Answers to the issues raised by the referees on Inference of recent admixture using genotype data

February 11, 2021

We thank both reviewers for their careful reading of the manuscript. Below are some clarifications. We start with some points which led to larger rearrangements. In the text, new text is in .

Reviewer 1

1. The paper claims to do a significance test, but this is not apparent. And should be further justified. I believe this can be done along the lines in the attached document.

2. Intro
   p1 l8: ADMIXTURE should be succeeded by [8]. I guess.
   p2 l20: [15,15]:=[15,25]
   p2 l3,4: I am not acquainted with the word leverage. But probably OK.
   Changed, and exchanged one instance of leverage.

3. Mat and met
   p3 l2: I would add the R-package LEA.
   Done. Section 2.1
   is unfriendly. It is a good practice to distinguish between vectors and numbers - where vectors are typically in bold. In particular q_k vs $\boldsymbol{q_n}$ should be distinguishable. But also $\boldsymbol{G}$, where I have a good guess of the meaning.
   It is also a bit overconfident to claim that the maximum is reached by the sketched algorithm. Do you prove that it converges? And in case of convergence - can we rule out, that we reach a saddle point/local maximum? At least you need to argue, that you have some empirical evidence supporting that the algorithm seems to work.
   We refrain from using bold letters for vectors, since this is (i) not standard in all fields and (ii) rather an editorial decision. (Nothing is said about this in the authors guidelines of FSI:G. If the editor decides this is the way to go, we will alter our notation.)
   You are right, we have no formal proof of convergence to a global maximum. We spelled this out more carefully now. At least, in all our numerical examples, convergence always happened.

4. Section 2.4
   (6): This is not a common way to evaluate the distance between distributions. Usually Kullbach-Leibler or Brier is the choice. What is the motivation for this particular choice? It should be explicit that e.g q_k is an estimate, typically by the notation $\hat{q}_k$. Which leads to my next point.
   blue

5. Why don't you consider an information criterion? E.g AIC(Akaike). E.q for the admixture model
   $$-2(\log(\hat{q}) - \log(q_{TRUE}))) + (K-1)$$
   and recent
   $$-2(\log(\hat{q}^M, \hat{q}^P) - log(q^M_{TRUE}, q^P_{TRUE})) - 2(K-1).$$
   This is the standard way to compare models.
   blue

6. Section 2.5
   This is the Achilles heel of the paper. Later on you claim to do a test, which you clearly renounce on in this section. Why?? Standard asymptotic theory says that $-2\log(\Delta\ell)$ has a chisquare distribution with $(K-1)$ degrees of freedom. That should have been investigated. E.g. by simulation. If the approximation is bad, there is still an alternative. $-2\log(\Delta\ell)$ is a sum of M independent contributions, so it is an obvious alternative to do a Wald test.
   blue

   Section 3
   I am not confident with the conclusions. I would like to see analyses based on information criterion and significance tests and not arbitrary numbers. E.g. in section 3.5, where "highly significant results" are reported. On p10 l9: "The likelihood ratio test indicates that ..". What test? If you consider the logLRT test statistic it seems to be 6.747? Which is insignificant in a chisquare with 3 degrees of freedom. Maybe this approximate distribution is inappropriate, but this really needs to be investigated.

   Section 3.3
   I dont like the comparison of errors. A more general model is expected to have a lower error. Again, I would prefer e.g AIC.

   Section 3.4
   As described in MM??

   Section S1
   p1 l3: probabilitiy I was really annoyed by all the α's and would stick to the apparant model (S2), where I would be more explicit and use $q \cdot p_m$ instead of $\beta_m(q)$. But maybe the notation serves its right in the RA(recent admixed) model.

   Section S2.1
   Hard to evaluate. RA should perform better as it has $(K-1)$ more degrees of freedom. But is

it a significantly better performance?

Reviewer 2

1. The authors mention biogeographic ancestry (BGA) multiple times and present it as their main motivation for their work. However, it is nowhere explictly defined in the manuscript. The authors need to define what exactly they consider as their aim. Is it a classication problem with a predefined set of classes (i.e. ancestry groups) or is it a refined modelling of ancestry proportions from a given (or yet to be inferred) set of ancestry classes?
   Correct. The second paragraph of the introduction was meant to discuss exactly this point. We rephrased the corresponding sentences to make this even clearer. In addition, we added a sentence to the only instance of a classification task in the manuscript, which is in Section 3.1.

2. The mathematical apparatus is well developed, in sufficient detail and without any obvious mistakes. Given the readership of the journal, it may be helpful to move formulas (1)-(5) to the Supporting Information and instead give a non-technical description of the main ideas of both the 'admixture' approach and the 'recent-admixture' approach in the main text. Besides, it should be made clear that formula (1) refers to a haploid system (i.e. a single chromosome), where (2) refers to the diploid genotypes of independent markers.
   We understand that our manuscript has several mathematical aspects. However, when moving all formulae to the SI, we will lose readers which are just as interested in a rough idea on the formulae, but not as much as looking into the SI. The readers not interested in the formulae will skip them anyway, if they are in the main text of in the SI. However, we added some explanatory sentence at the beginning of Section 2.
   We added haploid versus diploid calculations in (1) and (2).

3. Minor issues with the mathematical notation are: (a) The side condition of $\sum_k^K q_k = 1$ should be mentioned with formula (1). (b) The notation of α_{mkl} in formula (S1) is confusing with the simulatenous use of α_{m1k} and α_{m2l}. Perhaps using α^1_{mk} and α^2_{ml} would make the notation easier to read.
   Ok, done.

4. Furthermore, the assumptions made for the modelling should be spelled explicitly out, both in the main text and the appendix. This includes the use of autosomal data only, random mating between individuals, no linkage disequilibrium between markers, but also homogeneity within ancestry groups; perhaps even more.
   We added some sentences at the beginning of Section 2.

5. The authors use a particular distance measure (section 3.2) to compare methods in their simulations. However, it is nowhere introduced, just cited. Given the central role of this measure, the authors should introduce their quality measure explicitly in the manuscript and give an in-

terpretation and examples for it.

6. A fundamental issue not with the mathematical approach of recent-admixture itself but its application and interpretation is the assumption of homogeneity within ancestry groups. The authors use Hardy-Weinberg proportions in their approach and interpret any deviation from it as evidence for past admixture between individuals from different ancestry groups. However, the 1000 Genomes groups are not necessarily so homogenic. At least the African (AFR) and South Asian (SAS) groups feature substantial internal heterogeneity and clinal allele frequency changes. The question is then if the proposed method picks up effects of past admixture or just group heterogeneity. This affects in particular the results presented in Figure 2. Furthermore, the proposed method could actually infer admixture where no admixture has taken place, i.e. result in false-positive results, just due to internally heterogenous ancestry groups. The authors need to discuss this and should also perform simulations that can differentiate between these two effects.

   xxx todo. Note that substructure would lead to an excess of homozygotes whereas recent admixture leads to an excess of homozygosity.

7. A further issue that has not been appropriately addressed is the choice of the number of ancestral groups (K). An inherent limitation of the admixture model is the need to choose a value for K before the analysis. Since the true value is usually unknown, this renders the application of the admixture model explorative. It is unclear how misspecifications of K affect the 'recent-admixture' model. The authors should either present theoretical evidence that this is not an issue or perform simulations that their approach is robust against the choice of K.

   We stress that in our application, K cannot be chosen. The reason is that the reference database just has K different origins of their samples. Since we are only dealing with the analysis of a single *trace*, we are not doing a classical analysis with STRUCTURE. We added some lines and a reference to the discussion.

8. The simulation results from section 3.1 are not very convincing. The authors look at only 10 ancestry-informative markers, while at the same time assuming a 'sky-rocketing' human migration rate of 2.5%. For a population of 10 millions, this would imply a generational influx of 250,000 individuals. No wonder the performance is rather poor. The authors should repeat this simulation with more realistic, justified values.

   Ok, there are some misunderstandings here. We take a *sample* of 400 individuals per deme, where the actual population size is way bigger. In fact, assuming a per-site per-generation mutation rate of $\mu \approx 10^{-6}$, we take $\theta = 4N\mu = 1000$, leading to $N \approx 10^9$ (haploid) individuals. (This θ produces the amount of SNPs reported in the paper.) So, the probability that a single (diploid) individuals migration in one generation is $20 \cdot 10^{-9}$, which is rather small. All this comes with the standard setting in population genetics and coalescent theory.

9. In the Discussion (p. 13), the authors claim that STRUCTURE would represent the gold-standard in BGA inference. This is simply not true. Again, the authors are not clear if they pur-

sue a classification-like approach (categories) or the estimation of ancestry proportions. These are two different concepts. The authors need to clarify the aim of their manuscript.

10. The authors distinguish between paternal and maternal ancestry in their in silico data generation of admixed individuals (section 2.3 p. 4-5). Since the presented method considers only autosomal data, it is not clear why this distinction is necessary. Please clarify.