Reviewer: 1

Comments to the Author

The goal of this work is to develop a method for assessing the relationship between ancient sample and modern populations. It is designed to accommodate some of the challenges associated with ancient DNA sequencing by modeling fitting a population split model between one ancient and one modern sample.

Major Concerns:

This paper would be substantially improved if biologically novel conclusions could be drawn from the data. This would showcase the power of this method and improve the appeal of this work for broader audiences. Minimally, some expanded discussion and interpretation about the specific t2 estimates obtained (figure 4) should be included.

**Spend some more time on this? Maybe it’s worth brining up the bound on ratio of the effective sizes (t1/t2)**

Line 115. That the error probability is identical between derived and ancestral sites is problematic. Given that the central purpose of this method is to analyze lightly sequenced ancient DNA sequencing datasets, and that errors are very often nucleotide specific in aDNA, this error model should be updated to accommodate these additional considerations.

**A->D and D->A error rates? Two more parameters per sample… could also just maybe reference Fernando’s finding that it’s irrelevant.**

Line 130. This seems like a strange choice for simulating the dataset. I realize that diffusion will not accommodate linkage, but the goal is to apply this to real datasets—and especially human datasets with small Ne. Linkage is going to be really important and widespread across the genome in these cases. Therefore, to evaluate this method for its intended application requires simulation of a realistic linkage model. Msprime can certainly do this, so it’s a little strange that the author has decided to generate an enormous set of unlinked loci instead of realistic chromosome scale simulated genotypes.

**Redo simulations by simulating whole chromosomes and subsampling to the density of a SNP chip? Maybe leave this for now, see if it comes up again.**

Minor Concerns:

Line 88. As I understand it, the software is accepting an allele frequency for the modern populations. However, allele frequencies are not known with certainty, it could be preferable to model uncertainty in modern allele frequencies as a function of genotype counts in modern samples. This is likely to be particularly problematic for modern populations that are lightly sampled where substantial error in allele frequencies estimates might be expected to swamp any signal of allele frequency change.

**Gotta put in the stuff about the beta prior. I guess this is important.**

Line 124. I encourage the author to share the scripts used to simulate base calling errors as either supplemental files or within the github repo.

**It’s there, just be clearer.**

Line 168-175. How would inbreeding effect these inferences, and is there evidence for inbreeding from within these ancient samples?

**Removed known related samples. Inbreeding should look like bigger t2, which is really all drift is...**

Line 364. The github repository could use a little cleanup. There’s a ton of files in there and most appear to be unnecessary. Given that this is a methods paper, I consider the repo a part of the paper and it is a little challenging to review in the current state. I also highly encourage the author to include a complete example dataset so that users can confirm that they are running the program correctly.

**Okay. Reorganize the repo, this needs to happen anyway.**

It seems likely that many users of this program would be interested in estimating the confidence intervals for t1 and t2. Can the author suggest how this can be done? Perhaps this could be included in this software package?

**Bootstraps. Don’t want to do this for the whole dataset, because it’ll take eternity… maybe figure out a subset of samples to demonstrate a bootstrap with?**

Figure 4. I know very little about these specific samples. Nonetheless, there are some fairly big outliers for t2 in panel A. It would be worth discussing these outliers and plausible explanations for the large drift estimates.

**These could be the ones that I do a bootstrap with?**

Figure 4. Axis labels and text in the key is really small. I recommend a little figure reformatting to get this whole thing cleaned up.

**Need to fix**

Reviewer: 2

Comments to the Author

The manuscript describes a diffusion theory-based maximum likelihood method that can be used to test a relationship between an ancient population (represented by a set of ancient DNA samples) and a modern population. This is a field of interest to many, and the major advance is the ability to jointly analyse data from aDNA samples from a single population by integrating over the available sequencing reads at each genomic position of interest. The results show that having multiple aDNA samples at low coverage is often more valuable for population genetic inference than having a single sample at a high coverage, a useful insight. The applicability of the method is demonstrated on a previously published set of aDNA samples from various ancient West Eurasian populations. The author argues that none of the tested populations represent direct ancestors of present day individuals of European ancestry.

Current analysis of ancient DNA generally relies either on obtaining a high coverage genome, which is often not feasible, or so-called “random read sampling” of low coverage data. The later treatment of low coverage samples avoids the difficulties of genotype calling, but discards a significant portion of the available sequence information. This manuscript points towards a critical direction for future work in aDNA population genetics. The results might also help in decisions on prioritizing aDNA sampling, as it is an open question whether it is better to try to obtain multiple samples, even at very low coverage, instead of trying to generate a single higher coverage sample.

However, while potentially being of high interest to researchers in the field of aDNA studies, the manuscript in its current form suffers from a few flaws that make it difficult to assess the reliability of the estimation procedure, particularly how would it perform in practice.

Major comments

1. The method models sequencing errors by estimating a single error parameter, described as “base calling error”, which generally refers to errors in next-generation sequencing (NGS). However, the author neglects a larger error characteristic to ancient DNA – post-mortem damage (PMD) substitutions. These aDNA false substitutions often have orders of magnitude higher rates than other sequencing errors and have dramatically different characteristics – they are position specific, occur more often towards the ends of the reads and will likely be correlated (if a forward read carries a PMD C->T substitution at a given genomic position, another forward read is more likely to carry a C->T too). This makes PMD substitutions a non-trivial issue and have to be modeled explicitly (for an example see “Inferring Heterozygosity from Ancient and Low Coverage Genomes” by Kousathanas et al.) and its intricacies cannot be captured by a single parameter. It would be best if the model addressed this in some way, but barring a change to the model, the manuscript certainly should acknowledge the issue, and the two sources of error.

**I think I need to just pay lip service here, and maybe cite Fernando. I don’t mean for the parameter to be just sequencing error, but rather also PMD. Clarify that.**

Regarding the proper modeling of the aDNA PMD substitutions – it is important that the author demonstrates that his results on the data from Mathieson et al. are not confounded by aDNA damage. A simple test would be to re-run the analysis on a subset of sites, filtering out SNP positions that might be influenced by PMD false substitutions (i.e., restricting to transversions).

**Do this.**

Of particular concern, incorrect modeling of error would be expected to inflate t2, which is the primary “surprising” result of the manuscript.

2. In the first results section the author states that he used “scripts to simulate base calling error and low coverage data,” but it would be good to have a more thorough description of the simulations. Did he use sampled sequence data generated with msprime and then simulated NGS reads from those, followed by introducing random sequencing errors? It would be good to include a paragraph describing the generation of low coverage reads and sequencing errors in more detail. Again, regarding to point 1. above, it is important to address the difference between base calling error and aDNA damage patterns.

**Be more specific, the other reviewers have a similar complaints that the simulations are not obvious**

3. In the section describing analyses of the Mathieson et al. data set, the author makes several claims that are not always intuitive and it would be helpful if he could elaborate or provide a more detailed explanation. For example,

a. In 8/172 “[…] t2 >> t1, despite the fact that the ancient samples must have existed for fewer generations since the population split than the modern samples. This suggests that all of the ancient populations are characterized by extremely small effective population sizes.” and also “we see strong positive correlation between t2 and sampling time” in 9/185

– It seems reasonable to assume that the “true” ancestors of present-day human samples had a similar effective population sizes as the ancient populations from Mathieson et al. data, at least during a similar time period. Isn’t it expected then, that the lineage towards present-day individuals should experience similar amounts of genetic drift during this time?

b. In 9/183 “the relationship of these populations to the CEU is complicated and not summarized well by the age of the samples.” – It would be helpful for the reader if the author elaborates what demographic factors might complicate the relationships.

**Be more clear about this. I think that a Ne ratio discussion would be helpful.**

4. It would be interesting to see or discuss how the results are affected by alternative demographic models.  For example, a model as described above where both population one and population two experience the same bottleneck after they split, with a subsequent expansion in the modern population

1.  Alternatively assuming the ancient population is a true ancestor of the modern population, how would admixture from another, unrelated population, into the present-day human linage after the time of the ancestral ancient population affect the inference? Perhaps most important, what if the ancient samples, which are assumed to come from the same population, are not actually from a single population.  Presumably, this would also artificially inflate t2.

 This, combined with the issue with error and t2 from point 1, makes me hesitant to accept the authors strong claim of rejection of direct lineage from the ancient populations.  At the very least, unless these issues can be adequately addressed, it would be good to give caveats to this claim.

**Do two simulations: 1) demographic crap, 2) admixture. Probably for resubmission, admixture is most important. Do two different tests:**

**1) Ghost population into modern population, but ancient sample is truly directly ancestry (maybe plot p(reject) as a function of ghost admixture proportion?)**

**2) Ancient sample into modern population. This is where I expect that t1 -> 0 as admixture gets bigger, but t2 is still pretty easy to interpret.**

5. The author should mention what kind of data the method is applicable to. An ancient population and a modern population only? How would it perform on two ancient populations? In particular, what degree of separation is allowed between the two studied populations? Could this work using modern humans as a reference, and low coverage Neandertals as the ancient population (presumably not).  This should be clarified very early.

**Be clear that this works on Neandertals…**

Minor comments

- While both are generally synonymous, using “base-calling error” ( 2/43, 4/5, etc.) and “sequencing error” ( 1/33, 9/30, etc.) in the same text might be confusing to the reader.

**Be much more specific about this stuff**

- How sensitive is the method to the estimate of x in the “reference” population? Is the method applicable even in a case of comparing the relationships of two ancient populations, where the estimate of x wouldn’t be necessarily as accurate as it is the case in modern populations (even to the extreme of x being inferred from a single high coverage genome)?

**The beta stuff…**

- I understand that Python and popgen literature in general is not exactly famous for having the most beautiful graphics, but it would be nice to improve the aesthetics of the plots a little bit (Figures 2 and 3 specifically). Also, please increase the size of text in Figures 4, 5, 6 – it’s quite difficult to read.

**Do this, all the reviews commented on how shitty the figures are**

- Should the p\_{n,k}(x\_l) term in the equation (3) be a capital P from equation (1)?

**Check this**

- It is not clear what a capital “I” means in equation at 6/110.

**Check this**

- It is not clear what the indices i, j mean in the description of the two Q matrices.

**Check this**

- It would be useful to contrast this method more with the method of Racimo et al – why does this work with low coverage data, and the previous does not?

**This is an extension of Racimo’s method, need to be explicit about that.**

- In the introduction, the author is perhaps overly critical of other ancient DNA methods.  It is clear that advances are necessary.

**lol**

Reviewer: 3

Comments to the Author

This manuscript presents a method for determining whether one population (represented by an ancient DNA sample) is directly ancestry to another (represented by a modern sample). The paper is well written and very easy to understand, and the theory is very clearly presented and easy to follow. However, I have a number of concerns about the paper – largely about the practical usefulness of the method, and about the interpretation of the real data, and I think the paper would need an extensive overhaul before being suitable for publication. More extensive tests on different datasets would add to the paper.

Issues with the design of the test:

1) My first concern about the design of the method is that the null model – that one population is truly ancestral to another – would almost never be appropriate. I doubt there’s ever a real case where t\_2 is exactly zero. Therefore it seems like it would make more sense to reinterpret the test qualitatively – asking how much of the drift is shared, rather than whether it’s all shared, which is always false. As it stands, I do not think it has been shown that the method is of any practical use, since I think it would always reject the null in practice (unless it was underpowered because there was no data). Why not try the method on an example where t\_2 is claimed to be zero (for example the cited example from Rasmussen et al)?

**I guess I should include the Anzick sample, but that is annoying. Do I have any decent size Sardinian samples? Could check them against hunter gatherers…**

2) It seems like the test could be phrased in terms of f3 statistics, specifically whether f3(Outgroup; Pop1, Pop2) = f3(Outgroup; Pop2, Pop2). Is that fundamentally different to this approach? Uncertainty about genotype etc could be interpreted into the f3 statistics as well (although I’m not aware of a method that currently does this).

**Blech blech blech. Think about this. I don’t think it’s true, because the T\_{1,2} term would be different. You’d expect even if P1 were ancestral to P2, then T\_{1,2} > T\_{2,2}, because the sample from P2 has to wait before it can coalesce.**

3) If I understand correctly, each site is treated as independent, which is clearly not appropriate for the density of sites used in the paper. It seems like the likelihoods would thus be anticonservative.

**Another linkage complaint… think about some words about this.**

4) Is it really appropriate to treat the modern allele frequencies as known? For small allele counts the uncertainty is very large, and it seems like the difference between seeing 0 or 1 (or 2) copies of a given allele in the modern sample would make an enormous difference to the result.

**Beta stuff**

5) Does the method have power to separate out the effects of drift and sequencing error/contamination? Is there a correlation between the error rate estimate and t\_2? I don’t understand intuitively how that’s possible but I could certainly be wrong about that.

**Check this on the real data; I’m doubtful. It definitely works on simulated data…**

6) Please provide more information about the simulations. In particular, how uniform is the coverage? For example, if heterozygotes are important, then seeing half the sites at 2x and half at 0x is better than seeing all the sites at 1x. In reality, the coverage of these data tends to be very nonrandom and over dispersed compared to Poisson, for example.

**Clarify the simulations.**

Issues with the interpretation of the data:

7) The conclusion is that none of the tested ancient populations is directly ancestral to modern populations is unsurprising since all the modern populations are admixed – including from populations close to those ancient populations tested. It’s not really correct to see these ancient populations as splitting off the “main stem” CEU lineage – rather CEU is formed as a mixture of ancient populations.

**Right, this makes CEU a complicated situation, as Fernando pointed out. Sardinians would be more ideal…**

8) The observation that t\_2>>t\_1 could also be interpreted as a higher error rate in the ancient samples. Simple test: the error rate at transitions (specifically CpG sites) should be much higher than the error rate at transversions, so split the data in two and check. This would also be a nice demonstration that the method has power to separate out the effects of sequencing error and drift. Suggested sanity check: Test data from a modern population (e.g. a few subsampled 1000 Genomes FIN/GBR/IBS/TSI samples) and see if t\_1=t\_2.

**No no no, t2 would go DOWN with more error. T2 is influenced by heterozygosity! Need to do the transition/transversion masking thing, though.**

9) It seems like the correlation between age and estimated t\_2 is largely driven by the four oldest hunter-gatherer samples. But that would make sense, because those are the populations that have low diversity according to Skoglund 2014. I’m not sure that the other (late, southern, Neolithic) populations are expected to have particularly low diversity / small population sizes compared to present-day Europeans.

**The correlations are not useful, get rid of them. Just confusing.**

Other minor comments:

- Line 38 – “modern Native Americans are directly descended from the Clovis culture that inhabited North America over 10,000 years ago. This is not correct, at least not for all Native Americans: See, for example Reich 2012, Skoglund 2015. Maybe “some Native Americans”? Also maybe “people associated with the Clovis culture”, since you can’t descend from a culture.

**Fix.**

Fig. 1 legend – specify units of date (“years before present”?).

**Fix.**

Fig. 5; can you split this error rate up into transitions vs transversions?

**Split the data up and see what happens?**

In general is it approporiate to compute p-values for correlations between the points (e.g. age vs t\_2) assuming independence, since they are not independent (related populations share some drift).

**Good point. Get rid of the correlation.**

Line 58 – Are these approaches actually more affected by sequencing error and contamination than other approaches?

Line 50 – I think Green et al 2010 (for use in f statistics etc…) and Skoglund et al 2011 (for use in pca etc.) are better citations for pseudo-haploid data, but I take your point that the other citations show that it’s common.

**Add these citations**