general rule of plot for fig1,3-6, better to have legend out of box.

We understand your point and we tried to plot the legend outside of box but the result was more confusing. In addition, having the legend inside the box is what is conventionally done (see Li & Durbin 2011 , Schiffels & Durbin 2014 and Terhorst et al 2017). We therefore tried to keep the same layout.

Reviewer 3 :

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.

  Thank you for this valuable suggestion. We added supplementary tables 1-6 with the mean and coefficient of variation of the mean-squared error (MSE) of each analysis in the study. We agree that the MSE is a good indicator for statistical performance and therefore have used the calculated MSE to better support our interpretations. However we also observe that the MSE alone cannot fully measure the performance of the tested methods. The definition of MSE itself can lead to confusion since the MSE requires integrating the error over time and thus would not capture the method's accuracy in recent time but rather its performance in the far past (since there are more generations in the past). To avoid this problem we integrated over the log10 of time (which corresponds to what is plotted) to calculate the MSE. In addition, a shift in time might lead to a high MSE although the shape of the demographic history is correctly estimated. Furthermore, poor accuracy on the edges of the time window will lead to a high MSE although the demographic history can be estimated fairly well within the rest of the time window. We propose a simple MSE computation here as a first step, but further investigations are needed to develop more accurate statistics.

- 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.

As you explained, regularization is a complex topic which would require a complete study. Nonetheless, we implemented regularization penalty in eSMC2 and added supplementary figures 17 and 18 to test the effect of regularization penalty on past demography inferences. Yet, we agree that further studies are required.

- 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.

We added supplementary figure 25 to test the effect of removing SNPs under a certain MAF threshold. We generally find that, the more SNPs removed, the poorer the estimations in recent times, so the higher the MAF threshhold, the poorer the estimations. Note however, that microarrays are mainly used for model species or crops with a long history of genomics. For non-model species studied in ecology, the development of microarrays is very limited because reference genomes do not exist in the first place. It is nowadays less expensive and easier to perform RAD-seq studies and full genome sequencing than to develop microarrays.

- 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.

In the rho/theta ratio there is information beyond the recombination rate. Life-history traits, such as self-fertilization (and inbreeding), parthenogenesis, clonality, large variance in offspring production, and dormancy, but to name a few, can affect the observed rho/theta (at the chromosome scale, see the Book by M. Lynch on Genome Organization for example). Although these are of little consequence when working with mammalian data, they are present in a majority of animal, plant and prokaryote species. Thus from the discrepancy of the recombination rate measured experimentally and the one inferred by SMC methods, one can account for these traits and sometimes estimate the prevalence of certain ecological behaviours (see Sellinger et al 2020 https://doi.org/10.1371/journal.pgen.1008698, for more details). This is of interest to understand the effect of life-history traits on genome evolution and genetic diversity (see the review by Galtier and Ellegren in Nature Genetics, https://doi.org/10.1038/nrg.2016.58).

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.

We simulated the data using the original ms command lines of the arbitrary sawtooth demographic scenario mentioned by the reviewer (although using scrm and msprime to stay consistent). We analyzed the simulated sequences and added the results in supplementary figure 12, which are very similar to our alternative arbitrary scenario.

- 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".

We understand the confusion and have modified its name to what you have suggested : "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.

We changed the text and have zoomed out the plot to better display SMC++ results.

- 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.

The performance of MSMC-IM (based on an HMM machinery) strongly suggest its transition matrix could be used to infer migration and population size at least for 2 populations although it needs to be clearly demonstrated. In addition a recent study uses and extends the IICR to simultaneously infer migration and past demography (Arredondo et al BioArxiv 2020, https://www.biorxiv.org/content/10.1101/2020.09.03.282251v1).

- 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"

All fixed