# Answers to reviewers

## Faster model-based estimation of ancestry proportions

Comments and questions from the reviewers are colored black, while our answers are colored green.

## Reviewer #1

The authors introduce a software tool, fastmixture, which infers ancestry proportions and allele frequencies within the same likelihood framework used by the frequently used ADMIXTURE software. They propose three novel computational enhancements to speed up the analysis of large datasets. These improvements include:

- An SqS3 acceleration scheme for the EM algorithm,
- A randomized singular value decomposition (SVD) for better initialization of allele frequencies and ancestry proportions,
- Mini-batch updates for the EM algorithm.

Although these improvements do not reduce the computational complexity, the authors state that together, they result in a 20-fold speedup. (Could not check this yet, see comment 5 below)

We thank the reviewer for their time and efforts in reviewing our manuscript, leaving us with the opportunity to provide clarifications that have significantly improved our work. We have addressed their comments and questions sequentially.

**Main Comments:**

1. It would be valuable to understand which of the three improvements contributes most to the performance gains. If the authors could provide details on the individual impact of each enhancement, this would add useful context.

Authors' reply: We thank the reviewer for the excellent comment. Since our initial submission, we have implemented multiple major computational updates to our *fastmixture* software. Specifically,  we now employ a quasi-Newton acceleration scheme, similar to *ADMIXTURE*, instead of the SqS3 scheme. Additionally, we have now expanded our analyses by incorporating a new, more complex demographic model (Scenario C) in our simulations. This model includes evaluations starting with random parameter settings in *Q* and *P,* allowing us to assess the effectiveness of the SVD initialization and mini-batch updates. The random parameter initialization will capture the effectiveness of the mini-batch updates in comparison to a standard accelerated approach like *ADMIXTURE*. Furthermore, the difference in runtimes between random initialization and SVD initialization runs illustrates the efficiency gains from our SVD initialization. We have included the new analysis and added the following text to the Results section:

*“We further evaluated the effectiveness of our SVD initialization by comparing it to random parameter initialization inside the fastmixture framework for Scenario C. We reported computational runtimes, log-likelihoods, RMSE, and JSD measures in Table S7, where the two initializations performed similarly but the SVD initialization approximately halves the runtime on average in comparison to having a random initialization. Therefore, our observed runtime gains relative to ADMIXTURE could largely be attributed to our proposed mini-batch optimization.”, page 9, paragraph 4.*

2. The performance improvement is substantial and could significantly enhance the workflow of many large-scale genomic studies, without requiring the adoption of an entirely new and potentially less comparable modeling framework. However, for fastmixture to become a viable replacement for ADMIXTURE in future studies, additional tests and direct comparisons with ADMIXTURE would be beneficial.

For instance, I am curious whether the SVD initialization step affects the number of modes (see https://doi.org/10.1093/bioinformatics/btw327) inferred by FastMixture in comparison to ADMIXTURE.

Authors’ reply: We thank the reviewer for the comment. In all analyses performed in our study, *fastmixture*, *ADMIXTURE,* and *SCOPE* consistently converge to the same solution (mode) across runs. There is only one exception, where *ADMIXTURE* found a single suboptimal solution in the new Scenario C for one of its K = 4 runs (model misspecification). However, *Neural ADMIXTURE* is repeatedly finding different suboptimal solutions (five different modes) in all scenarios, as the only software.
Our SVD initialization step enhances *fastmixture* robustness by yielding similar parameter initializations across different seeds, thereby significantly reducing the variance of the obtained solutions. This can also clearly be seen in the very low variance in log-likelihoods and other assessment measures across the runs reported by *fastmixture*.

Additionally, over- or under-specifying the number of populations, K, might affect fastmixture differently than ADMIXTURE.

Authors’ reply: We thank the reviewer for the great comment on model misspecification. Evaluating the impact of underspecifying the true ancestral courses is indeed insightful, as results should ideally reflect the demographic processes (population splits) in the simulated scenario. Consequently, we have run all software for the new Scenario C using K = 2, 3, 4, while the number of true ancestral sources was K = 5. However, we find that using an overspecified K is less meaningful for evaluation purposes, as it does not correspond to a single correct or optimal solution, making such results challenging to interpret. We have added a subsection to the Results section named “Robustness to model misspecification”, where we report the results for each software and include admixture plots for K = 2, 3, 4. We show that *fastmixture*, *ADMIXTURE,* and *SCOPE* behave as expected by detecting older population splits. However, *SCOPE* introduces more noise, likely due to the increased complexity of the simulation scenario.

*“For most scenarios in ancestry estimation, the true number of ancestral sources is rarely known. We therefore tested and compared all software and their capabilities to deal with model misspecifications related to the number of ancestral sources used for Scenario C, which had a ground truth of K = 5. Here we would expect the ancestry estimations to capture older events in the demographic model for K < 5. The results comparing the software for K = {2,3,4} are displayed in Figures S6, S7 and S8, respectively, and their corresponding log-likelihoods are reported in Table S8. We note that ADMIXTURE only found the optimal solution in four out of five runs, thus showcasing its vulnerabilities due to random parameter initialization and a standard optimization approach. Due to the increased complexity of the simulation scenario, SCOPE exhibited an even further increased noise level in its ancestry estimates across all three values of K.”, page 9, paragraph 5.*

3. The simulated scenarios appear to be fairly narrow, and expanding the range of population structures could provide more insights. For example, all the scenarios presented (Figures S1 and S2) involve just one admixed population, with variations only in the number of non-admixed populations. There seems to be no ongoing migration between simulated populations. It would be interesting to see whether FastMixture performs similarly to ADMIXTURE in more complex population histories, such as those with multiple admixed populations or constant migration.

Authors’ reply: We thank the reviewer for the great suggestion and the opportunity to expand our manuscript. We have constructed a new simulation scenario with multiple admixture events, including continuous migration between the non-admixed populations. We simulate five non-admixed populations, a population with four-way admixture, a population with three-way admixture, and a population with two-way admixture. As anticipated by the reviewer, this scenario provided further insights into the benefits of the likelihood-based approaches in comparison to the likelihood-free approach. *fastmixture* and *ADMIXTURE* once again perform comparably. We have added text regarding the simulation in the Methods, Results and Discussion sections:

*“In Scenario A, B and D, we sample 1000 individuals, while in the more complex Scenario C, we sample 1,600 individuals. We perform standard filtering on minor allele frequencies at a threshold of 0.05, resulting in datasets consisting of 689,563 SNPs, 687,107 SNPs, 685,592 SNPs and 500,114 SNPs for Scenario A, B, C and D, respectively.”, page 6, paragraph 2.*

*“We further evaluated the different software in a more complex simulation scenario, Scenario C, which includes five ancestral sources (K = 5) with symmetric migration patterns and three admixed populations(Figure 1). Consistent with results from the simple scenarios, fastmixture*

*and ADMIXTURE outperformed the two other approaches, with fastmixture being ~28x times faster than ADMIXTURE. Due to the increased complexity of the simulation scenario, SCOPE exhibited an even greater increase of noise in its ancestry estimates, while Neural ADMIXTURE again failed to detect one of the unadmixed population sources and modeled two admixed populations as ancestral sources. Examining the accuracy of the ancestry estimates in each of the eight populations, we observed that fastmixture and ADMIXTURE performed similarly across the unadmixed and admixed populations, whereas SCOPE inferred more accurate ancestry estimates in the admixed populations in comparison to the unadmixed populations (Table S4 and S5).", page 8, paragraph 4.*

*"Our findings suggest that the added noise in the ancestry proportions estimated in SCOPE are likely to increase further in scenarios with a larger K or more complex demographic models, as demonstrated in scenario C. This limits the utility of SCOPE in association studies and precision medicine.", page 12, paragraph 3.*

4. I am also curious why NeuralAdmixture performs poorest among the evaluated methods. In the NeuralAdmixture paper, performance did not seem to drop significantly for admixed populations. Perhaps the authors could provide more insights here.

Authors' reply: We thank the reviewer for raising this concern, which we share. The initial results from *Neural ADMIXTURE were unexpected, prompting us to* reach out to the authors regarding its poor performance on the 1000 Genomes Project. Unfortunately, after over a year, our concerns remain unresolved (https://github.com/AI-sandbox/neural-admixture/issues/20). The poor performance appears to stem from their defined convergence criterion, in which log-likelihoods are averaged across batches, samples and variants, leading to suboptimal optimization.
In the GitHub issue, the first author of *Neural ADMIXTURE* tested the software on the 1KGP data, yielding 10 different results (modes: based on the *pong* tool), whereas *ADMIXTURE reliably* finds the same solution 10 out of 10 times with a much better log-likelihood on the same data. We have found it hard to validate their results from the original paper as all methods appear to have been run once, thus using a single seed. In most of their analyses, they use a combined dataset of the Human Genome Diversity Project, Simons Genome Diversity Project and the 1000 Genomes Project, where they defined ground truth labels by super population definition, thus not a real ground truth. They hereby completely disregard the accuracies of estimated ancestry proportions in admixed individuals and punish accurate methods in their assessments (these results can be found in their Supplementary Material).

5. The tool on GitHub was easy to install. However, the script produced an error when we ran it on our example files. This could be due to issues with the local package versions, but currently, there is no way to verify this, as I couldn't find clear version requirements in the github repository. It would be helpful if the GitHub repository included explicit requirements and a minimal working example to make installation verification easier.

Authors' reply: We thank the reviewer for the comment and bringing this to our attention. We have now provided detailed guidelines on how to install from GitHub directly, using PyPI. Additionally, we have provided a Conda environment (including version requirements of packages) and a Docker container to streamline installation and ensure compatibility.

The Manuscript:
The manuscript is well-organized, with a clear explanation of the motivation behind the research and the methods employed. I found it easy to place the manuscript within the broader context of related work. Other than the lack of detail on the impact of individual improvements, the manuscript does a good job explaining the concepts. A short paragraph on the principles behind SVD, similar to the description of the SqS3 algorithm, could make the ideas more accessible to readers.

Authors' reply: We thank the reviewer for their suggestion. We have now provided more information on the principles behind SVD and why it is useful in our case.

*"We initialize Q and P using individual allele frequencies estimated from randomized singular value decomposition (SVD) performed on the genotype matrix, combined with an alternating least squares (ALS) approach. SVD is a widely used dimensionality reduction approach in population genetics, which infers continuous structure by extracting axes of genetic variation.", page 4, paragraph 2.*

Apart from this, I found the manuscript easy to follow and enjoyed reading it.
Further Comments Regarding the Figures:
Figure 1: Sorting individuals within each subpopulation by the ancestry proportions inferred by fastmixture could help make the distribution of ancestry proportions in the admixed population more visually clear. This suggestion applies to the other figures as well, especially Figure 2.

Authors' reply: We thank the reviewer for this input. We have now sorted the ancestry estimates of individuals in the same superpopulation based on the *fastmixture* results for the two 1000 Genomes Project datasets (Figure 2 and Figure S9).

Figure 3: This would be more effective as a table.

Authors' reply: We thank the reviewer for the comment and apologize for any confusion, as the table (Table S1) already contains the relevant information. To clarify, we have now referenced this table in the figure caption. Since the first submission, we have made major computational updates to our algorithm and all the runtimes of *fastmixture* are updated. Figure 3 now includes an additional zoomed-in runtime plot, which makes it easier to compare the runtimes of *fastmixture*, *Neural ADMIXTURE* and *SCOPE*.

## Reviewer #2

1. Does the abstract present the main findings of the study? Not completely, expansion of the abstract is needed

Authors' reply: We thank the reviewer for the comment. The abstract now provides a more accurate summary of our findings by addressing the issues of noise in the likelihood-free approaches and updated runtime numbers.

2. A more detailed introduction

Authors' reply: We thank the reviewer for the suggestion. In response, we have now expanded on previous literature in the introduction:

*"Due to scalability issues, the Bayesian approach was later replaced by maximum likelihood models, which were optimized using expectation-maximization (EM) and block relaxation algorithms, this includes the widely used software ADMIXTURE [4,5].", page 2, paragraph 1.*

*"The SCOPE software [10] has gained increased popularity due to its efficient implementation of an ALS approach, which is well-suited for biobank-scale datasets.", page 2, paragraph 2.*

3. Perhaps build in a step to convert large-scale whole genome sequencing VCFs to binary plink files.

Authors' reply: We thank the reviewer for the comment. We expect researchers to have already performed preprocessing on their data prior to running *fastmixture*. The binary PLINK file format is one of the most used data-formats in population and statistical genetics, and since the *PLINK* software itself has very easy and efficient options to perform the conversion from VCF or BCF to binary PLINK files ("plink2 –bcf <file.bcf> –make-bed –out <file>"), we therefore advise researchers to use *PLINK* for this conversion during the data preprocessing.

4. A more detailed explanation on what preprocessing steps are assumed to be completed is necessary. This will affect the algorithms accuracy.

Authors' reply: We thank the reviewer for their suggestion. We point researchers to follow standard preprocessing steps for population genetic analyses. This includes standard quality control steps for sequencing or SNP chip data, variant and sample filtering to obtain a final curated genotype dataset in binary PLINK format of common variants in unrelated individuals. In our study, we only focus on the highly curated 1000 Genomes Project and simulated datasets. We have expanded on the expected preprocessing steps in our manuscript:

*“We assume that the user has performed standard quality control and preprocessing, (e.g., variant filtering based on a minor allele frequency threshold and to only include unrelated samples).”, page 5, paragraph 2.*

5. I would recommend increasing the sample number for each demographic scenario used for the simulations.

Authors’ reply: We thank the reviewer for the comment. We have now included a more complex simulation scenario that includes 1,600 individuals distributed across 5 ancestral populations and 3 admixed populations. We have kept the rest of the sample sizes at 1,000 individuals for the other simulation scenarios due to simple feasibility and complete comprehensiveness for the *ADMIXTURE* software.

*“In Scenario A, B and D, we sample 1000 individuals, while in the more complex Scenario C, we sample 1,600 individuals. We perform standard filtering on minor allele frequencies at a threshold of 0.05, resulting in datasets consisting of 689,563 SNPs, 687,107 SNPs, 685,592 SNPs and 500,114 SNPs for Scenario A, B, C and D, respectively.”, page 6, paragraph 2.*

*“We further evaluated the different software in a more complex simulation scenario, Scenario C, which includes five ancestral sources (K = 5) with symmetric migration patterns and three admixed populations(Figure 1). Consistent with results from the simple scenarios, fastmixture and ADMIXTURE outperformed the two other approaches, with fastmixture being ~28x times faster than ADMIXTURE. Due to the increased complexity of the simulation scenario, SCOPE exhibited an even greater increase of noise in its ancestry estimates, while Neural ADMIXTURE again failed to detect one of the unadmixed population sources and modeled two admixed populations as ancestral sources. Examining the accuracy of the ancestry estimates in each of the eight populations, we observed that fastmixture and ADMIXTURE performed similarly across the unadmixed and admixed populations, whereas SCOPE inferred more accurate ancestry estimates in the admixed populations in comparison to the unadmixed populations (Table S4 and S5).”, page 8, paragraph 4.*

*“Our findings suggest that the added noise in the ancestry proportions estimated in SCOPE are likely to increase further in scenarios with a larger K or more complex demographic models, as demonstrated in scenario C. This limits the utility of SCOPE in association studies and precision medicine.”, page 12, paragraph 3.*

6. Whilst I understand the rationale behind not including the results for the ADMIXTURE run for full 1000 Genomes dataset, I still think including the results is needed for transparency. How long did ADMIXTURE take to run the full 1000 Genomes dataset? Please include the log-likelihood in Table S3.

We thank the reviewer for the comment and great suggestion. We have now provided the results of a single *ADMIXTURE* run for the full 1000 Genomes Project, which took > 40 hours to complete further showcasing the scalability issues of *ADMIXTURE*. We have added and edited the text in the Results and Discussion sections of our manuscript to accommodate the changes:

*"For the full 1KGP dataset, we only performed a single run for ADMIXTURE due to its excessive computational runtime of > 40 hours, in comparison to the other software with runtimes of < 2 hours (Figure 3 and Table S1). Here fastmixture was ~30x faster on average than the ADMIXTURE run.", page 11, paragraph 1.*

*"The results of ADMIXTURE are only based on a single seed in the full 1KGP SNP set due to exhibiting a prohibitively excessive computational runtime of more than 40 hours, in contrast to ~74 minutes using fastmixture. We have therefore evaluated all software on a downsampled version as well, with 10x less SNPs, which is a common procedure in population genetic studies.", page 12, paragraph 1.*

2. An explanation of the admixture plot populations abbreviations is needed.

Authors' reply: We thank the reviewer for the comment. We have now included a description of the population abbreviations in the caption of the 1000 Genomes Project ancestry plots to enhance clarity for readers.

*"AFR: African, EUR: European, EAS: East Asian, AMR: American and SAS: South Asian ancestry.", captions of Figure 3 and S9.*

## Reviewer #3 (Oscar Lao Grueso)

The method proposed in this study combines various machine learning and optimization algorithms to significantly reduce the time required for estimating ancestry proportions while producing cutting edge results. The article presents an innovative approach to minimizing computation time by integrating different techniques from the field of machine learning, which I found very insightful and enjoyable to read. I agree with the authors that this methodology has the potential to "be the preferred alternative to ADMIXTURE in future population genetic studies".

Authors' reply: We thank the reviewer for their kind words, time, and efforts in reviewing our manuscript. We have addressed their comments and questions sequentially.

However, my main concern lies in how the proposed methodology compares with other existing algorithms. Many established methods assume marker independence and unrelated individuals (for instance, see Methods Mol Biol. 2020; 2090: 67–86. doi:10.1007/978-1-0716-0199-0_4). In contrast, based on the simulation study, it appears that only markers with a minor allele frequency (MAF) below 0.05 are excluded from the analysis. Even when the authors use a subset of markers from the 1000 Genomes Project, this subset is selected randomly. I believe that some of the discrepancies observed between methods could stem from this bias.

Authors' reply: We thank the reviewer for the excellent comment. We agree that LD pruning has been a common approach for reducing the SNP set, even when working with structured or admixed populations. However, it has recently been shown that standard LD pruning biases downstream measures for population differentiation as it removes variants that contribute to large allele frequency differences between the ancestral sources (https://doi.org/10.1101/2024.05.02.592187). The study shows that this also affects the $F_{ST}$-measures of *ADMIXTURE*. The variants removed are therefore not in LD in the ancestral populations. Standard LD pruning will artificially make unadmixed populations more genetically similar, and should therefore be ill-advised.

Additionally, the simulated models considered in the study raise some concerns. While these models are relevant for studying human demography, they seem rather specific. It would be beneficial to include other, more complex models, especially since the authors assert that 'Our findings suggest that the added noise in the ancestry proportions estimated in SCOPE will only increase for scenarios with a larger K and for more complex demographic models.

Authors' reply: We thank the reviewer for the great suggestion and opportunity to improve our manuscript. We have constructed a new simulation scenario with multiple admixture events including having continuous migration between the non-admixed populations. We simulate five non-admixed populations, a population with four-way admixture, a population with three-way admixture and a population with two-way admixture. As the reviewer also expected, this scenario provided more insights into the benefits of the likelihood-based approaches in comparison to the likelihood-free approach. *fastmixture* and *ADMIXTURE* perform comparably once again. We have added text regarding the simulation in the Methods, Results and Discussion sections:

*"In Scenario A, B and D, we sample 1000 individuals, while in the more complex Scenario C, we sample 1,600 individuals. We perform standard filtering on minor allele frequencies at a threshold of 0.05, resulting in datasets consisting of 689,563 SNPs, 687,107 SNPs, 685,592 SNPs and 500,114 SNPs for Scenario A, B, C and D, respectively.", page 6, paragraph 2.*

*"We further evaluated the different software in a more complex simulation scenario, Scenario C, which includes five ancestral sources (K = 5) with symmetric migration patterns and three admixed populations(Figure 1). Consistent with results from the simple scenarios, fastmixture and ADMIXTURE outperformed the two other approaches, with fastmixture being ~28x times faster than ADMIXTURE. Due to the increased complexity of the simulation scenario, SCOPE exhibited an even greater increase of noise in its ancestry estimates, while Neural ADMIXTURE again failed to detect one of the unadmixed population sources and modeled two admixed populations as ancestral sources. Examining the accuracy of the ancestry estimates in each of*

*the eight populations, we observed that fastmixture and ADMIXTURE performed similarly across the unadmixed and admixed populations, whereas SCOPE inferred more accurate ancestry estimates in the admixed populations in comparison to the unadmixed populations (Table S4 and S5).", page 8, paragraph 4.*

*"Our findings suggest that the added noise in the ancestry proportions estimated in SCOPE are likely to increase further in scenarios with a larger K or more complex demographic models, as demonstrated in scenario C. This limits the utility of SCOPE in association studies and precision medicine.", page 12, paragraph 3.*

I have some additional comments and questions regarding the tests conducted:
Please provide references for the values mentioned, 'constant recombination rate of 1.28 × 10−8 and a mutation rate of 2.36 × 10−8.' Additionally, it would be helpful to describe (where do they come from, for example) the demographic parameters in the Materials and Methods section and include the msprime code.

Authors' reply: We thank the reviewer for the comment and for bringing it to our attention. We have now added citations, and the code for replicating the entire simulation study (including the *msprime* code) is available in the data repository on Zenodo as stated in the "Code availability" statement.

In Scenario C (Figure 1), using the American-Admixture demographic model, the study states that 'we consistently observed that ADMIXTURE and fastmixture perform similarly in accuracy, with results closest to the ground truth (Table 1 and Table S2).' However, I believe the standard error should also be taken into account. ADMIXTURE shows a standard error ten times smaller than fastmixture.

Authors' reply: We thank the reviewer for the comment. We note that we are reporting the standard deviations for all assessment measures and not standard errors. We agree that examining the standard deviations is useful for assessing the variation across different runs, however, in our results for Scenario D (formerly C), the standard deviations for both *fastmixture* and *ADMIXTURE* are <1e-5, which amount to differences on the fifth and sixth decimal places between runs, thus next to none.

To present the results from Table 1 more visually, consider calculating the KL divergence between the predicted admixture values and the ground truth for each individual in each population and study for each evaluated method. This could help identify if any particular method exhibits more bias toward certain genetic backgrounds. For instance, the Neural ADMIXTURE paper observed that their method produces harder cluster predictions compared to ADMIXTURE, potentially impacting admixture proportions in mixed populations, as suggested by the authors of this study.

Authors' reply: We thank the reviewer for their input, which provides a great approach to further explore the differences between the software. We have now added the population-specific performance metrics in the new simulated Scenario C with 5 unadmixed and 3 admixed populations. The results are reported in Table S4 and S5 for RMSE and JSD, respectively.

*"When looking closer at the accuracy of the ancestry estimates in each of the eight populations, we could observe that fastmixture and ADMIXTURE performed similarly across the unadmixed and admixed populations, whereas SCOPE inferred more accurate ancestry estimates in the admixed populations in comparison to the unadmixed populations (Table S4 and S5).", page 8, paragraph 4.*

In the legend of Figure 2, please specify that it uses the downsampled version (I understand this applies to all methods, not just ADMIXTURE).

Authors' reply: We thank the reviewer for the comment. With the addition of an ADMIXTURE run on the full 1000 Genomes Project, the ancestry plot on the full dataset is now included as a main figure in our manuscript, while the ancestry plot on the downsampled dataset is in the supplementary material. The distinction is described in the captions of the ancestry plots.

It would also be very interesting to evaluate the performance of the proposed algorithm concerning the hyperparameters it requires.

Authors' reply: We thank the reviewer for the comment. We have now evaluated a broad range of initial mini-batches used in *fastmixture* for the complex simulation scenario, Scenario C. We have tested B = 8, 16, 32, 64, 128, where B = 32 is the default choice in *fastmixture*. We show that *fastmixture* is very robust to changes in its hyperparameter, where all the different choices of initial mini-batches converge to the same optimal solution across all seeds. We have plotted the admixture plots and reported assessment measures in Figure S4 and Table S6, respectively. We have also added the following text to the Results section under a new subsection "Testing hyperparameters":

*"The number of initial batches in fastmixture, used for its mini-batch optimization, is a hyperparameter. We tested the effect of changing the number of mini-batches in the more complex simulation scenario, Scenario C, having multiple admixture events and five source populations. We utilized B = {8,16,32,64,128}, including the default choice of B=32, and reported the computational runtimes, log-likelihoods, RMSE and JSD measures. Our results showed that fastmixture was robust to changes in B, as all evaluated choices consistently captured the same solutions with highly comparable assessment measures (Figure S4 and Table S6). Based on these findings, we conclude that B=32 was an optimal choice, balancing both fast runtimes and highly accurate ancestry estimations.", page 9, paragraph 3.*