Reviewer #1 (Comments for the Authors (Required)):

In the article "Assessing the relationship of ancient and modern populations" by Joshua Schraiber, the author develops a new method to statistically test whether a set of ancient samples belong to a population that is directly ancestral to a population that a set of modern samples belong to. The method uses diffusion theory to derive the likelihood to observe a particular allele frequency y in the ancient population conditional on observing an allele frequency of x in the modern population. In addition, an explicit binomial sampling scheme with an error model is imposed to allow for analysing error-prone low coverage individuals with low samples sizes. The method is tested with simulations and applied to real data.

I think the article is very relevant and timely, and I see no major problems with theory, data analysis or the way it's presented.

**I am grateful to the reviewer for their favorable view of the manuscript and helpful comments.**

I still have some (hopefully minor) points to consider for improvement:

1.) On Figure 2: I am a bit unhappy that the reader is presented only with RMSE plots to assess the quality of the inference from simulations. I think it would be better to also present some simple "true vs. inferred" type of plots, where the reader can see how good or bad the inferred parameters are in comparison to the true simulated ones. One idea would be to put the different sampling schemes currently distinguished by colour along the x Axis and then plot the inferred values for t1 and t2 onto the y axis, with the true simulated values indicated as horizontal lines or so. That's just a suggestion, I'm sure there are other types of plots or tables to show this. The idea is that this way the reader gets a more direct idea of the relative uncertainty and in addition of a potential bias in the parameter estimates.

**I appreciate the reviewer’s concern about the ambiguity inherent in a plot of RMSE. However, because the method is approximately unbiased for large enough sample sizes and coverage, attempts to plot true vs. inferred result in overlapping lines only distinguished by error bars. Thus, RMSE plots overall are reflecting decreased sampling variance, with a minimal impact of bias. I think that as a main display item, the RMSE plots will be the most useful. To address the reviewer’s concern, I now provide a supplemental table breaking down the bias and variance for all sampling schemes.**

2.) On the method: I am actually surprised that there is much power to distinguish between t1 and t2. My intuition tells me that the sum of t1 and t2 should be very accurately estimated (as for example also by F2 statistics or Fst). Indeed, equation 2 and the definition of Q and Q\_arrow look almost the same, suggesting that the conditional probability to observe frequency y given x is approximately similar to a simple diffusion process connecting frequencies x and y over a time t1+t2. My prediction would be that the errors on t1 and t2 would involve a substantial covariance component that makes the errors on t1 and t2 correlated in order to keep the error on t1+t2 low. I don't want to suggest much extra work here, but perhaps there is a quick way to asses the covariance between these errors from existing simulations in a simple manner...

**I, too, was surprised that there is sufficient power to distinguish t1 and t2, and I agree with the reviewer that there is good reason to expect the sampling distribution of t1 and t2 to be negatively correlated. To examine this, I added a Supplementary Figure showing that the errors between t1 and t2 are negatively correlated. I also added to the text to address this point.**

3.) I think it would be extremely useful to not only do maximum likelihood estimates of the parameters but also give confidence intervals, for example by trying to compute a Posterior probability distribution or by bootstrapping. For computing the posterior: the likelihood function is right there, and a posterior probability over parameters t1 and t2, in fact the \_joint\_ posterior probability including above mentioned covariance is straight forward to write down and e.g. evaluate using MCMC. If that's too computationally expensive, a simple alternative to get the covariance and error would be a weighted block-jackknife.

**The review is absolutely correct that a measure of confidence is essential for any statistical method. In fact, the software does support nonparametric bootstrapping, although I did not include a demonstration in the manuscript. This is primarily because it would be extremely time consuming to do for the example of the Mathieson et al data considered in the text, which takes 3 days to analyze on 50 cores; to do 100 bootstrap replicates would take almost a year on the hardware I currently have. Hence, I have opted not to include a demonstration of bootstraps in the manuscript, and instead stated that bootstrap confidence intervals are available from the software in the Discussion.**

4.) In line 69, there is a reference to "?".

**Fixed**

5.) In the unnumbered lines at the bottom of p4: "[...] the ancient sample, given that it is at frequency x, " -> add "in the modern population".

**Fixed**

6.) In Figure 4, perhaps indicate the true values for t1 and t2 as horizontal lines. Also, please list the exact scenario in the legend. For Figure 3, a whole family of scenarios was simulated, so it's confusing to then refer to the very first simulation scenario again for Figure 4 without explicitly mentioning it.

**Added horizontal lines and clarified the specific scenario that is occurring in the main text.**

7.) In lines 194-197: "As expected, we see that low-power sampling strategies [...] are relatively unimpacted by ghost admixture". In my opinion this is a slightly misleading statement, and so is the legend text below Figure 5. It sounds as if rejecting continuity with ghost admixture is kind of an artefact or something that is to be avoided, and hence having low power in the ancient genotypes is a good thing. Clearly this is not what the author means. I actually think that under ghost admixture, a continuity test \_should\_ be rejected, and failure to do so is a bad thing (but explainable by low power). I suggest to rewrite this sentence plus the legend in Figure 5 to make clear that rejecting continuity is not "false" as indicated in the legend, but would be the "correct" thing here.

**I appreciate the reviewer’s comment on this point. I modified to text to be more value neutral in the Results and the figure caption.**

8.) In line 210: What does it mean if "t2 is significantly greater than 0"? Since there are no error bars, there is no way to assess significance, is there?

**The significance was assessed using the Likelihood ratio test described in the main text. I now make this more clear in the main text.**

9.) In line 213: "[...] the ancient samples must have existed for fewer generations since the [...] split" -> Bad wording, you mean lineages or populations, but not samples. Samples typically exist for 2-3 generations or so.

**I appreciate the reviewer’s assistance with wording this sentence, as I found it somewhat hard to express concisely. I modified the text to try to use more clear language.**

10.) In line 234: "suggesting a slightly downward bias in estimating t1" -> This could be directly tested in the simulations when plotted the way I suggested, i.e. plotting true vs. inferred values, instead of just the RMSE. Even better would be to also plot error bars from jackknife or the posterior.

**I modified the text to reference supplementary table 1, which shows no such downward bias. Instead, I comment that this may suggest some heterogeneity in each panel.**

11.) Finally, I think the discussion is too long, in particular in relation to the rest. I would try to remove some of the parts that just summarise what was done and focus on a three or four major points to discuss, but I admit that's somewhat up to taste.

**I appreciate the reviewer’s stylistic criticism, but I opted to leave the discussion largely as-is. I come from a school of thought that says that you can never reiterate a point too many times in a scientific manuscript ☺**

Reviewer #2 (Comments for the Authors (Required)):

Schraiber's manuscript "Assessing the relationship of ancient and modern populations" presents a new method for inferring certain demographic parameters taking into account special challenges of ancient DNA. The main advance is a methodological framework for performing inference based on raw read data with errors, using results from diffusion theory. Overall I thought it was an interesting paper, but I had a few comments and questions on some of the details.

**I appreciate the reviewer’s positive assessment and constructive comments.**

1. Regarding the genetic drift results and sampling equations, first, I was wondering if there would be an impact from new mutations? The author mentions this on line 364 in regard to mutations in the ancestral population, but I believe some of the polymorphic sites in the present-day population might be new too? I could be wrong, but my interpretation of the appendix was that the sites were assumed to be polymorphic in the common ancestral population (frequency z), but in fact some of the mutations could be "lost" (going backward in time) more recently than the split time.

**The derivation done in the text is exact for sites that are known to be segregating in the ancestor. The reviewer is correct that actually ascertaining these sites can be somewhat difficult; however, the results from the simulations show that that this has little practical consequence; this is likely because on the time scales of ancient DNA, new mutations are likely to be low frequency, and those would be unlikely to show up in the ancient sample anyway.**

Second, the author writes on line 251 that the method is robust to ascertainment in the modern samples, but I was confused as to how this is the case. The derivations (lines 399-412) make use of the equilibrium 1/x form of the frequency spectrum, which is probably a reasonable assumption for the set of all polymorphic sites in the genome, but in practice one often works with a subset of sites for which the spectrum may be different. For example, if looking at relatively common SNPs from a genotyping array, then in a population related to the one used for ascertainment, the alleles with high derived frequencies would not obey a Wright-Fisher process going back in time.

**I appreciate the reviewer’s point that in the previous draft of the manuscript, I merely asserted that there was robustness to ascertainment, rather than providing any evidence. In the revision, I added an appendix containing a proof of robustness to ascertainment in the modern population. Obviously, the method is not expected to be robust to ascertainment in the ancient population or in populations more closely related to the ancient population than they are to the modern population. I added to the discussion to better reflect this fact.**

2. When dealing with genotype errors, does the method take into account the fact that post-mortem damage produces certain specific error patterns (especially apparent C-to-T substitutions)? If not, does that skew the results at all? I was also wondering about contamination, which is mentioned in the introduction but then not addressed as far as I could tell. It might be good at least to run simulations in which a small to moderate amount of ancestry from the present-day population is artificially introduced into the ancient data to see what effect that has on the inferences.

**I agree with the reviewer that there are several additional complications to the error model that are not considered in this paper. To address concerns about contamination, I performed simulations with contamination. I now add a supplemental figure showing that contamination from the modern population will result in a downward bias in both t1 and t2. As noted in Racimo et al (2016), the impact of including a separate parameter for C-to-T postmortem damage induced mutations is minimal. Including explicit models of these sources of error will be left to future work.**

3. The result about greater power for a larger number of relatively low-coverage individuals is interesting, but one point that I think should be added at least in the Discussion is that not all ancient samples have equal quality or are equally feasible to sequence to a given level of coverage. For example, if one has two samples from an ancient site with equal endogenous DNA content, then (at least for some analyses) sequencing both to 1x would be more cost-effective than sequencing one of them to 4x, but if the first sample has much higher endogenous DNA levels than the second sample, it could be more efficient instead to sequence only the first one to greater depth.

**I appreciate the reviewer’s insight and updated the text accordingly.**

4. The scenario of post-split admixture into the ancestors of the present-day population (bottom half of page 9) seems likely in a lot of historical examples and is perhaps even more generally applicable than the primary model in the paper of a split from a common ancestor with no subsequent admixture. In the main text, this section is framed in the context of robustness to false rejection of continuity, but arguably if the modern population has 10% or more ancestry (as in the examples) that wasn't present in the ancient population, then the two groups are not truly continuous, regardless of the post-split genetic drift. Similarly, for someone studying the history of these populations, such admixture might be the most interesting phenomenon, and one might want a method that can detect the new ancestry as a positive feature. I appreciated the fact that these points were raised in the Discussion, but I think it would be worth considering whether the main exposition might perhaps be tailored more along these lines.

**I agree with the reviewer that these situations of ghost admixture are of great interest to the community. I am actively working on approaches to detect such situations. However, this manuscript is mainly focused on situations where direct continuity is the question of interest, which it is in many cases. Hence, I opted to leave the main exposition as is, with a strong focus specifically on rejecting continuity and not testing alternative models.**

5. For the real data, the result of t2 >> t1 for almost all ancient populations does seem very surprising. The hunter-gatherers plausibly had low population sizes, but most of the samples come from agricultural groups, which one would expect to have had fairly high population densities (as mentioned in the last line of the abstract, although as currently stated, this appears only to be referring to the present day). We also have some existing knowledge about the relationships between some of these ancient populations and present-day Europeans; for example, from Haak et al. 2015, CEU could be modeled as a mixture of Neolithic European farmers and Steppe pastoralists (both of which are represented in this paper). As noted by the author (lines 179-180 and 282), admixture can lead to increased levels of heterozygosity, and that process could perhaps be contributing to the lower estimated drift time in the present-day CEU (and generally the lower t2 values as a function of time as in Figure 6C). Additionally, to the extent that some of the ancient samples seem to be fairly good proxies for one component (but not all) of present-day European ancestry, the split model may not be an ideal representation of the relationships (as in comment #4 above).

**I expanded the Discussion to reflect these points.**

Another possible factor, as described in comment #1, could be ascertainment bias. I believe the set of 1.2 million SNPs being analyzed contains as a subset the sites on the Affymetrix Human Origins Array, of which one panel consists of SNPs ascertained as heterozygous in a French individual. Such sites would be expected to show high heterozygosity in CEU. An empirical test could be to restrict the analyses to the Human Origins panels ascertained in Yoruba and/or San, which would be (approximate) outgroups to the populations in this study, and see if the results change at all.

**Because of the proof that the methods are robust the ascertainment, I opted not to perform this test.**

Overall, I think alternative explanations should be considered more thoroughly in addition to that of universally low population sizes (e.g., line 215).

6. Finally, two minor comments: on lines 69 and 361, there are compilation errors with citations; and in equation 2, in the exponents, is that simply multiplication between the Q matrices and the scalars t1 and t2?

**Yes.**