Comments by recommender:

A key concern raised by all three reviewers is the difficulty in reproducing your simulations and results. To address this, the authors must provide the complete code, with sufficient commentary, along with the specific options used in your pipelines, to ensure that the study can be fully replicated.

We agree that we did not provide all details in the previous version of our manuscript. We have now explained parameter settings and justified our choices. Additionally, we have included the missing code into our github repository.

Additionally, the justification for focusing solely on ascertained biallelic sites in your study—particularly in the context of paleogenomics—needs to be strengthened. A more general comparative analysis that includes all sites, and contrasts with tools such as snpAD, might provide a more comprehensive understanding.

Restricting to ascertained biallelic sites is very common in paleogenomics, due to the difficulty of acquiring any sequencing data and the high error rate due to post-mortem damages and contamination. Moreover, SNP capture arrays are commonly used in ancient DNA, resulting in a natural situation in which to condition on ascertained biallelic SNPs. We added a statement to the introduction to justify it better:

"Restricting to known biallelic SNPs is a common practice in the population genomic analysis of ancient DNA data as low-coverage and post-mortem damage usually limit the possibility of calling novel SNPs for most individuals (see e.g. Günther and Jakobsson, 2019), and methods like snpAD are restricted to very few high quality, high coverage individuals (Prüfer, 2018). Instead, most studies resort to using pseudohaploid calls or genotype likelihoods at known variant sites (Günther and Jakobsson, 2019); using ascertained biallelic SNPs is particularly relevant when ancient DNA is enriched using a SNP capture array (Rohland et al., 2022). This choice also allows us to estimate mapping bias locus-specific rather than using one estimate across the full genome of the particular individual."

Since your approach seeks to mitigate mapping bias effects, it is also important to consider how varying mapping quality (MQ) values impact the analysis. In this study, MQ was fixed at 30, but assessing different MQ thresholds may provide additional insights into the robustness of your findings.

We have added information to the discussion section on other studies which explored different mapping quality thresholds. Additionally, we repeated a subset of our simulations (with depth 0.5X) employing a cutoff of 25, which, consistent with the literature for bwa aln, reduces mapping bias substantially. A cutoff of 30 remains widely used, however, so we keep this as the underlying cutoff for most of our simulations while we highlight that bwa aln users would be better off using a cutoff of 25. Finally, we highlight that different mapping tools have different ways to calculate mapping quality, meaning that they are not directly comparable while an adjusted genotype-likelihood can be employed for any mapper and any quality threshold.

In the Results section, it would be beneficial to introduce clear subsections to distinguish between mapping biases in simulated data versus empirical data, and between the estimation of admixture proportions in these different contexts. The current structure may give the impression that empirical data and simulations are used arbitrarily across in analyses.

We have included the type of data used in the sub-headings of the results section.

Finally, the Discussion section highlights that the improvement brought by your algorithm is relatively modest in comparison to other challenges in inference. As some reviewers have suggested, the Discussion would benefit from a more detailed exploration of the advantages of using this algorithm, including its computational cost, and the specific conditions or analyses where it proves most effective.

Thank you for this suggestion, we added: “In contrast to other approaches to alleviate mapping bias, such as employing pangenome variation graphs (Martiniano et al 2020, Kopetkin et al 2023), it does not require establishing a separate pipeline. Instead, only reads mapping to a set of ascertained SNP positions need to be modified and remapped which only represents only a fraction of all reads and consequently will require a small proportion of the original mapping time. Our Python scripts used to calculate the genotype likelihoods could be optimized further, but this step is of minor computational costs compared to other parts of the general bioinformatic pipelines (~1 minute per individual in the empirical data) in ancient DNA research. The corrected genotype likelihoods can then be directly used in downstream analyses using the same file structures and formats as other genotype likelihood-based approaches.”

Reviewer 1:

Mapping bias poses a significant challenge in the analysis of ancient DNA data. This study introduces testable hypotheses that address the impact of mapping bias on allele frequency estimates and admixture proportion estimation, particularly in ancient DNA research. By testing the effect of mapping bias, the study clearly demonstrates its influence on allele frequency estimation in empirical data. The corrected genotype likelihood approach shows the best correlation with “true” allele frequencies. The research further shows that while mapping bias can substantially affect ancestry proportion estimates, the adjusted genotype likelihoods can mitigate this issue. It also emphasizes the critical role of method selection, with some methods exhibiting considerable variability in results. These findings help refine methodologies in the field, making it possible to obtain more reliable results from low-coverage ancient DNA data and thus moving the field forward.
Global impression
The article makes a valuable contribution to the field by introducing a novel method for reducing mapping bias in ancient DNA analysis. It effectively outlines the problem and current challenges, with the proposed approach appearing both innovative and promising. The use of high-quality SNP array data adds value, as it provides a reliable control. Although it would have been interesting to see the effect of mapping bias on real data, the decision to simulate admixture data seems like a good choice to address this. However, while the impact of the corrected genotype likelihood on allele frequency and admixture estimation is significant, it looks very minor when compared to the standard genotype likelihood method. A more detailed biological interpretation of these results would be helpful to clarify why the modified genotype likelihood only has such a modest effect on mapping bias. With this in mind, a discussion of other potential sources of biases is still lacking.

We thank the reviewer for this assessment and we agree that certain signals were not discussed in detail. We expanded our discussion section substantially. We list more specific changes as response to the corresponding comments below.

Overall, the study provides valuable results to address the initial research question, but further investigation is needed to fully explain and contextualize these findings.
Major comment

Introduction

The introduction is well-constructed, providing a clear understanding of the challenges associated with mapping bias, the current strategies proposed to address these issues, and the new approach for mitigating mapping bias and assessing its impact. However, given that this project focuses on ancient DNA, it would be beneficial to dedicate more time to introducing ancient DNA and explaining the specific challenges involved in mapping this type of DNA.

Thank you for the suggestion, we added more information to this sentence: “The effect of mapping bias is exacerbated in ancient DNA studies due to post-mortem DNA damage such as fragmentation and cytosine deamination to uracil (which is sequenced as thymine) (Orlando et al 2021) which increases the chances of spurious mappings or rejected reads due to an excessive number of mismatches relative to the fragment length.”

Additionally, the detailed description of algorithms for estimating admixture proportions is more appropriate for the methodology section, specifically under "2.4 Estimating Admixture Proportions."

We prefer to have this description here as the comparison of the methods is an important point of our study.

Methodology

Regarding the methodology part, the four sections are relevant and well described. However, there are instances where the choice of certain parameters or values could benefit from more detailed justification or references such as bwa and ANGSD parameters. Additionally, I have significant concerns regarding the reproducibility of the simulations, as I encountered difficulties running your code for simulating genomic data, because some of the required packages and modules seem to be internal to the author’s system without detailed information about their contents. To improve reproducibility, it is crucial to make these packages and modules available to the community and provide clear instructions on the specific commands and procedures used for the simulations.

We apologize for the oversight. The missing code was added to the repository and a motivation on parameter choices was added to the manuscript text.

Furthermore, I am concerned that selecting only SNPs with matching alleles in both pseudohaploid and SNP chip data might introduce a selection bias. This filtering approach excludes SNPs whose genotype is different due to methods, which might overlook important differences caused by different genotyping methods. Comparing allele frequencies between pseudohaploid data with all SNPs and pseudohaploid data filtered to match the SNP chip could reveal if the filtering process introduces significant biases and demonstrate that the SNP filtering does not significantly alter the results.

We appreciate the reviewer’s concern about the scope of our results, but as we explain above, we believe that restricting to sites that match ascertained alleles both reflects a common use case with ancient DNA, particularly when SNPs are captured. Thus, we believe that focusing on this situation is the most relevant to provide an interpretable and practical assessment of mapping and method bias.

Results

The three sections are relevant, but the analyses seem shallow. For example, it would have been interesting to investigate whether certain genomic regions are more susceptible to mapping bias. Is mapping bias more frequent in GC-rich regions, repetitive sequences, or complex genomic areas? Visualizing the locations of these potential differences and correlating them with specific genomic features would provide deeper insights into the sources of mapping bias.

We appreciate the reviewer's suggestions and agree that they would be useful analyses. However, we opted to limit the scope of our manuscript to mapping bias caused by SNP and damage-associated mapping errors as a way to focus on the specific phenomenon of mapping bias. Including more complex types of genomic features, such as repeats or structural variants, may introduce additional effects, and we leave the study of those features to future work.

In general, the results lack proper biological interpretation and discussion, especially regarding the admixture simulations, on aspects such as LD pruning and the choice of reference genomes in function of each case. As it stands it is mainly descriptive.

We attribute these differences to fundamental differences in the models underlying the different methods. We have added a new paragraph to the discussion explaining this:

"While the ancestry estimates depended slightly on the reference genome the reads were mapped to, they seemed more influenced by the choice of method or software. Methods differed by more than 10% in their ancestry estimates from the same source data. This highlights that other factors and biases play major roles in the performance of these methods. Depending on the method, the type of input data, and the implementation, they showed different sensitivities to e.g. linkage or the amount of missing data (which was on average ~37% per SNP for the 0.5X and ~3% for the 2.0X simulations). For non-pruned data, qpAdm performed best across all scenarios and did not show any method-specific bias in certain ranges of simulated admixture proportions. Multiple differences between the PSD and qpAdm methods may have contributed to the relative biases we observed. PSD models may propagate allele-frequency misestimation more than qpAdm because of their assumptions of linkage equilibrium and Hardy-Weinberg equilibrium. Indeed, we observed that LD pruning improved the performance of PSD models, but they are known to be sensitive to sample size and drift (Lawson et al., 2018; Toyama et al., 2020). More generally, because it is based on Patterson's f statistics (Patterson et al., 2012), qpAdm estimates ancestry from relative differences. If mapping bias affects all populations similarly, then their relative relationships remain more stable. In contrast, PSD models reconstruct exact allele frequencies for the putative source populations therefore emphasizing the impact of mapping bias. Finally, the ancestry proportions of PSD models are constrained to [0, 1] which is not the case for qpAdm. Indeed, we see negative estimates in a small number of simulations (3 runs with 0.5X depth and 50,000 generations divergence). This (biologically unrealistic) flexibility of qpAdm compared to PSD models drives the mean estimated admixture admixture proportion down, which may account for some of the reduction in upward method bias compared to the other methods."

For example, it would be valuable to discuss and possibly investigate why allele frequencies from SNP arrays show lower correlation with those derived from pseudohaploid genotype calls, while admixture proportion estimation with qpAdm, which uses pseudohaploid genotype calls as input, appear to perform best.

During the re-analysis, we realized that the correlations with the pseudohaploid calls were calculated using the wrong files. Using the correct files shows the highest correlations between array-based allele frequencies and pseudohaploid allele frequencies as expected from the distributions. We have modified the text accordingly.

Additionally, the paper would benefit from clearer conclusions at the end of each result section to highlight the really important information to take home.

Since we have separate Results and Discussion sections, we opted to have the Results mainly describe the results, while going into more detail about the implications of the Results in the Discussion. Nonetheless, we have added some additional signposting of key take-aways throughout the Results.

Some figures are not readable, particularly those comparing simulated and estimated admixture proportions, as the use of white text on a gray background makes it difficult to read the details.

We have changed the ggplot theme used for these figures.

Adding tables with the actual values of estimated admixture proportions would be helpful, as the small differences are hard to discern from the graphs.

We have added such a table as a supplementary Excel Spreadsheet (Data S1).

Including a table with correlations between allele frequencies values of SNP array and the different methods could also be valuable.

We have added such a table as Table 2 in the revised manuscript.

Additionally, it would be useful to include a figure showing the distribution of read balance values (r) as supplementary material. This could help illustrate the types of mapping bias (reference or alternate) and the ratio between them.

We added average read balance values per individual and per population as supplementary tables.

Discussion

The limits of this study are discussed, but the authors should clarify some practical points such as whether it is better to use a reference genome that is closer or more distant genetically in order to compute allele frequencies or compute admixture proportions, since these results seem to be contradictory between computation of allele frequencies and admixture proportion inference.

We apologize for some of the confusion caused by the error we had in the original version; now the results are consistent that pseudo-haploid calls are the most highly correlated for both allele frequency and admixture proportion estimation. In addition, we emphasize that the human reference genome, being a mosaic, is somewhat difficult to compare to the unadmixed reference genomes from our simulations. To address the patterns seen in our simulations, we added the following sentences to the discussion:

“In particular, we found that mapping bias and method bias even counteract each other in certain cases, leading to better estimates of the admixture proportion when mapping to one of the sources.”

The authors should consider adding some perspectives to this work, particularly in relation to the limitations of the study. While the issue of mapping bias has been reduced, it has not been completely resolved, as seen in the simulated data.

We expanded our discussion of limitations and made more explicit the comparison of our approach with other approaches for dealing with mapping bias.

Moreover, the impact on admixture proportion inference with read (*real?*) data remains unknown. As mentioned by the authors, mapping bias might have a greater effect on real datasets due to higher genomic variability, so the genotype likelihood correction could potentially reduce this impact more significantly. It would be valuable to evaluate the effect of the corrected genotype likelihood on non-simulated data.

We agree that the simulations are a simplified case as we also state in our discussion section. However, simulations make sure that we can use a known truth as control which is difficult with empirical data without known evolutionary history. Simulation of full human genomes and sequencing reads would exceed our computational resources as the current simulation of a small genome followed by simulation of reads, mapping and downstream analysis already requires a six-digit number of CPU hours.

The following comments were exported from the PDF, we apologize for the formatting:

p.4

Quote paper: precise

Reviewer Comment: *I agree you obtained precise estimates but it's still biased. Indeed, although allele frequencies calculated from corrected genotype likelihoods are more closely correlated with the "true" likelihood compared to other methods, they still don't allow for an accurate estimation of the "true allele frequency." In fact, as the results indicate, allele frequency estimates based on genotype likelihoods often tend to be higher than the actual frequencies.*

This final sentence summarizes the overall results of our study and does not specifically refer to our new approach alone. We believe that the readers of our study do receive additional information on how to better perform their studies. We have added a “more” to the sentence to reflect that the estimates are still not perfect or unbiased. The sentence now reads: “Overall, our study provides valuable insights for obtaining more precise estimates of allele frequencies and ancestry proportions in empirical studies.”

p.5

Quote paper: The effect of mapping bias 10 is exacerbated in ancient DNA studies due to post-mortem DNA damage such as fragmentation and 11 cytosine deamination to uracil (which is sequenced as thymine) (Orlando et al., 2021).

Reviewer Comment: *Since the research focuses on ancient DNA, spend more time introducing ancient DNA and explaining the challenges of mapping this type of DNA. For instance, to clarify the difficulties in mapping ancient DNA, you could add that resulting short fragments lead to a decrease in the number of accepted mismatches and an increase in the likelihood of multiple matching sites in the genome. You could also talk more about pseuhaploid data.*

We have added “which increases the chances of spurious mappings or rejected reads due to an excessive number of mismatches relative to the fragment length.” to this sentence and add a justification for pseudohaploid data and ascertained variants later in the introduction.

Quote paper: However, some downstream results can 45 depend on the specific genotype likelihood model selected (Lou et al., 2021).

Reviewer Comment: *I'm not sure if this sentence is helpful since you doesn't specify any model of genotype likelihood before and you don't address how the choice of a particular genotype likelihood model can impact downstream results*

We agree and we have removed the sentence from the manuscript.

Quote paper: genotypes all sites

Reviewer Comment: *I think this term is unclear. Does it refer to all possible genotypes {AA, AC, AG, AT, CC, CG, CT, GG, GT, TT}?*

We have rephrased this statement to: "...with the contrast that we are using a set of pre-ascertained biallelic SNPs and our aim is not to call genotypes at all sites across the genome including potentially novel SNPs. This choice allows us to estimate mapping bias locus-specific rather than using one estimate across the full genome of the particular individual."

p.6

Quote paper: Next, we examine the assignment of ancestry proportions. Most currently used methods trace 61 their roots back to the software STRUCTURE (Pritchard et al., 2000; Falush et al., 2003, 2007; Hubisz 62 et al., 2009), a model-based clustering approach modeling each individual's ancestry from K source 63 populations (PSD model). These source populations can be inferred from multi-individual data (unsu64 pervised) or groups of individuals can be designated as sources (supervised). Popular implementations 65 of this model differ in terms of input data (e.g. genotype calls or genotype likelihoods), optimization 66 procedure and whether they implement a supervised and/or unsupervised approach (Table 1). In 67 the ancient DNA field, f statistics (Patterson et al., 2012) and their derivatives are fundamental to 68 many studies due to their versatility, efficiency and their ability to work with pseudohaploid data. 69 Consequently, methods based on f statistics are also often used for estimating ancestry proportions in 70 ancient DNA studies. One method that uses f statistics for supervised estimation of ancestry propor71 tions is qpAdm (Haak et al., 2015; Harney et al., 2021). In addition to the source populations ("left" 72 populations), a set of more distantly related "right" populations is needed for this approach. Ancestry 73 proportions are then estimated from a set of f4 statistics calculated between the target population 74 and the "left" and "right" populations.

Reviewer Comment: *It's not necessary to describe the methods at this point. Save some details for the Materials and Methods section.*

We opted to describe the different admixture proportion estimation methods in the introduction as we believe that a substantial part of the conclusions drawn by readers will be the differences between these tools. Therefore, this is an integral part of the study which we prefer to include this early on.

Quote paper: P(bi|A) = ei 3 1 − ei if b = A if b = A

Reviewer Comment: *You have to reverse the two conditions.*

Thank you for catching this! We corrected this typo, and verified that the code used in the manuscript had the correct form.

Quote paper: ascertained biallelic 92 SNPs

Reviewer Comment: *Add information about which SNPs they correspond to. If possible, provide the VCF files for these SNPs.*

This part describes the general setup of our proposed approach. In practice the list of SNPs is defined by the user. We modified the text to: "...we restrict the following analysis to a list of pre-defined ascertained biallelic SNPs (list provided by the user)..."

p.7

Quote paper: https://github.com/tgue/refbias_GL

Reviewer Comment: *I was unable to run our code to simulate the genomic data. It appears that some packages, like sim_pop_ngs, are internal to your system. Please make these packages available to the community and provide the specific commands you used to generate your simulation.*

We apologize for the omission of this file from the github repository. The code was written exclusively for this project but we forgot to add it to the public repository. We have now added this file as well as the version of gargammel used for the project to the repository.

Quote paper: ten 108 FIN individuals and 10 YRI individuals from the 1000 Genomes project (Table S1)

Reviewer Comment: *Why didn't you include individuals from other populations, such as the Asian population? Were you constrained by the availability of chip genotype data?*

Thank you for this suggestion! We have added a comparison to a Japanese population to the revised version of our manuscript.

Quote paper: allele frequency of at least 0.2

Reviewer Comment: *You filter SNPs with a minor allele frequency below 0.2 to avoid rare SNPs, which can introduce statistical noise. This also ensures that the SNPs included are more likely to be common in both the FIN and YRI populations, improving the comparability and robustness of the results. Is this correct? Please mention briefly the reason.*

Thank you for the suggestion, we added our motivation to the text. The sentence now reads: "The SNP data was filtered to restrict to sites without missing data in the 20 selected individuals, a minor allele frequency of at least 0.2 in the reduced dataset (considering individuals from all populations together), which makes it more likely that the SNPs are common in both populations and both over- and underestimation of allele frequencies could be observed."

Quote paper: To make the 114 sequence data more similar to fragmented ancient DNA, each read was split into two halves at its mid115 point and each sub-read was re-mapped separately.

Reviewer Comment: *Why didn't you use gargammel as shown below?*

We did not consider using gargammel here since (in its default use) gargammel is starting with a "true" sequence for the individual in FASTA format while the data for the 1000 genomes was in FASTQs output from a sequencer. Thus, we could not have simulated fragmentation using gargammel. The reviewer is correct that we could have skipped the fragSim module and directly used the FASTQ sequences as input for the deamSim sub-program. Overall, we expect that such a formal simulation of deamination damage would increase mapping bias as shown by Martiniano et al 2019.

Quote paper: samtools

Reviewer Comment: *It could be useful to include your Samtools and ANGSD commands in your GitHub repository or as supplementary data.*

We have added the scripts used for the empirical data analysis to our github repository.

Quote paper: -n 0.01

Reviewer Comment: *Was this value chosen to account for general sequencing error? If so,, mention that. It's always benefit to explain the rationale behind your choice of parameters.*

Yes, that is correct, we have added "to allow for more mismatches and gaps due to post-mortem damages and increased evolutionary distance to the reference" to the text. These parameter choices are widely used and have been extensively benchmarked (see the cited Schubert et al 2012 and Oliva et al 2021).

Quote paper: Pseudohaploid genotypes were called 118 with ANGSD v0.933 (Korneliussen et al., 2014) by randomly drawing one read per SNP as described for 119 the simulations below and only SNPs with the same two alleles in pseudohaploid and SNP chip data 120 were included in all comparisons.

Reviewer Comment: *The differences observed between figures 2A and 2B should therefore be attributed to the variation in input data used to calculate allele frequencies (chip SNP data versus pseudohaploid data) rather than to mapping bias.*

The reviewer is correct in stating that the differences could be due to a combination of technology, mapping bias and different ancestries in the populations shown in figure 2. Therefore and also in response to a comment below, we have reduced the discussion of this figure and moved the distributions to a supplementary figure focusing on the per site differences in the main figure instead.

Quote paper: Remapping of modified reads and genotype likelihood calculation 121 were performed as described above. Allele frequencies were calculated from genotype likelihoods with 122 ANGSD v0.933 (Korneliussen et al., 2014) using -doMaf 4 and the human reference as "ancestral" allele 123 in order to calculate the allele frequency of the reference alleles. SNP calls from the genotyping array 124 and pseudohaploid calls were converted to genotype likelihood files assuming no genotyping errors, so 125 the allele frequency estimation for this data could be based on ANGSD as well.

Quote paper: Population histories are simulated

Reviewer Comment: *It might be helpful to explain why you didn't use non-simulated data. I assume it's to have a true positive control, but providing a clear justification for this decision could be valuable.*

Correct, the use of simulations is very common for methods testing since it allows to control the true values against the inferred results. We added: "To test the methods while having control over the "true" admixture proportions, population histories are simulated using msprime…"

Reviewer Comment: *This part seems repetitive. I suggest to first mention genotype likelihood calculation and then mention allele frequencies calculation using ANGSD.*

We agree. This section was restructured in the revised manuscript.

Quote paper: qpAdm

Reviewer Comment: *What is the use of qpAdm?*

qpAdm is a method for estimating admixture proportions as introduced during the Introduction section.

Reviewer Comment: *You should include references to support your justification for using 20,000 and 50,000 generations for t123, explaining how these values correspond to divergence times within and between (sub-)species*

Quote paper: Different values of 20,000 and 50,000 generations are tested for t123 approximately corresponding to divergence times within and between (sub-)species

Quote paper: Mutation rate was set to 2.5 × 10−8 and recombination rate was set to 2×10−8.

Reviewer Comment: *Please support the values you use for the mutation rate and recombination rate with relevant sources.*

Quote paper: The effective population size along all branches is 10,000. 138 For each population, 21 diploid individuals (i.e. 42 haploid chromosomes) with 5 chromosome pairs 139 of 20,000,000 bp each were simulated.

Reviewer Comment: *Did you rely on any existing simulations, or did you create it yourself? If it's the former, include a reference. If it's the latter, talk about your motivation for choosing these values.*

All simulations were set up for this project. We added motivation for our parameter choices: "Different values of 20,000, approximately falling in the range of the split of all human populations (Schlebusch et al., 2017) or the Neanderthal-Denisovan split (Rogers et al., 2017) i.e. approximating the divergence between distant populations or sub-species, and 50,000 generations, corresponding to a comparison between closely related species, are tested for t123. Mutation rate was set to 2.5×10−8 and recombination rate was set to 2 × 10−8, which are both in the upper part of the ranges for mammals and vertebrates (Dumont and Payseur, 2008; Bergeron et al., 2023). The effective population size along all branches is 10,000, a value often considered for humans (Charlesworth, 2009). For each population, 21 diploid individuals (i.e. 42 haploid chromosomes) with 5 chromosome pairs of 20,000,000 bp (corresponding to a short mammalian chromosome arm) each were simulated."

p.8

Quote paper: from the simulated true genotypes

Reviewer Comment: *I do not understand what are the simulated true genotypes*

We rephrased the sentence to: "To avoid ascertainment bias and the effect of damages, sequencing errors and genotype callers, biallelic SNPs were ascertained directly from the simulated genotypes, prior to the gargammel simulation of reads and mapping, and restricted to SNPs with a minimum allele frequency of 10% in the outgroup population S1."

Quote paper: parameters -checkBamHeaders 0 -doHaploCall 1 -doCounts 1 -doGeno -4 -doPost 2 -doPlink 2 163 -minMapQ 30 -minQ 30 -doMajorMinor 1 -GL 1 -domaf 1.

Reviewer Comment: *Explain the choice of the parameters.*

We added explanations for the parameters.

p.9

Reviewer Comment: *It's not entirely clear what you mean by the underestimation of the reference allele frequencies. What should I observe on figure 2A and 2B?*

We tried to clarify this part by consistently talking about non-reference allele frequencies when referring to the figure as the x-axis displays the non-reference frequency. Furthermore, in response to a justified comment below, we have reduced the discussion of this figure and moved the distributions to a supplementary figure focusing on the per site differences in the main figure instead.

Reviewer Comment: *I noticed that the allele frequency distribution is slightly shifted towards higher estimates of non-reference allele frequency, rather than towards lower non-reference allele frequencies. It is also the case for corrected genotype likelihood, you should therefore mention it.*

We state this observation in the text: "Overall, genotype likelihood-based point estimates of the allele frequencies tend towards more intermediate allele frequencies while pseudohaploid genotypes and ``true'' genotypes result in more alleles estimated to have low and high alternative allele frequency"

Reviewer Comment: *I don't understand your reasoning. I expected to have a reference bias (overestimation of the reference allele frequencies) for the YRI population since some alternative allele could not map on the reference but I don't understand how it can lead to alternative bias (underestimation of the reference allele frequencies) for the FIN population. In this case we shouldn't observe any bias. I think you should add more explanation about your reasoning.*

We thank the reviewer for this comment. After careful consideration, we agree that the patterns seen in these plots cannot be directly attributed to mapping bias. Therefore, we decided to focus the main text on the per-site differences between estimates while moving the distribution bar charts to a supplementary figure.

Quote paper: Mapping bias in empirical data

Reviewer Comment: *It could worth to add a plot with just read balance value (r), you could add on supplementrary data. It could be useful to see type and ratio of mapping bias (reference or alternate bias)*

We added average read balance values per individual and per population as supplementary tables.

Quote paper: estimate allele frequencies

Reviewer Comment: *These allele frequencies are derived from the default genotype likelihood and the corrected genotype likelihood calculated from ancient DNA simulated using 1kGp data. If I'm right, it should be mentioned, as it could enhance the understanding of the types of data used.*

We have added this information, we now write: "We used ANGSD to estimate allele frequencies from genotype likelihoods based on short-read NGS data (read lengths reduced to 36-54bp to better resemble fragmented aDNA data) and compare them to allele frequencies estimated from the same individuals genotyped using a SNP array and pseudohaploid genotype data."

Reviewer Comment: *The x-axis in Figures 2A and 2B represents the frequency of the non-reference allele. To enhance clarity, you should refer to non-reference allele frequencies in the text rather than reference allele frequencies or vice-versa. Consistency between the plots and the text will improve understanding.*

Quote paper: reference allele frequencies

Quote paper: underestimation

Thank you for the suggestion. We agree and aimed to make the terminology in this part more consistent in our revised manuscript.

Quote paper: a pattern which could be related to the fact that 202 most of the human reference genome has European ancestry

Quote paper: In both tested populations, the default version of genotype likelihood 204 calculation produced an allele frequency distribution slightly shifted towards lower non-reference allele 205 frequency estimates

Quote paper: The allele frequen206 cies estimated from the corrected genotype likelihoods exhibit a slightly better correlation with the 207 “true” frequencies in both FIN (Pearson’s correlation coefficient 0.9297 [0.9294, 0.9301] vs. 0.9310 208 [0.9307, 0.9313] for uncorrected and corrected, respectively; p = 2.14×10−7) and YRI (Pearson’s cor209 relation coefficient 0.9444 [0.9442, 0.9447] vs. 0.9459 [0.9457, 0.9462] for uncorrected and corrected, 210 respectively; p = 1.8 × 10−14). Notably, allele frequency estimates from pseudohaploid data display 211 the lowest correlation with the “true” frequencies in both FIN (r = 0.8571) and YRI (r = 0.8344) 212 indicating that while the distribution of allele frequencies seems close to the true spectrum

Reviewer Comment: *It would be beneficial to include a figure or table with correlation data, or to add values to an existing figure, as this information is very informative. Moreover, you could remove values from the text (it's not easy to read) and refer to the corresponding table.*

We have added a table displaying these correlations as Table 2.

p.11

Reviewer Comment: *You mentioned that there is significantly higher variation between pseudohaploid and true frequencies at each specific hint. Did you perform any statistical tests to determine if this difference is statistically significant? On which figure(s)/table(s) can we observe that? table s2 ?*

Quote paper: significantly

This statement refers to the wider box and whiskers in the boxplots (now Figure 2A-C). We have added a reference to this figure but removed the term “significantly” as this is merely an observation from the figure without performing a statistical test.

Reviewer Comment: *Instinctively, one would expect a reduction in mapping bias when mapping to an ingroup population. Indeed, if I map reads from a European individual onto an African reference genome, I would expect that European SNPs wouldn't align well, thus increasing the mapping bias. In addition to the observed results showing that mapping bias is actually reduced when aligning to a reference genome from an outgroup, it would be interesting to explain biologically why this occurs.*

Quote paper: We find a similar result here: the parts of 223 the reference genome that can be attributed to African ancestry (Green et al., 2010) display a mean 224 and median difference of nearly 0 in FIN but allele frequencies remain higher than array estimates 225 in YRI (Figure S1). In contrast, the European and East Asian parts of the reference genome show a 226 distribution of differences around 0 in YRI but positive means and median in FIN (Figures S2 and 227 S3). This confirms the utility of reducing the effect of mapping bias by mapping against a reference 228 genome from an outgroup.

The reviewer is correct that the overall number of reads mapping successfully will be maximized if the reference genome comes from an ingroup individual. However, in practice individuals have a history of admixture and different parts of the genome and the paternal and maternal chromosome will have *differing distances to the reference individual. If the reference genome instead represents an individual equally distant to all admixing sources, the mapping would not be biased towards either source population. We decided to add an explanation in the beginning of our simulation section since the connection of S1, S2 and S3 to the model shown in Fig 1 might be easier to understand than the complex human history behind the human data.*

Quote paper: Estimation of admixture proportions based on genotype calls

Reviewer Comment: *You note a difference between ADMIXTURE and qpAdm, but it would be helpful to provide a detailed explanation of these differences. Specifically, consider discussing which parameters influence the discrepancies between the two methods. Additionally, exploring alternative techniques that account for these parameters could provide further insights and potentially improve the accuracy of your estimates.*

qpADm and ADMIXTURE (or PSD methods in general) have fundamental differences in the underlying model which likely form the basis for the observed difference. We have expanded our discussion section to highlight this:

“While the ancestry estimates depended slightly on the reference genome the reads were mapped to, they seemed more influenced by the choice of method or software. Methods differed by more than 10% in their ancestry estimates from the same source data. This highlights that other factors and biases play major roles in the performance of these methods. Depending on the method, the type of input data, and the implementation, they showed different sensitivities to e.g. linkage or the amount of missing data (which was on average ~37% per SNP for the 0.5X and ~3% for the 2.0X simulations). For non-pruned data, qpAdm performed best across all scenarios and did not show any method-specific bias in certain ranges of simulated admixture proportions. Multiple differences between the PSD and qpAdm methods may have contributed to the relative biases we observed. PSD models may propagate allele-frequency misestimation more than qpAdm because of their assumptions of linkage equilibrium and Hardy-Weinberg equilibrium. Indeed, we observed that LD pruning improved the performance of PSD models, but they are known to be sensitive to sample size and drift (Lawson et al., 2018; Toyama et al., 2020). More generally, because it is based on Patterson’s f statistics (Patterson et al., 2012), qpAdm estimates ancestry from relative differences. If mapping bias affects all populations similarly, then their relative relationships remain more stable. In contrast, PSD models reconstruct exact allele frequencies for the putative source populations therefore emphasizing the impact of mapping bias. Finally, the ancestry proportions of PSD models are constrained to [0, 1] which is not the case for qpAdm. Indeed, we see negative estimates in a small number of simulations (3 runs with 0.5X depth and 50,000 generations divergence). This (biologically unrealistic) flexibility of qpAdm compared to PSD models drives the mean estimated admixture admixture proportion down, which may account for some of the reduction in upward method bias compared to the other methods.”

Quote paper: On the full SNP panel, the median estimated admixture proportion differs 247 up to ~ 4% when mapping to reference genomes representing either of the two sources (S2 or S3) 248 while mapping to the outgroup reference genome (S1) results in estimates intermediate between the 249 two.

Reviewer Comment: *The choice of the optimal reference genome seems to vary with the admixture proportion. Specifically, S3 performs best for an admixture proportion of 0.1, S2 is most accurate for a proportion of 0.5, and either S1 or S2 is preferable for a proportion of 0.9. This variability suggests that the choice of the reference genome is highly dependent on the level of admixture. It would be helpful to provide some explanations or hypotheses regarding why this variability occurs based on the admixture proportion.*

We added an introductory paragraph explaining why we see the results for the outgroup reference genome S1 as results unaffected by mapping bias, which is usually intermediate between the two other reference genomes. This does not mean that the estimates are completely unbiased as we also observe issues due to the models underlying the different inference methods (see also response to the previous comment). We have also added a sentence to the Discussion explaining that method and mapping bias sometimes counteract each other, leading to more precise estimates of the admixture proportion when mapping to an ingroup reference genome.

Quote paper: reduces mapping bias

Reviewer Comment: *Please clarify how mapping bias is defined in your results. In this context, mapping bias is represented by the variation in admixture proportion estimates across the three reference genomes.*

We hope this is clarified with the new introductory paragraph to this sub-section.

Quote paper: qpAdm (Haak et al., 2015; Harney et al., 2021), on the other hand, estimated all 251 admixture proportions accurately when the outgroup (S1) was used for the reference genome sequence 252 and when the full SNP panel was used. The median estimates of admixture differed up to 3% between 253 mapping to reference genomes from one of the source populations (S2 or S3).

Reviewer Comment: *It would be useful to explain why there is a difference between ADMIXTURE and qpAdm. While ADMIXTURE appears to have similar variation to qpAdm (4% vs. 3%) , it tends to be less accurate, often overestimating or underestimating admixture proportions. Do you know which differences between the two techniques might influence their respective levels of precision?*

We have added some speculations about the reasons behind this pattern to the discussion (see also response to previous comment).

Quote paper: The extent of mapping bias 256 decreases with lower population divergence across all methods (Figure S4), as mapping bias should 257 correlate with distance to the reference genome sequence

Reviewer Comment: *This is contrary to the previous results, where mapping bias decreased when using an output population as the reference. Please explain this contradiction.*

We do not see this as a contradiction. Mapping bias here refers to differences between ancestry estimates when mapping to reference genomes originating from one of the sources. Consequently, less evolutionary distance between them results in weaker mapping bias. A true outgroup is equally distant from the sources resulting in a limited effect of mapping bias on the ancestry estimates. We added "lower population divergence between the sources" to this sentence.

p.12

Quote paper: Notably, when employing the 272 corrected genotype-likelihood the estimated admixture proportions when mapping to S2 or S3 are 273 slightly more similar than with the default formula without correction, showing that the correction 274 makes the genome-wide estimates less dependent on the reference sequence used for mapping while 275 not fully removing the effect.

Reviewer Comment: *It would be helpful to include the simulation result values in the supplementary data. Indeed, it seems that only the estimated admixture proportions when mapping to S3 change with the use of corrected genotype likelihoods, making the results between S2 and S3 more similar. It's difficult to fully understand how the values vary just by looking at the graph.*

We have added such a table as a supplementary Excel Spreadsheet (Data S1).

Quote paper: Furthermore, employing the mapping bias corrected genotype281 likelihoods made the estimated admixture proportions less dependent on the reference genome used 282 during mapping.

Reviewer Comment: *Please provide the differences in admixture estimates when mapping to the different reference genomes. You only mention percentage values for standard genotype likelihood, but not for corrected genotype likelihood.*

We have added such a table (Data S1), see above.

p.13

Reviewer Comment: *To better assess mapping bias differences in real data when using different reference genomes, it could be useful to examine variations in allele frequencies across parts of the reference genome attributed to different ancestries (European, East Asian, and African), as illustrated in Figures S1, S2, and S3. Another approach would be to create a pseudo-reference genome specific to each ethnicity to more accurately capture group-specific variations. By analyzing the differences in allele frequencies associated with various reference genomes, I think you could gain insights into how mapping bias varies with the source of the reference genome in real datasets.*

Quote paper: The differences seen 315 in our simulations are likely underestimates of what might occur in empirical studies as real genomes 316 are larger and more complex than what was used in the simulations.

We thank the reviewer for this suggestion but we believe that this would be beyond the scope of our study. As also outlined above, analyzing and simulating sequence data for full mammalian genomes would come with an excessive use of computational resources.

Quote paper: sampling artifacts.

Reviewer Comment: *Where do you show or explain that in your results?*

We have removed the mention of sampling artifacts from the manuscript.

p.14

Reviewer Comment: *Could you provide further insight into the types of biases that might influence the performance of different methods for ancestry estimation?*

Quote paper: other factors 330 and biases play major roles in the performance of these methods

We are discussing this in our added discussion paragraph:

“While the ancestry estimates depended slightly on the reference genome the reads were mapped to, they seemed more influenced by the choice of method or software. Methods differed by more than 10% in their ancestry estimates from the same source data. This highlights that other factors and biases play major roles in the performance of these methods. Depending on the method, the type of input data, and the implementation, they showed different sensitivities to e.g. linkage or the amount of missing data (which was on average ~37% per SNP for the 0.5X and ~3% for the 2.0X simulations). For non-pruned data, qpAdm performed best across all scenarios and did not show any method-specific bias in certain ranges of simulated admixture proportions. Multiple differences

between the PSD and qpAdm methods may have contributed to the relative biases we observed. PSD models may propagate allele-frequency misestimation more than qpAdm because of their assumptions of linkage equilibrium and Hardy-Weinberg equilibrium. Indeed, we observed that LD pruning improved the performance of PSD models, but they are known to be sensitive to sample size and drift (Lawson et al., 2018; Toyama et al., 2020). More generally, because it is based on Patterson's f statistics (Patterson et al., 2012), qpAdm estimates ancestry from relative differences. If mapping bias affects all populations similarly, then their relative relationships remain more stable. In contrast, PSD models reconstruct exact allele frequencies for the putative source populations therefore emphasizing the impact of mapping bias. Finally, the ancestry proportions of PSD models are constrained to [0, 1] which is not the case for qpAdm. Indeed, we see negative estimates in a small number of simulations (3 runs with 0.5X depth and 50,000 generations divergence). This (biologically unrealistic) flexibility of qpAdm compared to PSD models drives the mean estimated admixture admixture proportion down, which may account for some of the reduction in upward method bias compared to the other methods."

Quote paper: ADMIXTURE and NGSadmix benefit from LD pruning 344 while LD pruning increases the method bias for fastNGSadmix and introduces method bias for qpAdm.

Reviewer Comment: *you should add some explanation why some methods work better with LD-pruned data while others get more biased. I suppose that Methods that improve with LD pruning could be benefiting from having fewer redundant SNPs, which helps them focus on useful genetic information. However, methods that get worse with LD pruning might need more SNPs to work well and can struggle when there are fewer SNPs.*

We are discussing this in the added paragraph (see above).

Reviewer Comment: *While the study shows that unsupervised methods using genotype likelihoods, such as NGSadmix, can achieve accuracies comparable to those of methods like ADMIXTURE that use pseudo-haploid genotype calls, they still do not reach the level of accuracy demonstrated by qpAdm. qpAdm, which also requires pseudo-haploid genotype calls, outperforms the other methods in terms of accuracy. In this context, it is not clear why one should use methods like NGSadmix, which can reduce mapping bias with corrected genotype likelihoods but still do not reach the accuracy of qpAdm.*

Quote paper: Nonetheless, this study highlights that unsupervised methods employing genotype-likelihoods 355 (NGSadmix) can reach similar accuracies as methods such as ADMIXTURE that require (pseudo-haploid) 356 genotype calls. Moreover, methods that incorporate genotype likelihoods have the added benefit that 357 the modified genotype likelihood estimation approach can be used to reduce the effect of mapping bias. 358 Furthermore, if some samples in the dataset have >1X depth, genotype likelihood-based approaches 359 will benefit from the additional data and provide more precise estimates of ancestry proportions while 360 pseudo-haploid data will not gain any information from more than one read at a position. Finally, 361 genotype likelihoods are very flexible and can be adjusted for many other aspects of the data. For 362 example, variations of genotype likelihood estimators exist that incorporate the effect of post-mortem 363 damage (Hofmanov´a et al., 2016; Link et al., 2017; Kousathanas et al., 2017) allowing to use of all 364 sequence data without filtering for potentially damaged sites or enzymatic repair of the damages in 365 the wet lab.

We describe the limitations of qpAdm in the preceding paragraph. In short, qpAdm input data cannot be corrected for mapping bias the same way as we did for the GL-based methods. qpAdm requires additional populations which might not be available for many non-human species. Finally, our simulated demography represents an idealized scenario for qpAdm while other studies (cited in that paragraph) have shown substantial limitations of qpAdm in more realistic scenarios.

Reviewer 2: Review by Michael Westbury, 30 Aug 2024 12:17

The manuscript by Gunther and Schraiber present a welcome and interesting addition to our knowledge of mapping/reference biases and how that can impact downstream analyses, especially relevant for palaeogenomic studies.

We thank the reviewer for this assessment and the constructive feedback.

Overall I found it a sound study but lacking in some details in the methods that I think could help the reader understand what was done better and replicate it if needed.

Here are my specific comments that I hope will be helpful:

113-115: What are the original read lengths? 100bp or 150? Maybe even up to 300 if merged PE reads?

We added this information to Table S1.

119: Why put this before the simulations? It seems strange to me to not give the details of the method first and just say "as described for the simulations below" since the reader has not read it yet. Normally I would put the details first and then write "see above"

Thank you for this suggestion. We have restructured the methods accordingly.

123: What parameter does this in ANGSD? -anc ?

That is correct. We added this information to the manuscript and included the code into our github repository.

124-125: How was this done? and I don't quite understand what is meant by "the allele frequency estimation could be based on ANGSD"

We added more information, this part now reads: “SNP calls from the genotyping array and pseudohaploid calls were converted to genotype likelihood files assuming no genotyping errors (i.e. the genotype likelihood of the observed genotype was set to 1.0, others to 0.0 whereas all three likelihoods were set to 1/3 if data was missing for the site and individual). This allowed us to also estimate allele frequency estimates for this data with ANGSD.”

127: Above it was past tense and now it is present tense. It would be good to keep it consistent

Thank you for noticing! We have now changed the entire Methods section to past tense.

142: What does "according to the msprime simulations" mean?

We changed this to: “As msprime does not produce sequences but positions of derived alleles at each haploid chromosome, we had to convert this information into a sequence. For each chromosome, a random ancestral sequence was generated with a GC content of 41% corresponding to the GC content of the human genome. Transversion polymorphisms were then placed along the sequence at the positions produced by the msprime simulations.”

142: What do you mean by first sequences?

We changed this section to: “The resulting sequences for each haploid chromosome were then stored as FASTA files. One of the 42 simulated sequences from populations S1, S2 and S3 were used as reference genomes. Out of the remaining sequences, pairs of FASTA files were then considered as diploid individuals and used as input for gargammel (Renaud et al 2017) to serve as endogenous sequences for the simulation of next-generation sequencing data with ancient DNA damage.”

144: gargammel uses a haploid fasta file to create the simulated reads. When/how was this sequence made? I assume you ran it independently for 2 individuals and merged the reads? A little more detail here would be helpful. Or did you directly use the msprime output and put it into gargammel? Since you mention this in the acknowledgments

See above.

147: What coverage was simulated?

We added the sentence: “We simulated coverages of 0.5X and 2.0X.”

150: This does seem on the low end as most studies use 30bp due to potential for spurious mapping e.g. https://doi.org/10.1093/bioinformatics/btae436. Could this impact the levels of reference bias?

Thank you for noticing! We have changed this setting to 30bp and repeated all simulations. We do not see a large difference likely because the simulated genome is much shorter than a mammalian genome and the absence of any environmental (and other) contaminant sequences which could be involved in spurious mappings.

151: Only merged reads were mapped?

All reads were mapped, we added this information.

155: How/when was genotype calling done? From the bam files?

Following your comment below and a similar comment by reviewer #1, we have restructured this part. The calling of pseudohaploid genotypes is now described in the sentence following after this one.

156-158: I don't really understand what was done here

We have restructured this part and added more information. It now reads: “To avoid ascertainment bias and the effect of damages, sequencing errors and genotype callers, biallelic SNPs were ascertained directly from the simulated genotypes, prior to the gargammel simulation of reads and mapping, and restricted to SNPs with a minimum allele frequency of 10% in the outgroup population S1. 100,000 SNPs were selected at random using Plink v1.90 (Chang et al 2015) --thin-count.”

163: -GL 1 is using the SAMtools algorithm but above your said you tested the GATK one so I am a little confused. I know this is the pseudohaploid call but maybe it was also used for GL calling?

The reviewer is correct that this setting calculates a different type of genotype likelihoods. We were not using genotype likelihoods calculated by ANGSD (but the program crashes if -GL is not provided) and only used this call to create pseudohaploid calls. All genotype likelihoods were calculated in the same Python script. To avoid confusion, we have still changed this to -GL 2 which would calculate GATK genotype likelihoods.

170: There is no mention how the GL were calculated? What tool was used?

We used the same Python script as for the modified method. We have made this more explicit in the revised manuscript.

170: An additional method that is gaining in popularity for ancestry estimation is admixfrog https://github.com/BenjaminPeter/admixfrog while only a suggestion it would be interesting to see how this performs relative to the other methods

We agree with the reviewer that admixfrog is an exciting new method that we have also started using in some projects. However, in contrast to the purely genome-wide, allele frequency-based methods included in our comparison, it has the main goal of calling segments of ancestry. Consequently, we consider it a different category of methods that is not directly comparable to the other methods included. Therefore, we did not include it in our revised analysis.

171: But in the table there are only 4 shown

We apologize for this typo. Thank you for noticing!

195: Why is the array less influenced by reference bias?

We modified the sentence which now reads: "As the genotyping array does not involve a mapping step it should be less affected by mapping bias…"

202: For someone who doesn't work with human data, what do these abbreviations stand for? Based on this I would assume YRI is more European as it shows more reference alleles?

We added a decoding of those acronyms to the methods section where they are first mentioned.

204: The phrasing here is a little difficult to interpret. Based on the plot it looks like GL finds the reference allele more relative to the "True" when the freq non-ref allele is low or high.

We rephrased this sentence to make clear that we are comparing the two GL-based allele frequency distributions. It now reads: "In all tested populations, the default version of genotype likelihood calculation produced an allele frequency distribution slightly shifted towards lower non-reference allele frequency estimates compared to the corrected genotype likelihood"

206: I am confused here. Looking at Fig2A and B the GL look the most different to the True in terms of counts whereas pseudohaploid looks more similar

This has been corrected in the revised version of the manuscript. The correlations with pseudohaploid calls were based on a wrong input file.

213: I assume this means the correlations were done on a site by site basis? What could explain these differences between distribution and individual estimates?

This has been removed/corrected in the revised version due to the mistake mentioned above.

220: what is a particular hint? a site?

This should read SNP or site. Thank you for noticing!

300-302: Which specific fig/table does this refer to? I thought for GL there was a high correlation? I assume this refers to the outliers mentioned?

We added that this is referring to (new) Figure 2 showing the allele frequency differences at individual sites. To make the connection to empirical studies from the preceding sentence clearer, we added that they were trying to find loci under selection.

303-305: This is from Fig 2? if most non-reference alleles segregate at low frequency and you found when non-reference alleles are at a low frequency there is a lower count, doesn't this mean the opposite?

We agree that this sentence was a bit confusing. We have removed it from the manuscript and changed the next sentence to: "Without high coverage data, genotype likelihood approaches without an allele frequency prior will naturally put some weight on all three potential genotypes at a site, ultimately collectively driving allele frequency to more intermediate values."

331: Where was the amount of missing data tested? I assume this is referring to the coverage of 0.5x and 2x?

We have added a per-site average of missingness to this sentence (~3% for 2.0X versus ~37% for 0.5X).

In regards to the suggested list

Title and abstract

Does the title clearly reflect the content of the article? [x] Yes, [ ] No (please explain), [ ] I don't know
Does the abstract present the main findings of the study? [x] Yes, [ ] No (please explain), [ ] I don't know

Introduction

Are the research questions/hypotheses/predictions clearly presented? [x] Yes, [ ] No (please explain), [ ] I don't know
Does the introduction build on relevant research in the field? [x] Yes, [ ] No (please explain), [ ] I don't know

Materials and methods

Are the methods and analyses sufficiently detailed to allow replication by other researchers? [ ] Yes, [x] No (please explain), [ ] I don't know
Are the methods and statistical analyses appropriate and well described? [ ] Yes, [ ] No (please explain), [x] I don't know

Results

In the case of negative results, is there a statistical power analysis (or an adequate Bayesian analysis or equivalence testing)? [ ] Yes, [ ] No (please explain), [x] I don't know
Are the results described and interpreted correctly? [ ] Yes, [ ] No (please explain), [x] I don't know

Discussion

Have the authors appropriately emphasized the strengths and limitations of their study/theory/methods/argument? [x] Yes, [ ] No (please explain), [ ] I don’t know
Are the conclusions adequately supported by the results (without overstating the implications of the findings)? [x] Yes, [ ] No (please explain), [ ] I don’t know

Reviewer 3

Following guidelines review:
Does the title clearly reflect the content of the article? [X] Yes, [ ] No, [ ] I don't know
The title "Estimating allele frequencies, ancestry proportions and genotype likelihoods in the presence of mapping bias" accurately reflects the content of the article. It captures the main focus of the study, which is the impact of mapping bias on various population genomic analyses and the proposed methods to mitigate this bias.
Does the abstract present the main findings of the study? [X] Yes, [ ] No, [ ] I don't know
The abstract effectively summarises the key findings of the study, including:
1. The impact of mapping bias on allele frequency estimates and ancestry proportions.
2. The proposal of an empirical adjustment to genotype likelihoods to mitigate mapping bias.
3. Comparison of different methods for estimating ancestry proportions under various scenarios.
4. The effectiveness of the adjusted genotype likelihood approach in mitigating mapping bias.
Are the research questions/hypotheses/predictions clearly presented? [X] Yes, [ ] No, [ ] I don't know
The introduction clearly presents the research questions and objectives of the study. The authors effectively outline the problem of mapping bias in population genomic analyses (lines 1-6) and introduce their proposed solution of adjusting genotype likelihoods.
Does the introduction build on relevant research in the field? [X] Yes, [ ] No, [ ] I don't know
The introduction provides a comprehensive overview of the relevant research in the field, citing numerous studies that have addressed mapping bias in various contexts (lines 1-6, 19-23). The authors also discuss different strategies proposed to mitigate mapping bias (lines 23-30), effectively situating their work within the existing literature.
Are the methods and analyses sufficiently detailed to allow replication by other researchers? [ ] Yes, [X] No, [ ] I don't know
The methods section provides detailed descriptions of the simulation procedures, data analysis, and software used. The authors include specific parameters for their simulations, detail the process of generating sequencing data, and describe the methods used for estimating admixture proportions. However, full reproducibility hinges on access to the pseudo-haploid calls, which are inherently random. To ensure complete replicability of the results, it is essential that the authors either share the file containing all pseudo-haploid calls or provide the seeds used in their software (if applicable). This level of detail should allow for replication by other researchers.

Are the methods and statistical analyses appropriate and well described? [X] Yes, [ ] No, [ ] I don't know
The methods and statistical analyses appear appropriate for addressing the research questions. The authors use a combination of simulations and empirical data analysis, which is a robust approach for testing their hypotheses. They provide a clear rationale for their choice of methods and describe the statistical analyses in sufficient detail.
In the case of negative results, is there a statistical power analysis (or an adequate Bayesian analysis or equivalence testing)? [ ] Yes, [ ] No, [X] I don't know
The study does not present negative results that would necessitate a power analysis or equivalence testing. The results generally support the authors' hypotheses about the impact of

mapping bias and the effectiveness of their proposed correction method.
Are the results described and interpreted correctly? [X] Yes, [ ] No, [ ] I don't know
The results are described and interpreted with appropriate caution and consideration of limitations. The authors present their findings clearly, using figures and tables to support their interpretations. They discuss the impact of mapping bias on allele frequency estimates and ancestry proportion estimates in a balanced manner, acknowledging both the strengths and limitations of their approach.
Have the authors appropriately emphasized the strengths and limitations of their study/theory/methods/argument? [X] Yes, [ ] No, [ ] I don't know
The authors provide a balanced discussion of the strengths and limitations of their study. They acknowledge that their simulations may underestimate the effect of mapping bias in real-world scenarios (lines 315-324) and discuss the limitations of their approach in fully removing mapping bias effects (lines 373-376).
Are the conclusions adequately supported by the results (without overstating the implications of the findings)? [X] Yes, [ ] No, [ ] I don't know
The conclusions are well-supported by the results presented in the study. The authors do not overstate their findings and provide appropriate context for their conclusions. They emphasise the modest but significant effect of mapping bias on ancestry estimates and discuss the implications of their findings for future research in population genomics.
Addition comments:

MAJOR COMMENTS:
- The simulation of ancient DNA (aDNA) reads was conducted meticulously, adhering to established standards and protocols to generate data closely resembling authentic aDNA datasets. However, this approach does not fully capture the complexity of real-world scenarios, where aDNA samples typically undergo various laboratory treatments. Notably, treatments such as UDG (Uracil-DNA Glycosylase) or partial UDG are commonly applied but were not simulated in this study. Incorporating these treatments into the simulated dataset would enhance its fidelity to real-life aDNA data used in research.
Despite this limitation, the findings presented in the paper remain significant and worthy of publication. Nonetheless, simulating these laboratory treatments would further improve the dataset's quality and increase its relevance to real-world aDNA analysis.

We thank the reviewer for this suggestion but, respectfully, we believe that such a test would be outside of the scope of our study. UDG mostly reduces damages causing a decrease in the number of mismatches which reduces mapping bias (see e.g. Martiniano et al 2020).

- Regarding reproducibility, it is crucial that the authors share either the pseudo-haploid calls files or the software seeds (if applicable). This step is essential because these calls are inherently random, and without this information, it would be impossible for other researchers to replicate the study's results precisely.

We set the seeds to the iteration number (from 1 to 50). Sharing pseudohaploid calls would not be sufficient for reproducibility since it does not include the sequencing reads needed for the GL calculation. However, more than 150 TB of FASTQ files were simulated in total and sharing such a large amount of data seems impractical. Therefore, we believe that setting the seed to the iteration count should ensure sufficient reproducibility.

MINOR COMMENTS:
- The authors' selection of Finnish (FIN) and Yoruba (YRI) populations might need further explanation. While one can infer that this choice aims to contrast reference bias between

European and African genetic backgrounds, this rationale is not explicitly stated in the manuscript. Given that the human reference genome was primarily derived from European and African-American individuals (Green et al., 2010), it would be valuable to include a third population with a distinct genetic background, such as an East Asian population. This addition would provide a more comprehensive assessment of the new methods' performance across diverse ancestries. This suggestion is particularly relevant considering that Figure S3 examines variants attributed to East Asian ancestry.

Thank you for this suggestion. We have added JPT, an East Asian population to the comparison in our revised version.

- The manuscript would benefit from a brief discussion (or comparison with other methods , maybe put a figure in SI?) of the computational requirements for implementing the proposed genotype likelihood correction method. This information would be valuable for researchers considering adopting this approach in their own work (which for example is one of the main struggles with pangenome graphs).

Thank you for this valuable suggestion. We added the following text to the discussion:
“In contrast to other approaches to alleviate mapping bias, such as employing pangenome variation graphs (Martiniano et al 2020, Kopetkin et al 2023), it does not require establishing a separate pipeline. Instead, only reads mapping to a set of ascertained SNP positions need to be modified and remapped which only represents only a fraction of all reads and consequently will require a small proportion of the original mapping time. Our Python scripts used to calculate the genotype likelihoods could be optimized further, but this step is of minor computational costs compared to other parts of the general bioinformatic pipelines (~1 minute per individual in the empirical data) in ancient DNA research. The corrected genotype likelihoods can then be directly used in downstream analyses using the same file structures and formats as other genotype likelihood-based approaches.”

- Line 34: Typo - "apporach" should be "approach"’

Corrected.

- Line 52: Typo - "not to call genotypes all sites" should likely be "not to call genotypes at all sites"

Corrected.

- Line 74: Redundant wording - "We simulate data sequencing data with realistic ancient DNA damage.." Remove the first "data"

Corrected.

- Line 107: Inconsistent number formatting - "for ten FIN individuals and 10 YRI individuals" Should be either "ten" and "ten" or "10" and "10"

Corrected.

- Line 308: Incorrect idiom - "as face values" should be "at face value"

Corrected.

- Line 400: Formatting error in reference - "O?Sullivan" should be "O'Sullivan"

Corrected.

In conclusion, this manuscript presents a significant contribution to the field of population genomics, particularly in addressing mapping bias in short-read sequencing data, with a focus on ancient DNA (aDNA) research. The authors' novel approach to mitigating mapping bias through empirical adjustment of genotype likelihoods is well-executed and shows promise. While I cannot provide expert feedback on the population genetics analysis pipeline, the methods appear sound and well-documented. However, the study could benefit from additional analyses, such as including an East Asian population and simulating UDG-treated reads (e.g., using tools like https://github.com/sbg/Mitty). These additions, along with addressing the previously mentioned typos and ensuring reproducibility by providing necessary files or software seeds, would further strengthen the paper. Despite these suggested improvements, the current results are sufficient and scientifically sound. If the reproducibility issues are addressed and typos corrected, the paper is suitable for publication

We thank you for reading our manuscript and for the constructive feedback.