Dear Aurélien Tellier,

Thank you for managing our submission and for your encouraging decision and we apologise for the long time it has taken us to elaborate the revision. We agree in general with the comments by the reviewers. Regarding the methodological shortcomings highlighted in their reviews, we are aware of them but, for practical reasons, we consider it better to address most of them in our future work. Our current implementation of the method is still quite computationally costly and we have started a complete rewriting of the code (https://github.com/mnavascues/ABC_TimeAdapt) that will make it faster, use less memory and be more flexible to incorporate additional elements in the model and sampling scheme. Unfortunately this is ongoing work and it is not ready to allow us to address all of the points raised by the reviewers. However, it is our opinion that our current results are important enough to be disseminated to the community, they offer a proof-of-concept of the approach and can stimulate further work beyond our team. Therefore, as you suggested, we have used our current code to explore only the effects of heterogeneous recombination along the genome. We have also revised our manuscript to take into account the suggestions and comments by the reviewers. Below we present a more detailed answer to each specific point made by the reviewers. We would like to remark that we have also noticed their positive comments which are very much appreciated, but, naturally, the reply focuses on their criticism.

On behalf of all authors,

Vitor Pavinato & Miguel de Navascués

Review 1 (edited to show only the points that need some comment)

- I find the ABC-rf to be well defined but not the data on which it was applied. In the main text (or in the appendix) I could not find basic information on the input data such as sample size of simulations or of the data (or if sample size at t1 was equal to sample size at t2, which does not seem to be the case according to what is written l 442). I believe the authors should add a table (in appendix or main text) describing what they simulated and the data they have used. Currently, the input data is very cryptic and I'm not sure results could be reproduced from the main text only.

Although the previous manuscript contained this information, the text was somehow confusing as it described the data and the model at the same time. We have modified the text to be more explicit and hopefully more clear (lines 227-239, for the model evaluation; lines 277-304, for the analysis of the temporal data of the feral populations of *Apis mellifera*).

The Authors assume census population size to be constant in presence of selection (if I understood correctly). I think it's necessary to run their ABC-rf on data that has been simulated with non-constant population size (e.g. bottleneck) but under neutrality.

See below our replies regarding the use of additional features in the model, including changes in population size and neutrality.

Secondly, and they mention it in the discussion, it would be interesting if they could run their approach on data simulated under neutrality, constant population size but in presence of migration. I believe those two analyses could help the authors to better understand the performance of ABC-rf on their data and strengthen their discussion.

We agree that further features in the model considered are important to fully understand the performance and robustness of the approach and also to evaluate the fit of the model to specific datasets. Unfortunately there are several factors to consider that could affect the analysis and implementing all of them into the framework requires substantial work. Here we try to strike a balance between the need to communicate the development of a novel methodological approach and exploring all its potential limitations. To this respect we followed A. Tellier's suggestion and we studied the effect of heterogeneous recombination rate but the effects of migration are not studied in our new version of the manuscript.

- The Authors mention summary statistics being used by the random forest to perform the inference, yet no summary statistic of the data are displayed in the manuscript or in the Supplementary files.

A table with the summary statistic values for the feral bee populations have been added as supplementary table S4.

As a sanity check I would add few Supplementary Figures/Tables displaying the "interesting"/"Selected" summary statistics of the data (based on S4 and S5), and the same summary statistics calculated from the simulations where parameters are set to the most likely (i.e. with highest density) one (Figure 4). If summary statistic calculated from data are similar than those of the simulated data, it would support their conclusions. On the opposite, divergence between both could indicate an underlying non-accounted process, which they also mention in the discussion.

In classical ABC analysis, one can perform a posterior goodness-of-fit evaluation by comparing some summary statistics from the simulations (or new simulations taken from their parameter values taken from the joint posterior distribution) that pass the rejection step with summary statistics of the target data. Good practice dictates that summary statistics used for goodness-of-fit must be different from those used for inference. Unfortunately, the use of random forest in ABC makes the process incompatible with such an approach as it uses all summary statistics and there is no rejection step or joint posterior distribution. In fact, the development of an equivalent posterior goodness-of-fit evaluation in ABC-RF is an active research topic (Arnaud Estoup, personal communication). Nevertheless, we completely agree with Reviewer 1 that some sanity check procedure is in order. So, instead of a posterior goodness-of-fit we have included an evaluation of the prior goodness-of-fit by applying a principal component analysis to summary statistics from the reference table and projecting observed summary statistics from the target data, which is also a classical prior evaluation approach in ABC. The “visual” fit of the model across the different populations is variable but all of them show PC axes for which the real data seems to be an outlier with respect to the model (Figures S8 to S14). AT the same time, many features of the real data seem to be captured in the model as many PC axes contain the observed data within the simulated variability. This highlights some shortcomings in the model that are likely the ones already discussed: admixture with other populations (which we know it happened) or other selection and demographic processes (like non-constant population sizes and selection on standing genetic variation.

-l 24, what do you mean by "realized potential" ?, I would rephrase

It was meant to say “realized adaptation”. Corrected.

-l 50, I would define "type I error rate" in the text for clarity

Substituted for “false positive rate”.

-l 80-95, you mention your methods to be accurate and that random forest has solved the computation issue, yet you didn't presented your input data. It would be usefull to have a sentence/paragraph describing the input data.

As replied above, description of the data has been rewritten to make their description more clear in the methods section, lines 225-254 for the model implementation; lines 277-304, for the analysis of the temporal data of feral populations of *Apis mellifera*).

-l 107, if you assume a constant population size of N, please mention it more clearly here.

We reworded our description of the model to explicitly state that population size was constant on each period (lines 106-110).

-l 126, I think you should also add the sample size here.

In this part of the text we describe the general model. Precise values for prior distribution of parameters and samples are given in lines 229-234.

-l 127-128, from the text it seems like you sampled individuals at time t1 and t2, with tau = t2-t1. But from Figure 1 A), it seems like you have been sampling between t1 and t2. In addition, tau is missing on the Figure 1. I would change Figure 1 A) by remplacing "sampling period" by "tau".

We interpret this comment as signalling the lack of clarity in the presentation of the model and sampling scheme, this has been hopefully solved by the rewritten of the description of the model as discussed above. We think that replacing “sampling period” for “inference period” should be more clear.

-l 133, I would change "c0" to "r0" for the recombination rate as I think it's a more classic notation.

Changed to "*r*" throughout the text.

-l 225-249, I do not think all variables have been clearly defined in the previous sections and there might be some inconsistent notations, e.g. is r="c0" ? I think it would help readers if the authors moved Table S1 into the main text (but r is missing from table S1).

The notation inconsistency noted by the reviewer has been corrected. We consider that Table S1 is better placed in the supplementary as it is redundant (we have defined everything in the text).

-l 285, Once the paper is accepted/recommended please add a link here in the text to the individual VCF files.

Data has already been made publicly available by their original authors. We have now included the database references as indicated in their original publication for easier reference to them.

-Table 1, what is written in column N (and add it to the table description) ?

-l 290-309, I didn't find where the sample size you are using was written. Could you add it in the text or in Table 1 ?

Table 1 contains the sample sizes, but they were wrongly labelled as “N”. This was another error in the notation that has been corrected.

-Figure 4, I think there is a typo in the legend, c) and a) should be exchanged.

The labels for the panels in the figure caption have been corrected.

-Figure 4, I think having a small table summarizing Figure 4 with the most likely (or whith highest density) ratio of Ne/N would help in understanding the results.

Numerical results are presented in Table S2 (Supplementary Material, page 6, including N_e/N ratio. We have included a reference to Table S2 in the caption of Figure 4 for clarity.

-l 440-498, The authors discuss the similarity between Ne and N and they mention the possibility of excluding lower values of theta_b indicating acting selection during the study (by giving explanation on why signature of selection might be harder to find). Could it really not be neutral ?

In principle, we cannot conclude anything about models that were not evaluated and a pure neutral model was not considered. However, simulations with low Θ_b are the closest to neutrality that we have considered and indeed the ones with lowest Θ_b contain very few or no loci under selection during the inference period. In fact, a significant proportion of the simulations present no fitness differences between individuals (L=0; see, for instance, lines 247-254), making them effectively neutral. With very few or no loci under selection there will be a lower proportion of outlier loci for statistics that are informative to selection. As the number of these outliers changes with the frequency of selection the overall genomic distributions of this statistics changes, and the summary statistics that we use to characterise them (such as the quantiles, and skewness) also change (and these are the statistics that are used for the inference of Θ_b, as seen in Figure S4). Therefore, we can predict that if the only thing that we change in the model is to remove selection, the resulting simulations would be similar to the simulations with low values of Θ_b that have low posterior probability density.

- I don't know if authors have read the recent paper of Barroso et al (https://doi.org/10.1101/2021.09.16.460667) on the inference of the mutation rate map from whole genome data. In their study they interpret the variation of the Ne along the genome as a the variation of the mutation rate. However, if you assume the mutation to be constant along the genome, then they would infer the variation of Ne along the genome. I wonder if the inferred mutation rate map would be an interesting summary statistic for you ABC-rf ? I would recommand the authors to read the study as it could be relevant in the light of their work (and maybe add it to the discussion).

We agree that heterogeneous mutation rate along the genome is one of the potential confounding factors in this approach. However, it is probably one of lesser concerns in our particular case. Summary statistics informing about the heterogeneity of genetic diversity along the genome (e.g. He) are not among the most informative (). As stated above, our objective in this paper was not to explore all these factors. In the new version of the manuscript we mention these factors.

Review 2 (edited to show only the points that need some comment)

In their preprint, Pavinato et al. describe an ABC Random Forest approach to jointly infer demography and selection in population genomics, two-time points temporal data. Overall I like the approach a lot, especially the fact that it can be extended to more realistic situations that for example include deleterious mutations. Reading the preprint I had an issue with semantics, that has to do with the fact that the authors do not really infer demography per se, but rather provide an approach that allows to estimate selection while also jointly estimating the influence of "whatever" past demography occurred. This is different from what most people in the field understand by "inferring demography", which usually means explicitly inferring population size changes over time. This needs to be made clearer throughout the manuscript.

As also suggested by reviewer 1, we have worked on the text to make it more explicit that population size is constant in each period of the model.

1) Recombination. The authors use uniform recombination in their approach. We know however that many genomes have highly heterogeneous distributions of recombination events, and those even often co-vary with with functional elements in genomes. For example in human genomes, recombination hotspots tend to co-coccur together with regulatory elements. This is an example where not accounting for this heterogeneity is a big issue given that any approach with uniform recombination will then likely underestimate adaptation that occurs in these functional elements. The authors need to test with simulations with heterogeneous recombination how much their approach is affected, especially when the recombination hostpots tend to occur near where the adaptation is expected in the first place.

We completely agree that heterogeneous recombination rate along the genome is one of the main potential confounding factors in this approach and we did not discuss it enough in the previous version of the manuscript. We have followed the recommendation of exploring the effects of heterogeneous recombination rate.

2) Deleterious mutations. In the same spirit, instead of just discussing very briefly about this limitation, the authors could actually run a few simulation to see how their approach performs when the test, simulated genomes actually include deleterious mutations. What happens when the deleterious mutations are distributed evenly across loci, and what happens when the deleterious mutations are heterogeneously distributed across loci. The deviations from the expected values especially for adaptation should then provide more solid ground to discuss the limitations in much more detail than it is done at the moment.

I am suggesting thes two things (recombination and deleterious mutations) since it would not take too many additional simulations with SLIM to get a more precise sense of how not accounting for them biases the estimates.

We agree that the presence of deleterious mutations is another potential confounding factor. However, as argued above, we believe that this will need to be addressed in some future work as doing it properly requires much more work than what is suggested by reviewer 2. As previously stated, we are following A. Tellier's suggestion and we will explore only the effect of recombination in the present work.

In addition I just have some minor comments about a few things that need to be better defined in the Introduction. First, the authors need to better describe how Random Forest works and why this works better in the context of ABC. Second, the authors need to better define early in the introduction what latent variables are, how they work, and what their usefulness is in the approach.

The introduction has been revised to give some more information on ABCRF and latent variables.

Review by Lawrence Uricchio (edited to show only the points that need some comment)

1. Overall I found the descriptions of the analyses to be clear, but how they fit with the existing literature on joint inference of selection and demography was sometimes less clear. Some important areas of the literature on joint selection/demography inference are not mentioned in the manuscript (described more in the next point). It would be very helpful if the authors could further clarify how the approach that they take (random forests with ABC) provides an advance over the range of previous approaches. Their method aims to directly account for the effects of linked selection (which I agree is an important goal), but whether their method actually performs better in a practical sense than methods that do not explicitly model linked selection does not seem to be assessed. The effect of the random forests on the performance relative to ABC without the inclusion of the random forests is also not mentioned, although this may be a less feasible comparison given computational efficiency limitations.

Lawrence Uricchio points out that some part of the literature regarding the inference of demography and selection has been neglected in our manuscript (see next point for further discussion). However, when looking specifically at methods that address the inference of selection and demography, from temporal population genetics data without additional information/assumptions on neutral and selected sites, there are not many methods to compare to. Methods that target that type of data to address the inference of selection, such as WFABC (Foll *et al*. 2014; doi:10.1111/1755-0998.12280), focus on the detection of outlier loci or the estimation of selection parameters at a single locus, while our approach focuses on estimating genome wide parameters (or latent variables). The methods are complementary but the results are not comparable. These methods also address the estimation of the effective population size and our manuscript includes the comparison of effective population size estimated from temporal F_{ST} as a generally equivalent approach to current available methods (which are based on allele frequency changes through time).

Regarding the comparison between ABCRF and classical ABC, we think the results from the original articles describing ABCRF (cited in our manuscript) are quite clear about the advantages of ABCRF. The reduction of almost two orders of magnitude in the number of simulations required for inference was a major incentive to develop this work. In our opinion, trying to replicate the analysis in classical ABC is a waste of computational resources.

2. The authors state in the introduction that "Methods for demographic inference assume that most of the genome evolves without the influence of selection and that any deviation from the mutation-drift equilibrium observed in the data was caused by demographic events". I agree in the sense that the effects of linked selection are often ignored in such analyses, but there are many studies that attempt to disentangle selection/demography by dividing the genome into putatively selected sites and putatively neutral sites.

Such methods are derived from the PRF (Poisson Random Field) and/or MK (McDonald-Kreitman) framework and assume that demographic events shape patterns of variation at putatively neutral sites (often taken as synonymous alleles within genes) while both selection and demography shape putatively selected sites (sometimes taken as nonsynonymous alleles within genes). Patterns of variation at the "neutral" sites are used to fit the demographic model while the "selected" sites are used to fit the selection model. There are many such papers (e.g. one could start with Williamson et al 2005 PNAS and Keightley & Eyre-Walker 2007 Genetics, or a more recent paper such as Racimo & Schraiber 2014 Plos Genetics).

The realization that putatively neutral sites may not be sufficiently 'neutral' for such analyses (due to direct or indirect selection) has also been a long-standing topic of discussion, and some studies have sought to address this issue. For example, Gazave et al 2014 (PNAS) sequenced regions far from genes in humans to try to fit demographic models that were less affected by selection. Torres, Szpiech, & Hernandez 2018 (Plos genetics) found that background selection has a substantial influence on demographic inference in humans. Messer & Petrov 2013 (PNAS) developed a method to infer adaptation rate and tested its robustness to linked selection and non-equilibrium demography. There are is also at least one paper that seeks to infer recurrent selection and demography using a combination of ABC and PRF methods (N. Singh et al 2013 Genetics, "Inferences of demography and selection...")

The effects of non-equilibrium demography on selection inferences using genetic time series data have also been studied (E.g., Schraiber et al 2016 (Genetics) uses an MCMC-based procedure and compares equilibrium/non-equilibrium demography).

This is just a subset of the papers in this area, and it's not my intent to be prescriptive about citing particular papers. Overall, it would be very helpful for the reader if the authors could put their method into context with this body of work.

There is a wealth of different approaches and methods in the population genetics literature to address the inference of demography and selection and our article does not intend to cover all of them. We have focused on making links to the approaches that are more relevant with our work.

The methods that divide the genome into putatively selected sites and putatively neutral sites have some interest in our work because they are methods that target the inference of genome wide selection parameters. However, they are in stark contrast with the objectives of our work in that (1) we target the effect of “recent” selection (recent in the sense of selection occurring during the period of time in which the samples were taken) while these methods describe the action on selection over long periods of time; (2) we have no assumptions regarding different types of polymorphism categories in the data analysis (the model, of course has different types of polymorphisms) and (3) we make inferences both on demography and selection parameters from the same model while these methods make demographic and selection inferences sequentially (with biased demographic inference potentially affecting the selection analyses; e.g. Messer & Petrov 2013; note that their approach does not target improving demographic inference but making selection inference robust to demography misspecification). Because of these important differences, we do not consider that this type of method needs to be discussed in our work.

3. The authors focus on a model of beneficial alleles and neutral alleles. I found this a bit surprising, given that negatively selected alleles are likely to represent a much larger fraction of mutations and have a substantial effect on the frequency spectrum (and other summary statistics).

The authors discuss the impact of background selection briefly in the discussion, but I think further justification for the relevance of this beneficial allele model (or simulation of a model that includes deleterious alleles, or an argument as to why negative selection will have limited impact on the authors' summary stats) would be helpful here. It is true that some recent papers (e.g. Schrider, Shanku & Kern 2016 Genetics) have focused on positive selection models, but their purpose was to describe the effects of positive selection on demographic inference directly, rather than develop inference procedures. It seems unlikely to me that a model of beneficial alleles alone can provide useful inferences on real data that are certainly affected by a mixture of beneficial alleles and deleterious alleles. Alternatively, if there are specific insights that a beneficial allele model can provide then it would help to further highlight these. I do appreciate that the authors were clear about this limitation in their discussion section.

It is important to remember that our model consists of two distinct periods. The first period is used to generate a realistic genetic diversity in the population and the inference period that comprises the samples. The estimated parameters and latent variables are calculated on this second period in order to characterise recent selection processes. Most of the summary statistics that allow us to estimate the parameters and latent variables characterise changes in genetic diversity among temporal samples and the heterogeneity of these changes across the genome (e.g. quantiles and moments of the F_{ST} distribution, see figures S4 and S5, Supplementary Material). Therefore, we are more interested in the short term effects of negative selection on allele frequency changes than in long term effects on the site frequency spectrum. We discussed this aspect given the scarce literature (to our knowledge) available on this aspect (i.e. temporal changes in allele frequencies) of background selection. We think background selection is an important factor to be considered and, as discussed above, we would like to explore its influence in our future work.

Regarding whether a model with only *de novo* beneficial mutations can provide insights in the analysis of real data we recognise that, with current implementation, the practitioner should interpret the results with caution. We believe that for many cases a model of adaptation from standing variation could be more appropriate. If that would be the case, the meaning of the parameters and latent variables, estimated from the model with *de novo* beneficial mutations, might have a less straightforward interpretation. We would like to develop and test more complex models under this framework, but given the workload required and that we consider our current results interesting enough by themselves, we would leave them for future work.

-Line 71: Is the '10,000' here years? Or generations?

It should say ‘100,000 years’. Corrected.

-Line 80: Another paper that may be of interest here is A. Stern et al Plos Genetics (2019)

This is a very interesting point that highlights something that we have not discussed in our manuscript (because it is a bit tangencial to our approach). We argue that we are interested in the recent history of population and that is the reason to use temporal samples. However, methods that allow the estimation of the ancestral recombination graph have opened up the possibility to study recent history from single-time samples too. Stern et al. (2019) is a great example of that. Still, these methods require phased data and some previous knowledge on recombination rates, which are not widely available for many species (Stern et al., 2019, applied their method to human populations). However, we think that discussing such methods is not necessary for the understanding of our work and that will make the introduction less linear and clear.

-Line 90: AABC is also a way to reduce the number of simulations, see Buzbas & Rosenberg 2015 Theoretical Population Biology.

It is possible that the performances of ABCRF and AABC are comparable, but we have no experience with AABC and, to our knowledge, there is no publication presenting a systematic comparison of both approaches. ABCRF reduces the number of simulations by better exploiting the simulations in the reference table and AABC reduces the number of simulations by adding additional approximations, which,in principle, we would like to avoid. There are also additional features (in addition to the reduction in the number of simulations) of ABCRF that make it a preferable approach compared to AABC, such as the automatic choice of informative summary statistics or that ABCRF does not need to set a rejection threshold.

-Line 120: I'm a bit confused by this description of burn-in. Generally the purpose is to reach the dynamic equilibrium state that is determined by the parameter settings.

It should say “the initial simulation state” not “the initial parameters set”. This has been corrected.

-Figure 1: Why are there two different colors for the neutral mutations? Do they represent different things?

They represent neutral mutations in the neutral regions and in the selected regions, which is an unnecessary distinction. We have simplified the figure to avoid confusion.

-Selection is determined by a Gamma distribution here, which has two parameters. Only the mean of the distribution seems to be mentioned here, which seems to leave one free parameter. How are these parameters selected (perhaps this is mentioned and I missed it)?

The implementation of the gamma distribution in SLiM takes two parameters: mean and shape. In our approach the mean was sampled from a prior and we set shape=mean. The previous version of the text did not contain this information and it has now been corrected.

-Line 190: I think 'oscillated' may not be the appropriate word here, perhaps fluctuated?

Corrected.

-Line 231: 10 generations seems like a very short time? What is the rationale for this number?

We have in mind sampling schemes that would be realistic for a range of situations such as resurrection experiments, experiential populations or historical collection (museums, herbariums) compared to modern samples. In those contexts, 10 generations is a realistic value. Different ranges of ages are also explored for the analysis of bee populations.

-Line 293: The transition from discussing real to simulated data seems very abrupt here?

We have revised the text to make the transition more clear.