nature portfolio

Peer Review File

Tracing Human Genetic Histories and Natural Selection with Precise Local Ancestry Inference

Corresponding Author: Dr Puya Yazdi

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The article “Tracing Human Genetic Histories and Natural Selection with Precise Local Ancestry Inference” presents a novel method for Local ancestry inference and shows that it outperforms existing methods by a large margin. Compared to some existing state of the art methods, their new method, “Orchestra” is shown to be able to distinguish between ancestries with more subtle levels of genetic differentiation. Indeed, for a panel of 16 ancestries, Orchestra was shown to have an accuracy above 75% for each ancestry, far surpassing the competing methods (and overall accuracy ~90% vs ~75% the rest). This is quite dramatic as the overall accuracy of LAI has not _drastically_ improved in the last decade.
Local ancestry inference (LAI) is an important tool to understand the demographic past of populations, including admixture, migration and selection. LAI is also proving to be a useful in the study of the genetic basis of disease and complex phenotypes (e.g. Genome wide association studies GWAS), so an improvement in LAI methods of this magnitude would be quite remarkable.
So normally, this would be an exciting situation, as taken at face value these results suggest a new generation of LAI inference methods is possible. However, after reading the article I remain quite skeptical of the results as presented. This feeling of skepticism is the result of multiple factors, so I will try to outline the causes here, and get into more detail below.
This is a method paper, as the main purpose is to introduce and demonstrate the effectiveness of a novel method. And yet, there is very little focus on making this method available as a resource / available to other scientists. Indeed it seems the opposite is more the case, this method is not really meant to be available to other scientists. What is the purpose of publishing a method that is not available to science?
AI / machine learning/ neural network methods have shown to be amazingly powerful at classifying and other tasks. But as we move away from biologically-inspired models and towards more black-box models there is increased potential for overfitting these complex statistical models, and this can happen in many different and sometimes subtle ways. In order to rule out the possibility of overfitting in Orchestra, there needs to be more clearly defined steps in generating distinct train / test / validation datasets, including the synthetic admixed individuals.
All of the authors are employed by the same for-profit company. This doesn’t necessarily signal a problem, but when combined with not following open-science principles - this paper can veer a bit too close to reading more like a press release than a peer-reviewed scientific article.

Taken together, these issues leave me with deep reservations about the current version of the manuscript, but these are issues that can be addressed.

Despite these issues there is quite a bit to like about this article. In general the writing is clear, the authors are familiar with the relevant literature and have gathered an impressive array of genetic data relevant to the questions they address. The ability to resolve sub-continental ancestry is really very appealing. Existing methods can do this to some extent, but the results shown here represent what could be a substantial improvement. Following on this, this manuscript shows some interesting applications of this with LA enrichment analyses. Figure 2 shows LA inference in Latino Americans, these are interesting result, but it not clear what if any of the country averages could have achieved with non-local ancestry methods e.g. PCA or ADMIXTURE type analyses. It would be useful to quantify how much is gain via LA-based analyses. The article also uses the orchestra results to also presents some interesting ways

I will now go into a bit more detail about the three main issues I listed above.

As noted, this is a methods paper, but it does not provide any way for others to apply or evaluate the described methods. There is no source code provided, no executable, and the statistical model is not even fully described. In this situation there is no ability for anyone outside the authors to evaluate the method - we must take them on their word. There is no code availability statement in the article, and the data availability statement reads ‘Orchestra is available for non-commercial use via the corresponding author’ this is not a sufficient way to make science accessible in 2024. It is simply too easy for these requests to get ignored or delayed. The reporting summary document offers some hope and says “our method will be published as a CodeOcean capsule” this could be a good solution, but it is impossible for me to evaluate, this is not currently available. It may end up containing the full source code and method, but I don’t know.

The potential for overfitting is present in many types of statistical models, and the risk increases as the models get more complex. The neural network-based methods used in Orchestra are complex statistical models with many parameters and so are especially at risk of overfitting. The effect of overfitting would be that evaluation on the ‘test’ data set would show improved perforation relative to truly novel input data. The actual model behind orchestra is not fully described, so it is hard to fully evaluate, but I see multiple possible sources of overfitting.
The first and most obvious possible source is that the same individuals are used as source haplotypes to generate the synthetic target data and then used as a set of reference haplotypes for inferring LAI. This leads to increased similarity between the target and reference haplotypes compared to a novel target individual/haplotype. This has long been recognized as an issue when evaluating LAI methods (see e.g. the original RFMix paper or the MOSAIC paper) as there is a sincere desire to generate realistic input data and so real haplotypes are used, but a set of reference haplotypes are also required, so there is a temptation to use one set of haplotypes for both purposes due to limited sample size. The author’s even recognized this “To decrease bias resulting from the inclusion of the initial samples in the reference sequences, Orchestra's base layer algorithm was modified to remove the best matching haplotypes from the entire training set first, using the same greedy matching algorithm”, but it is not clear that this modification removes this issue. There should be a complete separation between the haplotypes used as a reference and those used to generate synthetic target data, and close relatives should also be excluded if they are present at appreciable levels. Another way to deal with this issue is to work with simulated data (e.g. the FLARE paper) where sample size can be as large as needed, but this involves complex simulations and cannot fully match some aspects of real data sets.
Again the model / method is not fully described, so it is not clear if a trained model is specific to a particular set of reference haplotypes, or if it is portable and a different reference set can be used. The article presents a case where the ‘test’ data is perfectly matched to the perfectly ancestry-annotated training data. So the admixture proportions / generation of admixture / source populations / allele freq-distribution / SNP data QC / SNP array are all perfectly matched between the test and training data sets. This will not generally be the case for novel data - there will not be a perfectly matching set of local ancestry-annotated training data available. A general approach to a situation like this is to include a validation data set that is distinct from the train/test, so that the model is evaluated on how well it generalizes beyond the very similar test/train datasets. This might involve leaving out some of the source populations from the test/train and then using them for the evaluation. Along these same lines the QC steps taken to select variant sites are another possible source of overfitting, as batch effects may have been retained when retaining ancestry-associated sites.

Finally, all of the authors are from the same company. It’s great that the authors want to share their science and to get the stamp of approval that publishing in a peer-reviewed journal provides, but the embedded profit motive (including a relevant patent application) raises extra potential for conflicts of interest and so sets a higher bar for open science and reproducibility (e.g. the FAIR research software principles published by Nature https://www.nature.com/articles/s41597-022-01710-x, which this article does not try to meet. As noted above, there is no ability of other scientists to evaluate the claims in this manuscript, so it is too close to a press release to be comfortable.

Small comments
Line 21(Abstract): “We then deployed it to answer long-standing debates about 22 the ancestral origins of Ashkenazi Jews”. This language in the abstract is too strong. It is fine that this article sheds some light by applying the method to Ashkenazi jewish data, and shows broad consistency with other methods, but that is a long way away from “answering long-standing debates”. The main text does a better job here, highlighting that this result is broadly consistent with other articles.
Line 29: “The remainder, of which only 0.1% is due to SNPs” not fully clear what is meant by this.
Line 33: “allowing us to reference any single individual to well differentiated reference populations” I think it would be better to present LAI as a particularly useful tool - sometime it is useful to conceive of people’s ancestry as mixture of well-differentiated source populations, but this is not the only correct way to think about ancestry, nor are the source populations always clear.
Line 49: “Many human populations, despite being geographically close, are genetically heterogeneous.” I don’t quite get how this sentence works - what is the connection between being close and being heterogeneous?
Line 55: the author’s cite a paper for the polygenetic adaptation on height, but this finding has been called into question by a pair of papers in 2019 https://elifesciences.org/articles/39725 and https://elifesciences.org/articles/39702 I suggest the authors change this example. This is an important change in the manuscript, but does not really substantially change the case for the usefulness of LAI.
Figure 1 - there are no confidence intervals or sample sizes included in this figure, but these are important for interpreting the results.
Figure 1 - it is not clear the admixture history simulated for the evaluations shown here. What populations are mixing and at what rates?
Figure 2 - the sublabels a/b/c do not match the figure caption

(Remarks on code availability)

No code was provided or available.

Reviewer #2

(Remarks to the Author)
In this manuscript by Lerga-Jaso et al., the authors propose a new method for local ancestry inference (LAI) and put forth an aggregated reference panel they have curated from prior studies. They benchmark their method against other commonly used LAI tools using simulated data, and then apply it to empirical data to characterize the demographic history of multiple populations and investigate potential indications of past natural selection as indexed by local ancestry enrichment. Their new method indeed seems to outperform the LAI tools against which they benchmark, though I think there are several areas that could use expansion.

Major Comments:

1. A more complete description of the existing tools and how they compare to the proposed model is needed. The introduction does a good job at spelling out the utility of LAI, but currently there is no detail about how existing LAI methods work, where they fail, and what makes Orchestra different/better.
2. Does sample size imbalance in the reference bias prediction? Some populations contain less than 100 individuals, while the biggest are capped at 500 (presumably in an effort to reduce massive sample size imbalance). Did you quantify how big of a concern sample size imbalance is, and/or do you have solutions in place for addressing it?
3. In the benchmarking against other LAI methods, were the optimal settings for these methods used? For example, in RFmix, there is a flag to account for unbalanced reference sample sizes to improve accuracy. Were these used or were all tools simply run with the default parameters? Testing the competing LAI tools' performance when run in the best way for the use case would be the most fair comparison to running your new method with its optimal parameters.
4. What happens when a user paints a cohort with an ancestry that is not present in the reference populations? Does Orchestra output any metric of confidence for its assignment such that poorly painted regions could be filtered out?
5. The authors construct their reference panel with both sequence and array data, which is unusual; prior work has limited LAI reference panels to sequence data to capture haplotypes defined by less common variants. Given they later restrict to MAF>5% this may be a moot point, but I wonder at the choice for restricting to such common variants in the reference. This would limit to the use of only older variants that are more likely to be shared across populations; would you not get improved performance by including variants at a slightly lower MAF to capture more private alleles?
6. Also, imputation is known to perform less well on diverse populations compared to ones better represented in the reference panel. The authors do several QC procedures to try and account for batch effects, but I do not see INFO score directly used to restrict to high-confidence imputed sites – why is that? Also the authors note "To minimize the impact of potential biases due to imputation, we maximized the presence of directly genotyped SNPs in our final panel by filtering by the union of all genotyped SNP sets contained in the array-based studies employed". Would this not create more risk of batch effects by platform since some variants are directly genotyped and some are imputed depending on the study, and studies contain different and mostly non-overlapping populations? Typical procedure for mega-analyses in GWAS consortia is to impute across arrays separately and then intersect after INFO score filtering.
a. Related, data were imputed with 1kG, but there are larger and better performing reference panels out now; e.g. TOPmed, the joint 1kG/HGDP callset from Koenig et al. If this reference is to serve as the training data on which all LAI calls are built having the best-in-class imputation done would be vital.
7. Some discussion of grouping for ancestry labels is warranted. For example, why are multiple diverse African countries given a single ancestry label (e.g. GLS = Ghanaian, Ivorian, Liberian, & Sierra Leonan)?
8. The authors reference historical events explaining the presence of some cryptic ancestries in modern populations. Do the local ancestry tract lengths match the expectations for the timing based on these historical records? This would be most convincing; as it is one could conceivably explain the presence of most ancestries based on some past event, so demonstrating the timing of relevant genetic admixture events is concordant with records would be more direct.
9. The authors note a somewhat surprising lack of concordance in detection of selection signals between their work and a prior publication. They explain their failure to replicate some peaks as due to different reference panels, but is that really enough to completely change signals? Also they don't just fail to detect some peaks that were previously found, they also detected new peaks as well. Are these real or false positives? The presence of multiple new signals isn't discussed in the text at all.
10. There is currently no benchmarking of computational efficiency across this and competing methods. This is an additional important consideration for users given how computationally intensive LAI is, and some discussion of the runtime/computational load required for various methods would be valuable.

Other comments:
a. How did you decide on these dimensionality reduction techniques? As has recently come to the fore, UMAP is typically not the best choice for representing ancestry, as it will discretize populations more than they are in actuality (ancestry is continuous). UMAP is not featured here for anything other than reference panel curation, but the authors may be cautioned about pushback for its use in an ancestry context.
b. Nice strategy to leverage global ancestry in the smoothing to refine LAI predictions; this is similar to what is done in the EM iteration in RFmix.
c. Figure 1 Panel b has a ton of shapes going on - some other way to show this might be clearer…
d. Panels are blocking points in Extended Data Fig. 3.
e. Some abbreviations for geography are non intuitive and/or conflict with field standard ones. For example CSA is typically

Central and South Asian (per HGDP) but here is used to indicate Central, South, and Southeast Africa.
f. Methods of the algorithm are written very clearly.
g. Please add page and line numbers.

(Remarks on code availability)
I believe a more open data sharing strategy is warranted. For the empirical data, the aggregated reference dataset should be somehow made available rather than simply directing the user to the component studies. Aggregating this resource was a lot of work, as the authors indeed highlight in their main text, so is an unreasonable ask to task each subsequent user to replicate the procedure. I appreciate that some of the component studies have access restrictions/applications, but perhaps a streamlined data application for reuse could be implemented such that the aggregated data file could be directly shared.

On the software front, Orchestra itself should be distributed in a more scalable way than by emailing the corresponding author.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
This paper presents Orchestra, a machine learning-based method for Local ancestry inference (LAI). The results shown here represent a substantial improvement in LAI accuracy, which not really been seen in the last decade. The authors have comprehensively replied to the reviewers comments. They have made their code and method available, edited the text for clarity, and reworked some analyses. They show that their method is a clear improvement in local ancestry inference, improving accuracy and demonstrate that fine-scale analyses are possible

The authors have posted code on GitHub and Code Ocean, posted pre-trained models on AWS, and included two ready-made example analyses. This represents a substantial improvement in the availability and accessibility to the science and code presented in the manuscript.

I appreciate the author's clarification of their approach to train-test data splitting. It is my understanding that they have segmented the training and test data in a reasonable way.
I remain slightly skeptical that some of the headline results will fully generalize because the increased flexibility of the model in Orchestra (vs other methods) may be better able to capture (nearly unavoidable) data artifacts such as batch effects, compared to previous methods. However, it seems the authors have taken reasonable steps to try to address this.

I appreciate the added details on tract length distributions in the Alternative figure 2, and suggest that this is the better figure. But I find that the figures could be more clearly labeled and laid out.

Line 19 - "SNPs, which account for only 0.1% of our DNA" The authors have revised this sentence in this version, but I still find it unclear. Do the authors mean that approx. 1 in 1000 basepairs of DNA contains a SNP? This very much depends on the set of samples you look at and the frequency of SNP sites that are considered. There are very many SNPs with rare alleles. I still think the authors should consider what idea they want to convey here. Perhaps they are making a point about pairwise differences between individuals at SNP variant sites, but I'm not sure.

(Remarks on code availability)
The authors have made the analysis code publicly available, and have provided a small example analysis .

There are details installation instructions and code to facilitate installation and setting up the software environment and data.

Reviewer #2

(Remarks to the Author)
I am broadly satisfied with the response from the authors. They have provided a thorough and detailed comparison of tract lengths with historical records, which significantly enhances the credibility of their findings. Additionally, the revised selection scanning they employed offers a more robust validation of their results, further reinforcing the reliability of their outcomes. I also appreciate the improvements made to the language in the introduction and discussion sections. The authors have addressed my critiques to that effect effectively, particularly in their discussion of existing competing LAI tools and with respect to the computational requirements for Orchestra. These sections are now clearer and more informative, providing a better context for understanding the advantages and limitations of their method. Moreover, the authors have taken significant steps to enhance the accessibility of their work by openly distributing their code and pre-trained models based on the curated reference panels. This is vital, facilitates reproducibility and will also encourage broader use and validation of their method within the scientific community.

My only remaining gripe is their inclination to defer the consideration of accuracy estimates to a future publication. Accuracy

is highly important for LAI, so it would be fantastic to incorporate those analyses into this work if that effort is ongoing. At the very least, some discussion of the considerations for use with ancestry components not in the reference should be added to the text, since misclassification will occur in that instance and apparently cannot be easily filtered out.

(Remarks on code availability)
I have perused the github page for the release of this code and am satisfied with the availability of code and the usability of the documentation.

Version 2:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
I am satisfied with the revisions made by the authors in this most recent round of review. I appreciate the detailed response to my lingering prior question regarding accuracy estimates for ancestries, with a focus on the impact for ancestries that are not well represented in the reference panel. The added paragraph to the discussion which describes these considerations is an important facet for future users of the program.

(Remarks on code availability)

RESPONSE TO REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

The article "Tracing Human Genetic Histories and Natural Selection with Precise Local Ancestry Inference" presents a novel method for Local ancestry inference and shows that it outperforms existing methods by a large margin. Compared to some existing state of the art methods, their new method, "Orchestra" is shown to be able to distinguish between ancestries with more subtle levels of genetic differentiation. Indeed, for a panel of 16 ancestries, Orchestra was shown to have an accuracy above 75% for each ancestry, far surpassing the competing methods (and overall accuracy ~90% vs ~75% the rest). This is quite dramatic as the overall accuracy of LAI has not _drastically_ improved in the last decade. Local ancestry inference (LAI) is an important tool to understand the demographic past of populations, including admixture, migration and selection. LAI is also proving to be a useful in the study of the genetic basis of disease and complex phenotypes (e.g. Genome wide association studies GWAS), so an improvement in LAI methods of this magnitude would be quite remarkable.
So normally, this would be an exciting situation, as taken at face value these results suggest a new generation of LAI inference methods is possible. However, after reading the article I remain quite skeptical of the results as presented. This feeling of skepticism is the result of multiple factors, so I will try to outline the causes here, and get into more detail below.

1.

This is a method paper, as the main purpose is to introduce and demonstrate the effectiveness of a novel method. And yet, there is very little focus on making this method available as a resource / available to other scientists. Indeed it seems the opposite is more the case, this method is not really meant to be available to other scientists. **What is the purpose of publishing a method that is not available to science?**

Thank you for your valuable feedback on the accessibility of our method. We fully agree that ensuring scientific methods are openly available is essential for advancing research and promoting collaboration within the scientific community. At the time of our submission, we faced unexpected complexities in transferring our code from the computational platform where we conducted our analyses to GitHub. This process was more challenging and time-consuming than anticipated, which regrettably delayed the public release of our code.

We have since resolved these issues and are pleased to inform you that the full codebase for our method, Orchestra, is now publicly available on our GitHub repository: https://github.com/omicsedge/orchestra-paper. This repository includes:

- **Training Code**: The complete method used to train our models (framework for simulating admixed genomes for training, alongside Orchestra's base and smoothing layers).
- **Inference Code**: Code for applying the trained models to perform inference on new datasets.

We also provide pre-trained Models: Models trained on both the 16-population (1KGP-16pops) and the 35-population (custom-35pop) panels, as well as smaller models used for natural selection analyses in the manuscript through AWS:

- https://orchestra-paper-public.s3.us-east-1.amazonaws.com/1kg-16pops.tar.gz
- https://orchestra-paper-public.s3.us-east-1.amazonaws.com/custom-35pops.tar.gz

- https://orchestra-paper-public.s3.us-east-1.amazonaws.com/custom-35pops-latino.tar.gz
- https://orchestra-paper-public.s3.us-east-1.amazonaws.com/cosmopolitan-mexicans.tar.gz
- https://orchestra-paper-public.s3.us-east-1.amazonaws.com/fulani.tar.gz
- https://orchestra-paper-public.s3.us-east-1.amazonaws.com/makrani.tar.gz
- https://orchestra-paper-public.s3.us-east-1.amazonaws.com/peruvians.tar.gz
- https://orchestra-paper-public.s3.us-east-1.amazonaws.com/sahelian-arabs.tar.gz

Pre-trained models can be directly applicable to new datasets, facilitating immediate use without the need for extensive training times or significant computational capacity. We acknowledge that the training process for our method demands high-performance computing resources. By providing pre-trained models, we aim to make Orchestra more accessible to researchers with varying levels of computational capabilities.

Additionally, we have created a small toy example on CodeOcean that can be run interactively. This example allows users to explore and experiment with our method in a user-friendly environment without the need for specialized hardware. We have also uploaded scripts and data for a more complex toy example to the AWS link, where you can explore selection signals for adaptive admixture in Admixed Mexicans by estimating LAD and Fadm scores:

- https://orchestra-paper-public.s3.us-east-1.amazonaws.com/canonical_results.tar.gz
- https://orchestra-paper-public.s3.us-east-1.amazonaws.com/prepared_ref_and_target_panels.tar.gz
- https://orchestra-paper-public.s3.us-east-1.amazonaws.com/toy_example_scripts.tar.gz
- https://orchestra-paper-public.s3.us-east-1.amazonaws.com/observed_and_expected_frequencies_for_Fadm.tar.gz

We believe these steps significantly enhance the accessibility and usability of Orchestra, aligning with the open-science principles that we deeply value. We are committed to supporting other scientists in applying our method and welcome any further suggestions to improve its availability.

2.

AI / machine learning/ neural network methods have shown to be amazingly powerful at classifying and other tasks. But as we move away from biologically-inspired models and towards more black-box models there is increased potential for overfitting these complex statistical models, and this can happen in many different and sometimes subtle ways. In order to rule out the possibility of overfitting in Orchestra, **there needs to be more clearly defined steps in generating distinct train / test / validation datasets, including the synthetic admixed individuals.**

We have now included a new section in the supplementary methods titled "Analytical Framework Details and Orchestra Implementation Guide". This section provides a more detailed explanation, addressing how we defined the different datasets and generated simulated individuals using SLiM. Due to space limitations set by the journal, this additional information is included in the supplementary material, which we refer to for those seeking more details. We hope this makes everything clearer:

- Analytical Framework Details and Orchestra Implementation Guide
  To train Orchestra, we require a reference panel consisting of genotype data paired with a sample mapping file that links each sample ID to its population membership. This panel should consist exclusively of non-admixed samples. For benchmarking in the present study, we utilized two reference panels: (i) 1KGP-16pops, a WGS dataset from the 1000 Genomes Project; and (ii) custom-35pop, a broader and more diverse panel containing 35 populations (see ‘Reference Panel’ in the main Materials and Methods). Orchestra’s model, however, is designed to be

compatible with any reference panel, with no restrictions on granularity—allowing flexibility to accommodate continental, sub-continental, population, or sub-population level distinctions—and with any number of populations. Additionally, Orchestra can be trained with adjustable window sizes and a wide range of SNP inputs, including our ancestry-specific SNP set (as used with custom-35pop), array-based genotyping data or WGS, providing adaptability for a range of research applications.

Given the challenge of obtaining datasets with perfectly annotated ground truth, we split the reference panel into two independent subsets for benchmarking: 80% for training (serving as the reference panel for model development) and 20% for testing (serving as the target for accuracy benchmarking). To create synthetic admixed individuals for training, we simulated six generations of random admixture within the training subset using SLiM (see 'Simulated Admixed Individuals' in the main Materials and Methods). During Orchestra's base layer phase, initial non-admixed samples from the training set were used as reference to reconstruct the synthetic admixed haplotypes generated with SLiM, generating recombination distance vectors (see Figure 1a). To prevent overfitting, given that reference and target haplotypes overlapped within the training set, we excluded the best-matching haplotypes, assuming these represented the original reference genomes from which synthetic individuals were generated. These vectors served in turn to train the parameters of Orchestra's smoothing layer (weights and bias terms), with the simulated admixed genomes as target haplotypes with known local ancestry information. Once trained, the model could then be applied to any target cohort.

For accuracy benchmarking, the resulting trained model was applied to the defined testing cohort, where the original non-admixed samples were considered as generation 0. Additionally, synthetic admixed samples were generated across six generations using SLiM to test varying levels of admixture. This setup maintained a clear separation of haplotypes by using independent training and testing cohorts, thus avoiding any potential inflation in accuracy metrics.

Following benchmarking, Orchestra was trained on the full reference panel (without splitting) to maximize accuracy in downstream analyses. The model leverages both non-admixed individuals from the reference panel and simulated admixed individuals, which are necessary as input to train the smoothing layer parameters; thus, generating synthetic admixed samples is a required preliminary step before model training. Once trained, these models are fully portable and can be applied to any independent target cohort using Orchestra's inference module. Detailed usage instructions are available in our GitHub repository at https://github.com/omicsedge/orchestra-paper.

3.

All of the authors are employed by the same for-profit company. This doesn't necessarily signal a problem, but when combined with not following open-science principles - this paper can **veer a bit too close to reading more like a press release than a peer-reviewed scientific article**.

We understand your concern regarding potential conflicts of interest due to all authors being affiliated with the same for-profit company. To demonstrate our commitment to open science, we have made our full codebase, including training and inference tools and pre-trained models, publicly available on GitHub and AWS. These steps reinforce that our work is intended as a valuable scientific contribution, not promotional material. We hope this demonstrates our commitment to transparency, reproducibility and collaboration within the scientific community.

4.

Taken together, these issues leave me with deep reservations about the current version of the manuscript, but these are issues that can be addressed.

Despite these issues there is quite a bit to like about this article. In general the writing is clear, the authors are familiar with the relevant literature and have gathered an impressive array of genetic data relevant to the questions they address. The ability to resolve sub-continental ancestry is really very appealing. Existing methods can do this to some extent, but the results shown here represent what could be a substantial improvement. Following on this, this manuscript shows some interesting applications of this with LA enrichment analyses. Figure 2 shows LA inference in Latino Americans, these are interesting result, but it **not clear what if any of the country averages could have achieved with non-local ancestry methods** e.g. PCA or ADMIXTURE type analyses. It would be useful to quantify how much is gain via LA-based analyses. The article also uses the orchestra results to also presents some interesting ways

It is true that in our analysis of Latin American populations, we initially employed Orchestra primarily as a global ancestry inference method by averaging local ancestry results, without fully exploiting the software's local ancestry features until later sections that focused on selection signals. Reviewer #2 also noted this and suggested that incorporating local ancestry results at this step could significantly strengthen our findings, specifically by comparing local ancestry tract lengths with historical expectations. Indeed, detailing the local structure of admixture events between major and trace ancestries in modern Latin American populations provides a clearer demonstration of the benefits of local ancestry inference, thereby highlighting the effectiveness of the Orchestra tool in mapping the genetic makeup at a local level—something global methods may not distinctly resolve.

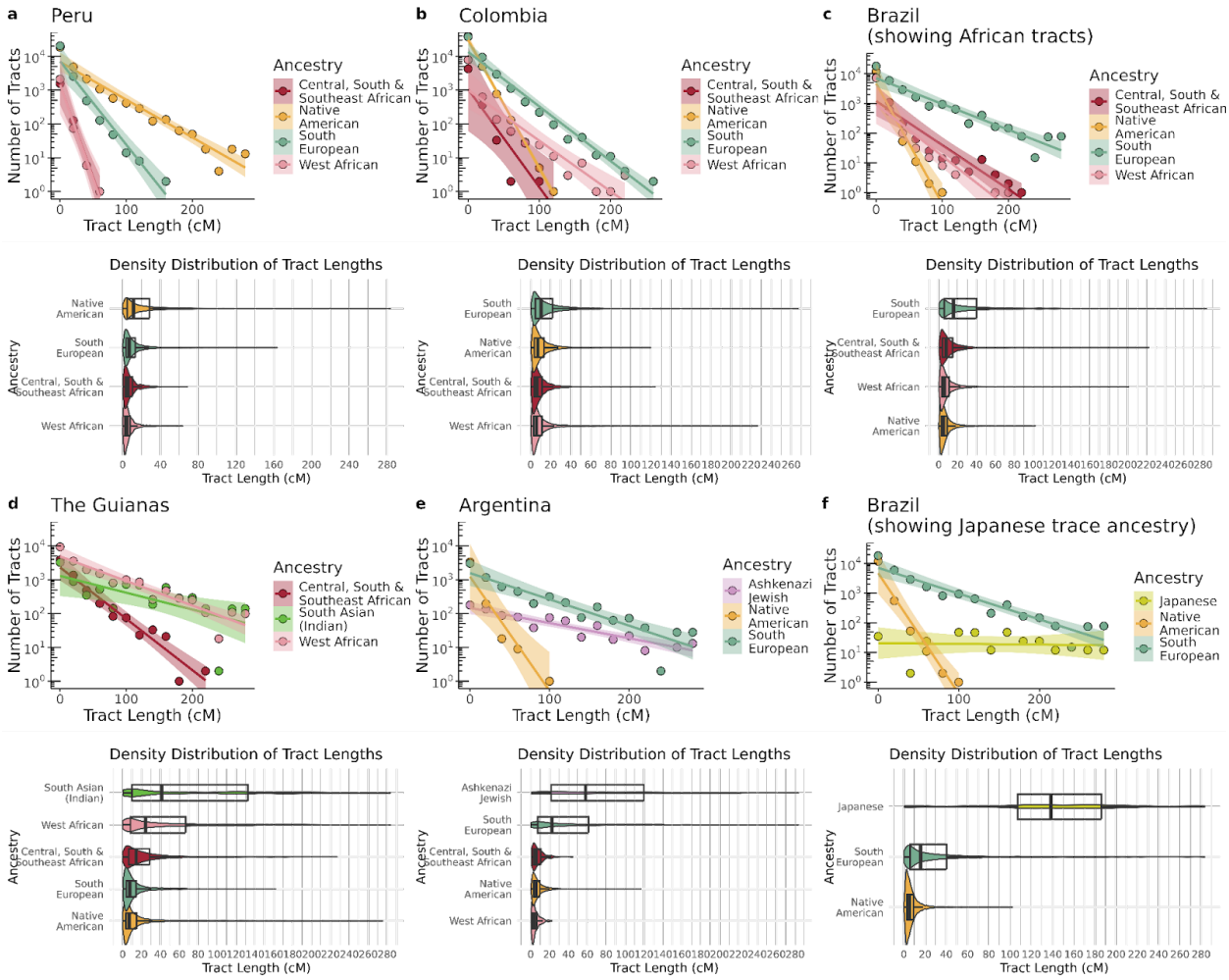
In response, we have added a new supplementary figure (see below) that illustrates the distribution of local ancestry tract lengths across various South American populations. These panels highlight the admixture contributions from Native American, South European and African ancestries (major sources), and South Asian (Indian), Ashkenazi Jewish and Japanese ancestries (trace ancestries). This analysis provides a more nuanced view of the historical admixture events that are specific to each country, including:

- Panels a, b and c (Peru, Colombia and Brazil): These show the distributions of Native American, South European and African admixture tracts, corresponding to historical events like European colonization and the transatlantic slave trade. Our findings corroborate prior studies by Moreno-Estrada et al. (2013) and Homburger et al. (2015).

  It is well established that longer chromosome segments indicate more recent admixture events. Notably, large Native American ancestry fragments in Peru align with a higher proportion of Native American ancestry, possibly indicating ongoing gene flow from indigenous groups. Across these three countries, the distribution of African ancestry tract lengths shows a consistent pattern of admixture, with shorter tracts corresponding to the era of the transatlantic slave trade (1500s to 1880s). In Brazil, the presence of longer tracts from Central African ancestries coincides with historical accounts of the later phases of the slave trade, which involved Central African populations.

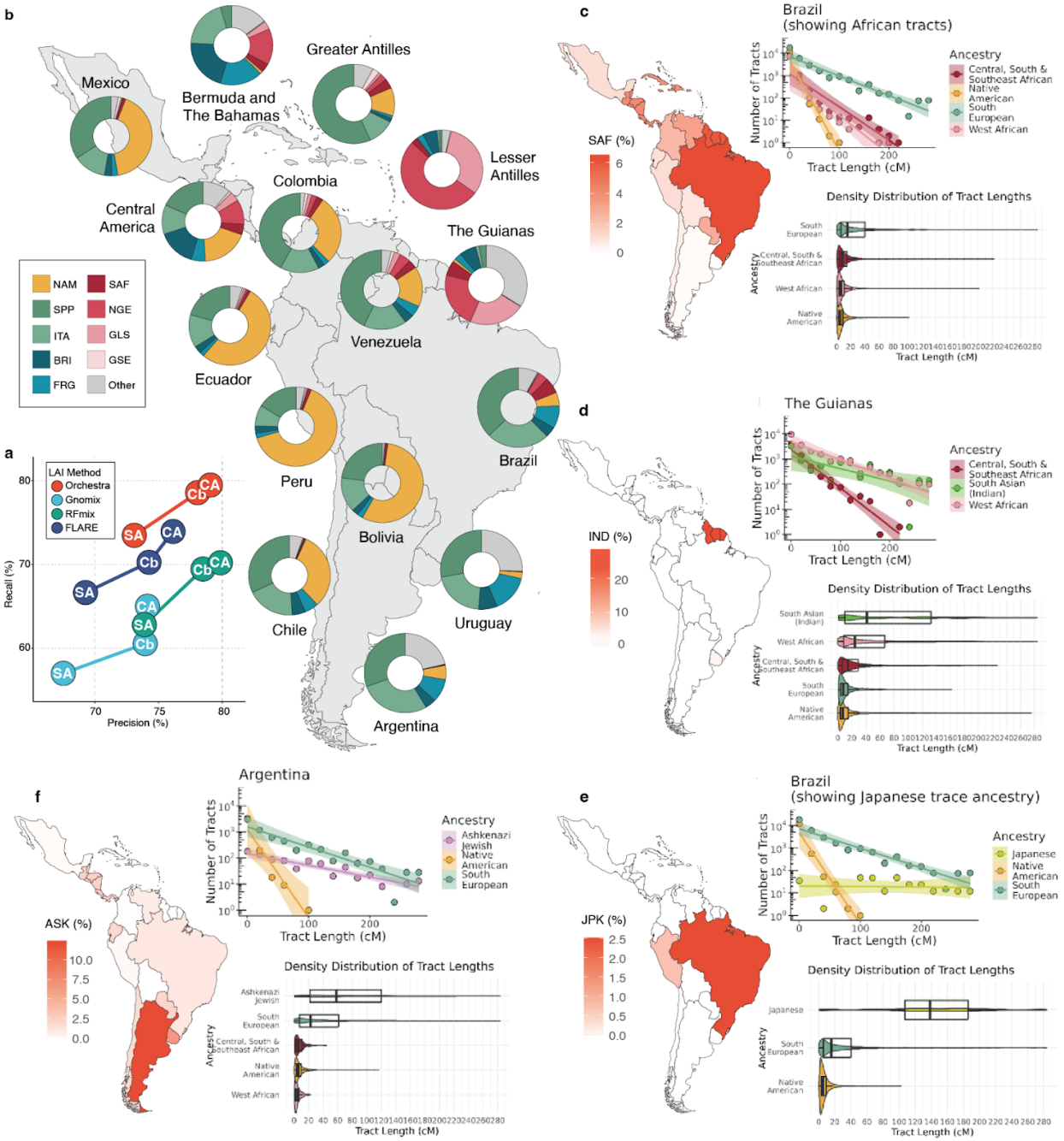
  In Argentina (panel e), longer European ancestry tracts reflect significant European immigration during the Spanish colonial period and later from Southern Europe in the 19th and 20th centuries.

- Panels d, e and f (The Guianas, Argentina and Brazil): These depict ancestries from more recent migrations. In The Guianas, South Asian (Indian) tracts trace back to indentured labor migration from India between 1838 and 1917, with moderate tract lengths suggesting admixture during this period. In Argentina, Ashkenazi Jewish tracts from immigration waves between 1880 and 1920. Similarly, in Brazil, the long Japanese ancestry tracts begin with immigration starting in 1908, especially during the late 1920s and early 1930s, reflecting the more recent introduction of this ancestry. The lengths of these tracts align with the timing of the historical events they represent, with longer tracts signifying more recent admixture.



This figure shows the distribution of local ancestry tract lengths across South American populations, with each panel displaying the number of tracts per ancestry as a function of tract length (in 20 cM bins) and their density distribution. The tract lengths correspond with the timing of these historical admixture events, with longer tracts indicating more recent admixture.

The analysis demonstrates that local ancestry-based methods reveal admixture dynamics and timings not captured by global ancestry techniques, providing precise correlations between chromosome segment lengths and historical events. We appreciate the reviewers' suggestions, which have significantly refined our work. Based on this, we have also created a new alternative Fig 2. (see below).



Alternative Fig. 2 | Ancestry Inference in Latino Americans. Clockwise: (**a**) Percent recall and precision for ancestry deconvolution by FLARE (navy), Gnomix (light blue), RFmix (green), and Orchestra (red) on Latino simulations (equivalent to 12 generations of admixture; we adjusted the simulations to mimic the actual genetic composition of different regions within the continent: CA = Central America, Cb = Caribbean, SA = South America). (**b**) Ancestral composition of 1KGP admixed American populations and UKBB participants that were born in the Americas revealed by Orchestra. Proportions of Native American (yellow), Southern European (green), Northern European (blue) and African (red) ancestries are shown. (**c-e**) Orchestra was also able to detect trace ancestries that reflect known historical population displacements and immigration events. **(c)** Central, South & Southeast African (SAF) ancestry in Brazil, **(d)** various Indian ancestries in the Guianas (IND), **(e)** Japanese & Korean (JPK) ancestry in Brazil and Peru, and **(f)** Ashkenazi Jewish (ASK) ancestry in Argentina. Graphs show the distribution of local ancestry tract lengths, as a function of tract length (in 20 cM bins) and their density distribution.

5.

I will now go into a bit more detail about the three main issues I listed above.

As noted, this is a methods paper, but it **does not provide any way for others to apply or evaluate the described methods**. There is no source code provided, no executable, and the statistical model is not even fully described. In this situation there is no ability for anyone outside the authors to evaluate the method - we must take them on their word. There is no code availability statement in the article, and the data availability statement reads 'Orchestra is available for non-commercial use via the corresponding author' this is not a sufficient way to make science accessible in 2024. It is simply too easy for these requests to get ignored or delayed. The reporting summary document offers some hope and says "our method will be published as a CodeOcean capsule" this could be a good solution, **but it is impossible for me to evaluate, this is not currently available. It may end up containing the full source code and method, but I don't know.**

As mentioned, we have uploaded the full code to GitHub and included toy examples on CodeOcean and AWS, both of which are publicly accessible. This ensures that anyone can access, evaluate and use our method.

We have expanded the Supplementary Methods section by adding a new section with more detailed explanations. This section covers related concerns raised about detailing the steps in generating training/testing sets (see above "Analytical Framework Details and Orchestra Implementation Guide"). We hope this addition provides clarity on the process and the operation of Orchestra.

6.

The potential for overfitting is present in many types of statistical models, and the risk increases as the models get more complex. The neural network-based methods used in Orchestra are complex statistical models with many parameters and so are especially at risk of overfitting. **The effect of overfitting would be that evaluation on the 'test' data set would show improved perforation relative to truly novel input data**. The actual model behind orchestra is not fully described, so it is hard to fully evaluate, but I see multiple possible sources of overfitting.

The first and most obvious possible source is that **the same individuals are used as source haplotypes to generate the synthetic target data and then used as a set of reference haplotypes for inferring LAI**. This leads to increased similarity between the target and reference haplotypes compared to a novel target individual/haplotype. This has long been recognized as an issue when evaluating LAI methods (see e.g. the original RFMix paper or the MOSAIC paper) as there is a sincere desire to generate realistic input data and so real haplotypes are used, but a set of reference haplotypes are also required, so there is a temptation to use one set of haplotypes for both purposes due to limited sample size. The author's even recognized this "To decrease bias resulting from the inclusion of the initial samples in the reference sequences, Orchestra's base layer algorithm was modified to remove the best matching haplotypes from the entire training set first, using the same greedy matching algorithm", but it is not clear that this modification removes this issue. There should be a **complete separation between the haplotypes used as a reference and those used to generate synthetic target data**, and close relatives should also be excluded if they are present at appreciable levels. Another way to deal with this issue is to **work with simulated data (e.g. the FLARE paper) where sample size can be as large as needed, but this involves complex simulations and cannot fully match some aspects of real data sets**.

To provide a comprehensive response to the concerns raised, we will address two main points: first, the separation of reference and target haplotypes to prevent data leakage; and second, clarifying the process described regarding how Orchestra handles training to avoid biases by removing the best matching

haplotypes. Below, we elaborate on these aspects:

1. We were meticulous in preventing data leaks throughout the process. It is incorrect to state that we used the same individuals to generate the simulated target genomes and as reference haplotypes. We divided our single-origin reference panel into separate training and testing sets in the benchmarking analysis (80% and 20% of samples, respectively). Simulations using SLiM were conducted independently for both sets: the training set for training Orchestra and the testing set as the target. Thus, there was complete separation between the reference haplotypes used for training and those used to generate the synthetic target data.

   We recognize that in the past, due to limited sample sizes, there was a temptation to use the same set of haplotypes for both training and testing. However, with access to over 10,000 single-origin individuals from 35 worldwide populations, we did not face such limitations.

   Moreover, if, hypothetically, there had been increased similarity between the target and reference haplotypes due to overlap, it would have impacted all the models tested in our benchmarking, not just Orchestra. Any potential bias or improved performance would have been reflected across all models under the same testing conditions. However, since we maintained complete separation between training and testing haplotypes, this scenario did not occur, preserving the validity of our results.

   Additionally, from the very beginning when designing the reference panel, we excluded close relatives, which meant they were not present in either the training or testing cohorts.

2. Regarding the statement “To decrease bias resulting from the inclusion of the initial samples in the reference sequences, Orchestra's base layer algorithm was modified to remove the best matching haplotypes from the entire training set first, using the same greedy matching algorithm”, we understand the potential confusion. This applies only to the training phase, where Orchestra adjusts parameters by using training set haplotypes as references and evaluates on the simulated genomes within the training set. That is, in summary, the situation you were describing in your concerns about using the same haplotypes in both reference and target sets. However, this is strictly confined to the training process only. Once the model is trained, it is applied to the separate testing set, reinforcing the separation of haplotypes and minimizing overfitting. We revised this section to clarify this point. The following paragraph has been relocated from ‘Simulated Admixed Individuals’ to the end of the ‘Orchestra’ section in Materials and Methods and revised for greater clarity:

- **Original paragraph:**
  “When training the model with a simulated dataset that simulates admixture across multiple generations, it is inevitable that subsequent generations will include direct descendants of the initial samples. To decrease bias resulting from the inclusion of the initial samples in the reference sequences, Orchestra's base layer algorithm was modified to remove the best matching haplotypes from the entire training set first, using the same greedy matching algorithm. It was assumed that those best matching haplotypes belong to the "ancestral" samples of the simulated sample.”

- **Revised paragraph:**
  “During the training phase, where Orchestra adjusts its smoothing layer parameters (weights and bias terms) using simulated admixed individuals, it is inevitable that subsequent generations will include direct descendants of the original samples (see ‘Simulated Admixed Individuals’). This means the same haplotypes are present in both the reference and target sets, potentially leading to an overfitted model. To minimize bias from using source samples of the synthetic admixed data

as reference sequences for population classification, Orchestra's base layer algorithm was modified to remove the best matching haplotypes from the entire training set using a greedy matching algorithm, under the assumption that these best matches represent the 'ancestral' samples of the simulated genomes. Once the model is trained, it can be applied to any separate testing cohort."

We hope this addresses any remaining concerns and highlights our rigorous efforts to ensure robust and unbiased results.

7.

Again the model / method is not fully described, so it is not clear if a trained model **is specific to a particular set of reference haplotypes, or if it is portable and a different reference set can be used**.

The method is not limited to a specific set of reference haplotypes. Orchestra can be trained on any set of reference haplotypes. In fact, in the benchmarking section of the manuscript, we present results using two different reference panels (1KGP-16pops and custom-35pop sets). Additionally, in the updated natural selection analysis section (see below answer to Major Comment 9 from Reviewer #2), we trained the model using different reference panels corresponding to the original source populations of the admixed individuals to identify selection signals. This demonstrates that Orchestra can be applied flexibly with various reference haplotypes. We have clarified this point in the new supplementary section (see above "Analytical Framework Details and Orchestra Implementation Guide").

8.

The article presents a case where the 'test' data is perfectly matched to the perfectly ancestry-annotated training data. So the admixture proportions / generation of admixture / source populations / allele freq-distribution / SNP data QC / SNP array are all perfectly matched between the test and training data sets. This will not generally be the case for novel data - there will not be a perfectly matching set of local ancestry-annotated training data available. A general approach to a situation like this is to include a validation data set that is distinct from the train/test, so that the model is evaluated on how well it generalizes beyond the very similar test/train datasets. This might involve leaving out some of the source populations from the test/train and then using them for the evaluation. Along these same lines the QC steps taken to select variant sites are another possible source of overfitting, as batch effects may have been retained when retaining ancestry-associated sites.

We do agree that, among the elements cited, admixture proportions and the generation of admixture could indeed have been sources of overfitting. The other items, such as source populations, allele frequency distribution, SNP data QC or SNP array, are consistent across training and test datasets for all evaluated software, so we believe there is no unique advantage for Orchestra in these areas. However, we acknowledge that matching admixture proportions and the schema for generating admixture between training and test sets could have provided an advantage for Orchestra. While our goal was to recreate complex admixture scenarios with varying degrees across generations to train Orchestra for diverse situations, this precise match may have unintentionally favored our method. Other software also simulate admixture using their own internal models (e.g., we configured RFmix to simulate six generations of admixture to align with our setup); nonetheless, since each tool generates synthetic admixed training data in distinct ways (all based on the same reference training haplotypes for fair comparison), this may still be a potential source of overfitting. We have included a note in the Discussion section highlighting this as a potential limitation of our study (see below the new paragraph for Discussion).

We also agree that including a distinct validation set would have been ideal, but in local ancestry inference, finding a perfectly annotated dataset with ground truth for validation is challenging. Therefore,

although we did not define a formal validation set, we conducted several analyses equivalent to validation using external samples not included in the training or testing sets and not simulated. Specifically: (i) we applied LAI models to over 10,000 UK Biobank samples with inferred single-origin individuals based on dimensionality reduction, where Orchestra outperformed other LAI methods in 91% of the 103 evaluated countries and was the second-best in the remaining cases (Supplementary Fig. 2c); (ii) updated natural selection analyses (see below answer to Major Comment 9 from Reviewer #2) replicated the majority of known selection signals found using other reference panels and datasets with RFmix, confirming the robustness of our results beyond the testing target set used in the benchmarking. In this regard, we have updated the sentence in the Methods section under 'Benchmarking' to clarify that the analysis using the 10,000 UK Biobank samples with inferred single-origin individuals could be considered an independent dataset distinct from the train and test cohorts, serving as a validation set:

- **Original sentence:** "We assembled a new panel using UKBB samples for additional benchmarking."
- **Revised sentence:** "We assembled a new panel using UKBB samples, providing a distinct and independent validation dataset to assess Orchestra's ability to generalize beyond the original test set"

We have also emphasized this in the Results section:

- **Original sentence:** "In addition to our two panels, we applied all LAI models to over 10,000 UK biobank samples that were not included in the custom-35pops panel"
- **Revised sentence:** "In addition to our two panels, we applied all LAI models to an independent set of over 10,000 UK Biobank samples not included in the custom-35pops panel"

Finally, we acknowledge that our QC steps for selecting ancestry-associated sites might have retained batch effects, despite our efforts to mitigate them. While this selection process might benefit Orchestra—as it was designed to maximize information from the chosen variant set—it may not be optimal for other algorithms. However, we also ran analyses using the 1KGP-16pops panel, a whole-genome sequencing dataset, where Orchestra still outperformed other algorithms. We have added this point to the Discussion to acknowledge this potential limitation:

- **New paragraph for Discussion:**
  "A potential confounder in our benchmarking was the matching of admixture proportions and the schema used for generating admixture between training and testing sets, which may have unintentionally favored Orchestra. Differences in how each software simulates synthetic admixed training data, even when parameters are matched and optimized for best performance on the tested data, could influence results. Despite this limitation, we tested Orchestra and other LAI tools under the same conditions on an independent set of over 10,000 UK Biobank samples containing inferred single-origin individuals identified through dimensionality reduction. In this analysis, Orchestra outperformed other LAI methods in over 90% of the 103 evaluated countries and was the second-best in the remaining cases, demonstrating higher performance at detecting more subtle levels of genetic differentiation, even when admixture was not involved. Additionally, while Orchestra was designed to maximize information from our QC-selected variant set—which may not be optimal for other algorithms developed for array or WGS data—we also tested these methods using a WGS reference panel, where Orchestra continued to excel ."

9.

Finally, all of the authors are from the same company. It's great that the authors want to share their science and to get the stamp of approval that publishing in a peer-reviewed journal provides, but the embedded profit motive (including a relevant patent application) raises extra potential for conflicts of

interest and so sets a higher bar for open science and reproducibility (e.g. the FAIR research software principles published by Nature https://www.nature.com/articles/s41597-022-01710-x, which this article does not try to meet. As noted above, there is no ability of other scientists to evaluate the claims in this manuscript, so it is too close to a press release to be comfortable.

Again, we hope that all the concerns raised have now been addressed. If there is anything further we can do to meet the high bar for open science and reproducibility, we are open to making the appropriate changes.

Small comments

Line 21(Abstract). *"We then deployed it to answer long-standing debates about 22 the ancestral origins of Ashkenazi Jews". This language in the abstract is too strong. It is fine that this article sheds some light by applying the method to Ashkenazi jewish data, and shows broad consistency with other methods, but that is a long way away from "answering long-standing debates". The main text does a better job here, highlighting that this result is broadly consistent with other articles.*

We agree that the original phrasing may have overstated our findings. To address your concern, we have revised the sentence in the abstract as follows:

- **Original sentence:** "We then deployed it to answer long-standing debates about the ancestral origins of Ashkenazi Jews."
- **Revised sentence:** "We then deployed it to weigh in on the debated origins of Ashkenazi Jews, highlighting their South European heritage."

We believe this revised sentence more accurately reflects the scope of our work and maintains consistency with the main text, where we highlight that our results are in agreement with previous studies.

Line 29: *"The remainder, of which only 0.1% is due to SNPs" not fully clear what is meant by this.*

We acknowledge that the original phrasing may have been unclear. Recent advances in genomics have expanded our understanding of the extent of genetic variation among humans. Studies like Pang et al. (2010) and large-scale projects such as the 1000 Genomes Project have revealed that structural variations—including insertions, deletions, inversions and copy number variants—although fewer in number than SNPs, cover a much larger portion of the genome. These structural variants can account for up to 1% to 1.5% of the genome, suggesting that any two humans share approximately 98.5% to 99% of their DNA sequence. This indicates a greater degree of genetic diversity than previously recognized. However, SNPs remain the most abundant type of genetic variation in terms of number. Although SNPs represent about 0.1% of the genome, because they cover fewer base pairs compared to structural variants, their high abundance and widespread distribution throughout the genome make them highly informative for population genetics and ancestry studies. SNPs are sufficient to reveal ancestral origins because they provide a dense set of markers that can distinguish between populations. This is the point we intended to convey in the first two sentences of the introduction. We recognize that this was too ambitious to try to fit into two sentences. We simplified our point as follows::

- **Revised introduction:** Despite vast differences in phenotypes, languages, and culture, any two humans living today are genetically remarkably similar. Yet SNPs, which account for only 0.1% of our DNA, can tell us from which part of the world an individual, or their ancestors, originated1.

Line 33: "allowing us to reference any single individual to well differentiated reference populations" I think it would be better to present LAI as a particularly useful tool - sometime it is useful to conceive of people's ancestry as mixture of well-differentiated source populations, but this is not the only correct way to think about ancestry, nor are the source populations always clear.

We agree that ancestry is complex and cannot always be fully represented as a mixture of well-differentiated source populations. Our intent was to simplify for context, as this is the foundational approach on which current LAI models operate. We appreciate your suggestion and have modified the sentence to present this in a more subtle way:

["Mating between individuals in geographical proximity, coupled with genetic drift and divergent demographic histories, helped shape our modern human genetic landscape,..."]

- **Original sentence:**
  "... allowing us to reference any single individual to well differentiated reference populations"
- **Revised sentence:**
  "... allowing us to reference any single individual to pre-defined reference populations"

Line 49: "Many human populations, despite being geographically close, are genetically heterogeneous." I don't quite get how this sentence works - what is the connection between being close and being heterogeneous?

Our intention was to highlight that geographical proximity does not always result in genetic similarity among human populations, which underscores the importance of considering fine-scale genetic variation (i.e. within-continent diversity) in LAI studies. Various factors such as historical isolation, cultural practices, linguistic differences and social structures can limit gene flow between neighboring groups, leading to significant genetic heterogeneity despite close geographical distances.

To address your concern, we have revised the sentence and the continuation for clarity:

- **Original sentences:** "Many human populations, despite being geographically close, are genetically heterogeneous. "
- **Revised sentences:** "We know that populations that are geographically close can exhibit significant genetic heterogeneity due to historical isolation and/or cultural or geographical barriers."

This ties in well with the preceding paragraph, which ends by expressing regret that the main applications of LAI have been restricted to cross-continental resolution.

Line 55: the authors cite a paper for the polygenetic adaptation on height, but this finding has been called into question by a pair of papers in 2019 https://elifesciences.org/articles/39725 and https://elifesciences.org/articles/39702 I suggest the authors change this example. This is an important change in the manuscript, but does not really substantially change the case for the usefulness of LAI.

Thank you for pointing out the recent studies that question the polygenic adaptation on height. In response, we have removed the example. We have linked a more recent paper that explores heterogeneity in complex traits in Europeans (https://www.nature.com/articles/s41586-023-06705-1), without delving into the selection vs. ancestry debate. We believe that either way, our point that geographically close populations can still be genetically quite different is supported by this evidence.

Figure 1 - *there are no confidence intervals or sample sizes included in this figure, but these are important for interpreting the results.*

Sample sizes for the analyses can be derived from the Supplementary Fig. 1, where the number of individuals in each population is given. Data was split: 80% training and 20% testing. Additionally, for each generation, we simulated twice the number of admixed testing samples as the original cohort. We have added this clarification to the Methods section. We have not included the information into the figure itself, as we consider it would make the figure too complex.

To address the comment on confidence intervals, we calculated 95% confidence intervals for the custom panel with 35 populations by dividing the dataset into 1,000 fragments of equal SNP count and recalculating recall and precision for each fragment. While these results provide insight into the variation in accuracy, we opted not to include this figure in the paper due to the already dense presentation of elements in the figure. Moreover, given the large sample size and high number of SNPs tested, it is our belief that the overall recall and precision metrics are highly reliable and sufficiently informative.

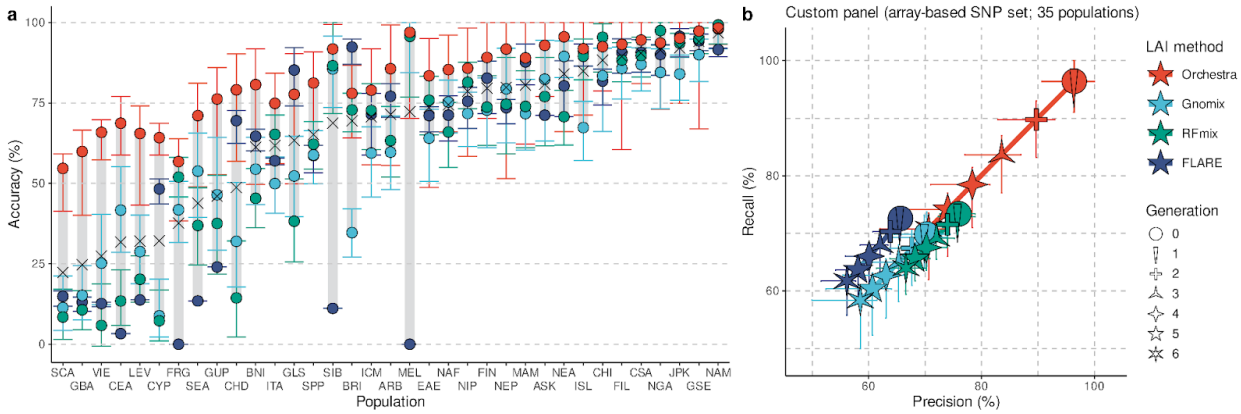


Figure 1 - *it is not clear the admixture history simulated for the evaluations shown here. What populations are mixing and at what rates?*

We apologize that the admixture history in our simulations was not clearly detailed. To address this, we have updated the 'Simulated Admixed Individuals' section in Materials and Methods with additional specifics on the SLiM simulations. As noted in the main text, we used a Wright-Fisher model with modifications: parents were chosen by random sampling with inbreeding avoidance (i.e., close relatives could not mate), and each generation was rerun independently so that parents of simulated individuals were not from prior generations. We have added the following paragraph to clarify further:

- **New paragraph for 'Simulated Admixed Individuals' section:**
  We simulated a fully intermixed scenario where all individuals from populations in the reference panel had an equal probability of contributing to mating, with no specific rates assigned to population mixing, as individuals were chosen entirely at random for each generation. By generation 6, each haploid genome could contain ancestry from up to 32 populations. The expected median number of admixed populations per haploid genome is approximately 1, 2, 4, 7, 13, and 21 for generations 1 through 6. We also implemented a more realistic non-random model where individuals preferentially mate within their continent (e.g., Europeans, Western Asians, and North Africans; Sub-Saharan Africans; Central and South Asians; and East and Southeast Asians), with a 10% migration rate per generation, though benchmarking results were nearly identical. Thus, we opted to train Orchestra in the fully-intermixed random-mating scenario to ensure robustness.

Thus, we have updated our methods to include the simulated admixture history. We chose to train Orchestra in highly admixed scenarios to better prepare it for diverse real-world applications, such as those found in the Americas, enabling it to capture subtle admixture patterns.

Figure 2 - *the sublabels a/b/c do not match the figure caption*

Thank you for pointing that out. We have now addressed the issue.

Reviewer #1 (Remarks on code availability):

No code was provided or available.

Code for training and inference, along with pre-trained models, is now available on our GitHub repository (https://github.com/omicsedge/orchestra-paper). Toy examples replicating the analyses from the manuscript can be found on CodeOcean and AWS. Pre-trained models are available on AWS.

Reviewer #2 (Remarks to the Author):

In this manuscript by Lerga-Jaso et al., the authors propose a new method for local ancestry inference (LAI) and put forth an aggregated reference panel they have curated from prior studies. They benchmark their method against other commonly used LAI tools using simulated data, and then apply it to empirical data to characterize the demographic history of multiple populations and investigate potential indications of past natural selection as indexed by local ancestry enrichment. Their new method indeed seems to outperform the LAI tools against which they benchmark, though I think there are several areas that could use expansion.

Major Comments:

1. A more complete description of the existing tools and how they compare to the proposed model is needed. The introduction does a good job at spelling out the utility of LAI, but currently there is no detail about how existing LAI methods work, where they fail, and what makes Orchestra different/better.

Our main focus in this study was to increase the granularity of the local ancestry. Even though the models we benchmarked were designed differently (RFmix employs a conditional random field; FLARE applies the Li and Stephens model; Gnomix has CovRSK (covariance reduction string kernel) -XGboost as the most successful classifier-smoother combination), we found that they all had the same shortcoming – as we increased the number of reference populations, they became less accurate. In other studies, these LAI methods were applied to 2-way up to 6-way admixture, which is practically continental level admixture (e.g. African, European, East Asian, South Asian, Native American). We wanted to create a LAI that achieves the granularity that was so far only possible with global ancestry. Our hope was that this would help shed light on admixture and evolutionary processes that we were not able to track well before, because it occurred between more closely-related populations (using intra-continental resolution). We have edited the introduction, and hope it now reads more clearly.

We found that out of the models we benchmarked, RFmix was the most accurate as the number of reference populations increased. The downside is that it achieves this accuracy when admixture time is known, which is often not the case with real-world samples. FLARE is more efficient, but that came at a cost of accuracy in highly admixed populations. Finally, Gnomix offered flexibility, but because of that

required extensive parameter tuning without clear guidance due to the continuous nature of ancestry and lack of definitive ground truth. Because they were designed to look at continental-level admixture, these tools have excelled at detecting deep historical admixture across broad geographical scales, but, as mentioned, they still falter with recent, complex admixture scenarios. Orchestra, by contrast, leverages attention and deep learning to prioritize recent admixture events, enhancing precision in more recent generations. This focus provides Orchestra with a unique advantage, although other non-window based methods remain more suitable for detecting deeper ancestral events. Additionally, Orchestra's self-training capabilities and the availability of pre-trained models simplify its application, making it a powerful and user-friendly tool for researchers targeting recent genetic history. We have expanded our discussion to include these points.

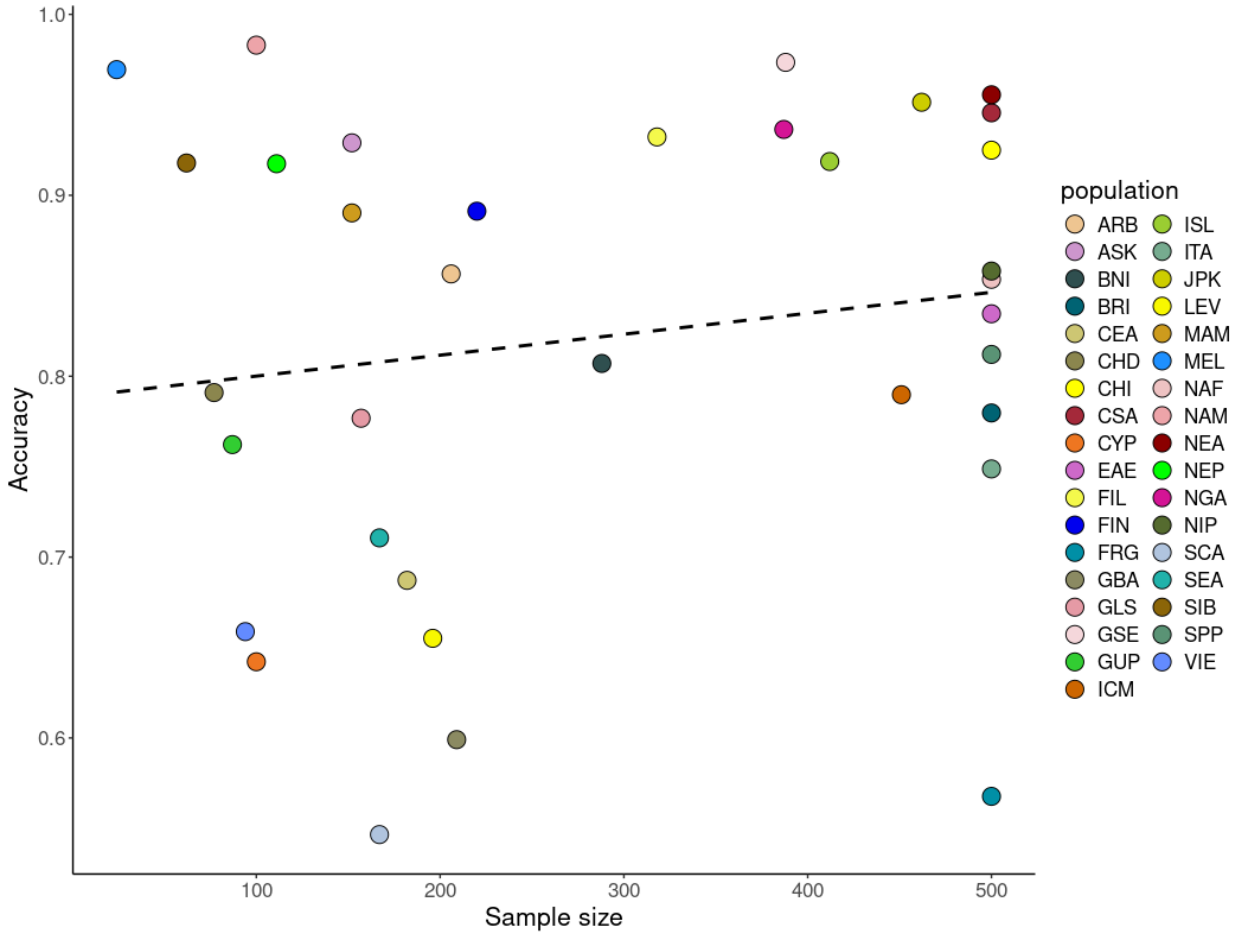
2. Does sample size imbalance in the reference bias prediction? Some populations contain less than 100 individuals, while the biggest are capped at 500 (presumably in an effort to reduce massive sample size imbalance). Did you quantify how big of a concern sample size imbalance is, and/or do you have solutions in place for addressing it?

Indeed, the sample size was capped at 500 to reduce significant imbalances across populations. If we had capped the sample size at 1000, only a handful of populations would have reached that requirement, whereas capping at 500 allowed 10 populations to meet this threshold. If we lowered the cap to 300, then 16 populations would reach the new threshold. Although reducing the capping threshold would help mitigate imbalances, it would also mean less data would be available for analysis. We recognize that imbalance in the reference panels is a concern, particularly when there are more samples from neighboring populations. Even when the reference panel filters out admixed individuals or those bordering others in the PCA, our base layer tends to include shared segments at the local level between populations. If there is a significant imbalance, this can lead to an overrepresentation of segments from neighboring populations being used to reconstruct the genomes of the population under study. Thus, capping at 500 samples per population was primarily a technical decision—a compromise to maintain a sufficient level of accuracy while not excessively increasing the already considerable training time. Increasing the sample size further would have resulted in even longer training times, which was not feasible given the computational demands of our method.

We conducted a small experiment where we capped Sub-Saharan African populations at 300 samples and reran the analysis. Interestingly, while the base layer accuracy decreased, the smoothing layer was able to recover this loss, resulting in accuracy levels equivalent to those achieved with 500 samples per population max. This recovery is likely due to our smoothing layer, which we designed to give higher weights to low-frequency classes (populations) in the loss function to handle class imbalance effectively. We appreciate the reviewer bringing this to our attention, as it reminds us to update the Methods section to include this detail.

We also observed that genetically distinct populations with small sample sizes were not significantly impacted by the lower sample numbers. For example, Melanesians (MEL) with fewer than 30 samples and Ashkenazi Jewish (ASK) with about 150 samples still yielded reliable results. We encountered more issues with populations that were more closely related, such as French and Germans (FRG) vs. British and Irish (BRI), despite having enough samples to fill the 500-sample cap. Indeed, as shown in the plot below, FRG is one of the groups with the full 500 samples yet has among the lowest accuracy. As commented, we found several populations with low sample sizes but very high accuracy, such as Melanesian, Siberian, Nepalese, Ashkenazi Jewish, Manchurian and Mongolian, Finnish, and Native American populations. However, there is a slight trend indicating that a higher number of samples generally improves accuracy, as seen in the group of populations with low sample numbers and lower accuracy, particularly when they are closely related to other populations (e.g., Scandinavian, Greek and

Balkans, Cypriot, Vietnamese, and Levantine). These cases could benefit from reduced imbalances, but decreasing the sample sizes of other populations doesn't seem like the appropriate solution.



This figure illustrates the correlation between the sample size of each population in the reference panel and the accuracy achieved for those populations in the testing set across the six generations evaluated.

In summary, while sample size imbalance is a concern, our approach balances the need for computational efficiency with accuracy. The cap at 500 samples is a practical compromise, and our method, especially the smoothing layer, is designed to mitigate the effects of sample size imbalance. We are grateful for the reviewer's comment, which has helped us refine our discussion and ensure that these considerations are clearly communicated in the manuscript.

3. In the benchmarking against other LAI methods, were the optimal settings for these methods used? For example, in RFmix, there is a flag to account for unbalanced reference sample sizes to improve accuracy. Were these used or were all tools simply run with the default parameters? Testing the competing LAI tools' performance when run in the best way for the use case would be the most fair comparison to running your new method with its optimal parameters.

Thank you for your comment. We gave our best effort to optimize each LAI tool as much as time permitted, given the extensive analyses and runtime for each experiment. The most thoroughly optimized

tool was Gnomix, where multiple models were run with different parameters until we identified the best configuration for our benchmarking.

As noted in the Methods section, we did adjust certain parameters: Gnomix was trained with specific settings (Large mode, smooth_size: 100, context_ratio: 0.10 and window_size_cM: 0.5), RFmix was run with -G 6 to specify generations and we applied a MAF filter to WGS data for RFmix to ensure computational feasibility (as done in FLARE). Specifying admixture time for RFmix, may put this method at an advantage in this case, because with samples of unknown origin, such a parameter would be unknown. However, due to time constraints and the complexity of parameter tuning, not all potential optimizations were explored, such as the suggested flag for unbalanced panels in RFmix, which we were unable to locate.

In the case of RFmix, its slower implementation meant we focused on evident optimizations, like adjusting the number of generations since admixture and filtering rare variants. RFmix was generally robust; but Gnomix, with its extensive parametrization, presented challenges in determining the optimal settings for each scenario, especially in real-life data testing.

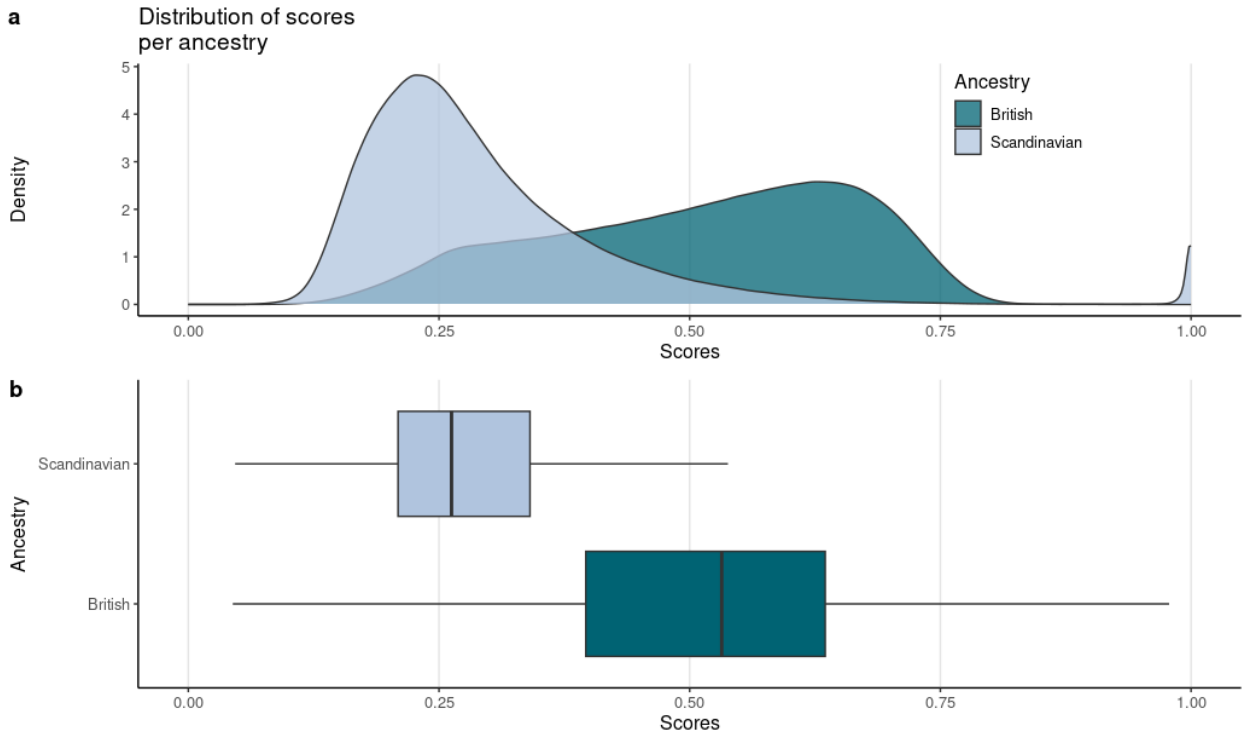
To be honest, we often found that methods that claimed to beat RFmix when they were published, actually performed worse than RFmix when we were testing them, and we believe that is because we did not run any model with default parameters – but tried actively to maximize each algorithm's accuracy. One of the main reasons we used a 16 population 1KG panel, is that this panel is easily obtainable, has been often used in papers, and makes it so it is easy to replicate our findings with any of the LAI algorithms, to obtain similar results.

We believe we made the best possible efforts with the resources and time available and we appreciate your suggestion. We will aim for further optimizations in future analyses based on your feedback.

4. What happens when a user paints a cohort with an ancestry that is not present in the reference populations? Does Orchestra output any metric of confidence for its assignment such that poorly painted regions could be filtered out?

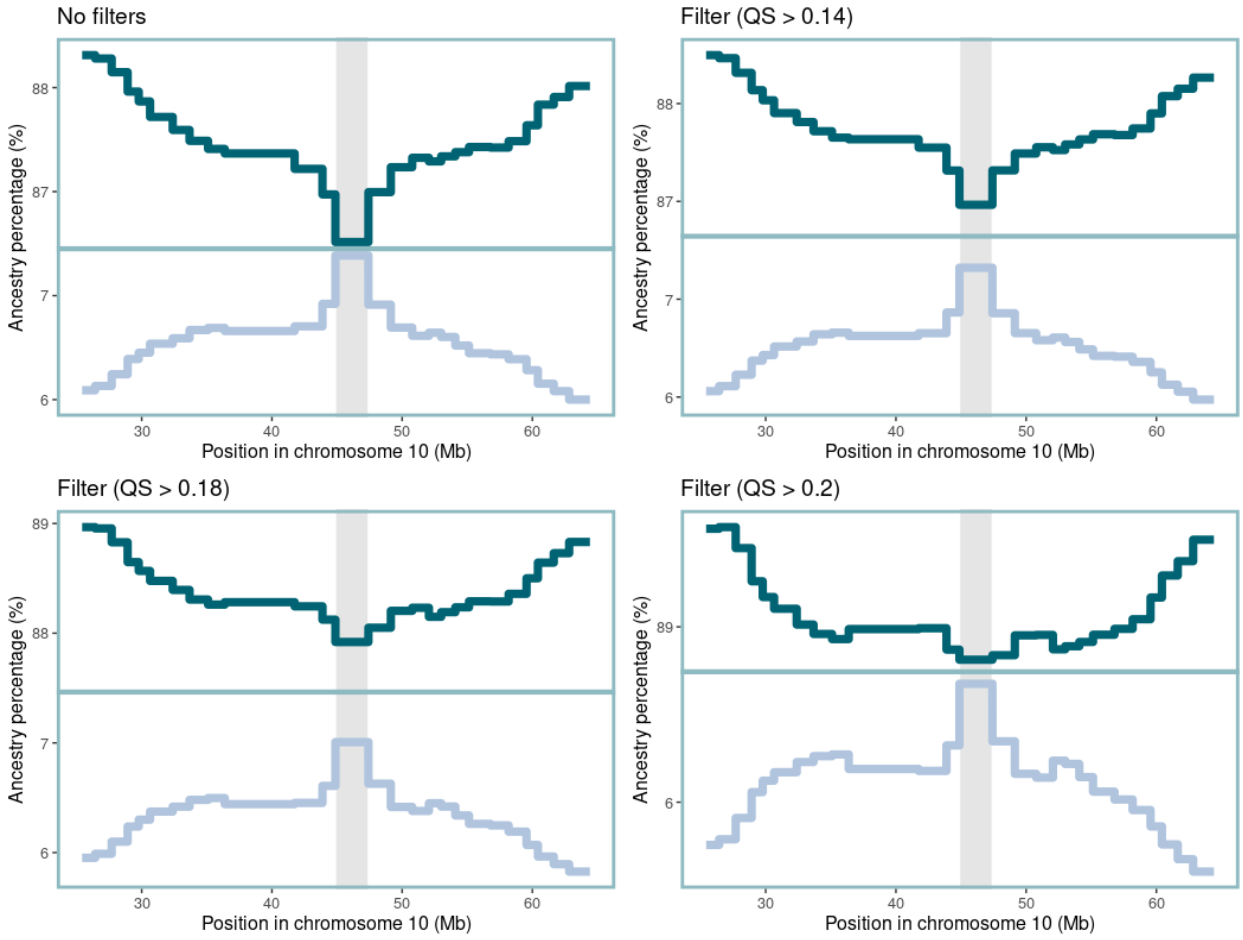
Thank you for your question regarding the assignment of ancestries not present in our reference populations and the related confidence metrics from Orchestra. Orchestra indeed provides a quality metric for ancestry assignments. However, this metric has not been calibrated yet, and therefore, we have opted not to rely on this measure in the current analyses presented within this publication. The existing metric is derived from a deep learning model trained to identify the most likely ancestry per genomic region rather than to accurately quantify the probability of each potential ancestry. Consequently, these scores are inherently biased towards the most likely ancestry and do not accurately reflect the probability of each ancestry.

We plan to address these limitations in a forthcoming publication by recalibrating the output to more accurately reflect the true probability of each ancestry assignment. As a preliminary demonstration, we analyzed chromosome 10 in British samples, where we identified an enrichment of Scandinavian ancestry. By applying different thresholds to exclude low-quality calls, we confirmed that our findings remain robust, even when filtering out less reliable data.



This figure shows the distribution of Orchestra's quality scores for British and Scandinavian ancestry calls within the UKBB British samples. The data indicates that British ancestry calls are generally more reliable, with higher quality scores, whereas Scandinavian calls tend to have lower scores, reflecting greater uncertainty.

The figure above demonstrates that filtering haplotypes based on lower quality scores disproportionately excludes those of Scandinavian ancestry, highlighting the relative difficulty in accurately determining Scandinavian ancestry compared to British ancestry. This is consistent with Fig. 1C from our manuscript, where the Scandinavian population showed the lowest accuracy among all tested populations. To evaluate the impact of quality scores on the observed enrichment, we applied filters to exclude haplotypes with quality scores below 14%, 18% and 20%, corresponding to the exclusion of 1%, 5% and 10% of Scandinavian haplotypes, respectively. This filtering disproportionately affected Scandinavian haplotypes, removing significantly more of them compared to British haplotypes, which generally had higher quality scores. For example, a quality score threshold of 20% led to the loss of less than 2% of British haplotypes, indicating their overall higher reliability. Despite this filtering, the remaining samples, which include only the highest-quality scores, continue to show enrichment on chromosome 10 (see below).



This figure displays the local percentages of British and Scandinavian ancestries within the chromosome 10 window, using different filtering thresholds to exclude low-quality haplotypes.

We recognize the need to further develop and calibrate the confidence metrics, and we plan to address this in future publications to enhance their applicability. The discussion has been updated to introduce this potential metric, while also highlighting current limitations—namely, that it is more focused on correctly identifying ancestry rather than providing calibrated probabilities. We caution users to apply this metric carefully and emphasize the need for refinement in subsequent work.

5. The authors construct their reference panel with both sequence and array data, which is unusual; prior work has limited LAI reference panels to sequence data to capture haplotypes defined by less common variants. Given they later restrict to MAF>5% this may be a moot point, but I wonder at the choice for restricting to such common variants in the reference. This would limit to the use of only older variants that are more likely to be shared across populations; would you not get improved performance by including variants at a slightly lower MAF to capture more private alleles?

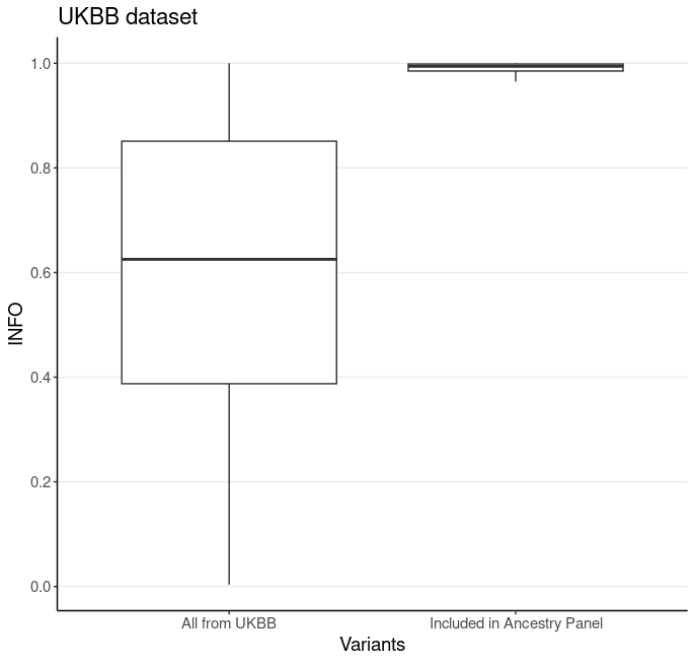
Yes, including variants with a slightly lower MAF could potentially improve performance by capturing more private alleles. However, this approach would limit the data available to us and reduce the usability of data obtained with genotyping chips, which are more affordable and commonly used. Moreover, imputation is generally more reliable for common variants when merging data from different datasets. We believe that by focusing on common variants and accessing larger, more diverse samples, we can more effectively distinguish ancestry than by including rarer alleles, which might introduce higher imputation

biases and result in less reliable ancestry inference due to fewer, less diverse samples. Moreover, this approach maintains a balance between accuracy and computational efficiency, which is crucial given the already heavy computational demands of our method. Nonetheless, we are exploring scenarios where we lower the MAF while trying to maintain data accessibility and we may adjust this approach as more WGS datasets become available in the future.

6. Also, imputation is known to perform less well on diverse populations compared to ones better represented in the reference panel. The authors do several QC procedures to try and account for batch effects, but I do not see INFO score directly used to restrict to high-confidence imputed sites – why is that? Also the authors note "To minimize the impact of potential biases due to imputation, we maximized the presence of directly genotyped SNPs in our final panel by filtering by the union of all genotyped SNP sets contained in the array-based studies employed". Would this not create more risk of batch effects by platform since some variants are directly genotyped and some are imputed depending on the study, and studies contain different and mostly non-overlapping populations? Typical procedure for mega-analyses in GWAS consortia is to impute across arrays separately and then intersect after INFO score filtering.

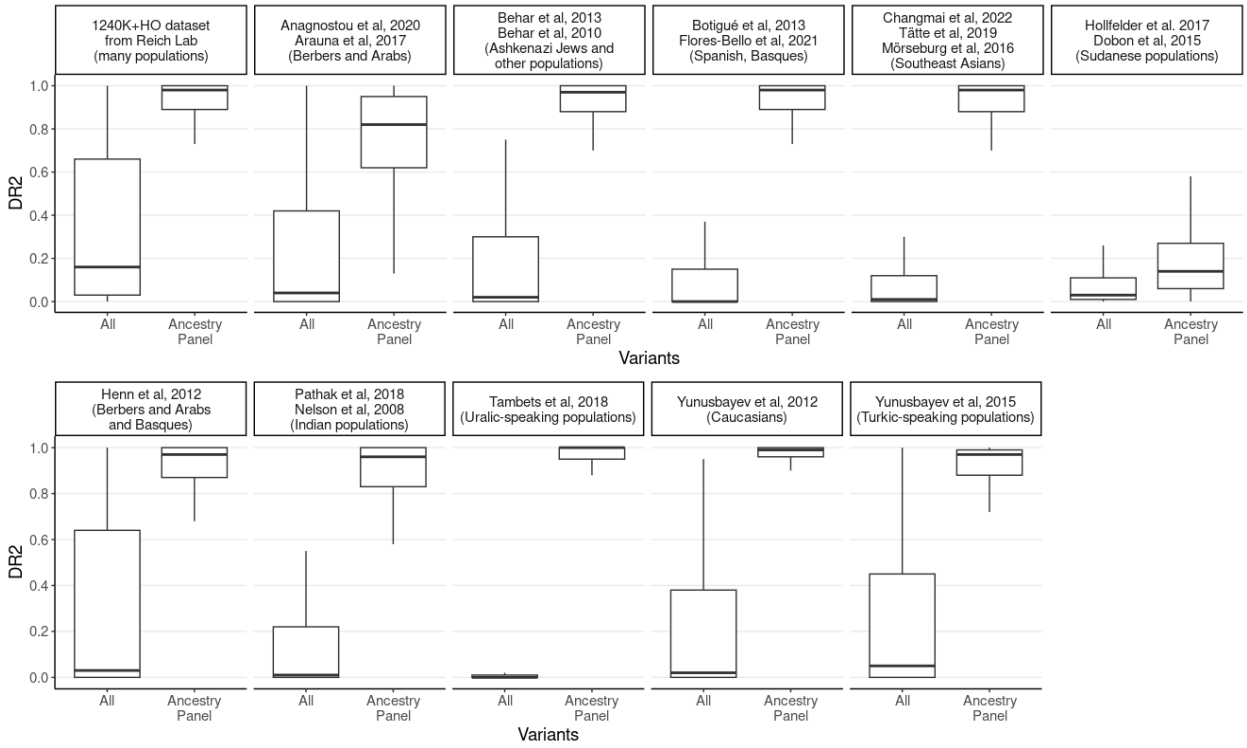
While we did not use the INFO score as a primary criterion for selecting variants in our ancestry panel, we conducted a thorough analysis to assess the imputation quality of these variants post hoc. Specifically, we examined the imputation INFO scores provided by the UKBB for both the full set of imputed variants and those included in our panel.

Our analysis revealed that 99.6% of the variants in our ancestry panel had an INFO score above 0.8, which is a commonly accepted threshold for high-confidence imputed sites. In contrast, only 31.2% of all UKBB variants met this criterion. Furthermore, the median INFO score for the variants in our panel was 0.994, compared to 0.625 for all UKBB variants. These findings indicate that our selection strategy inherently favored variants with high imputation quality, despite not explicitly filtering by INFO score initially (see figure below).

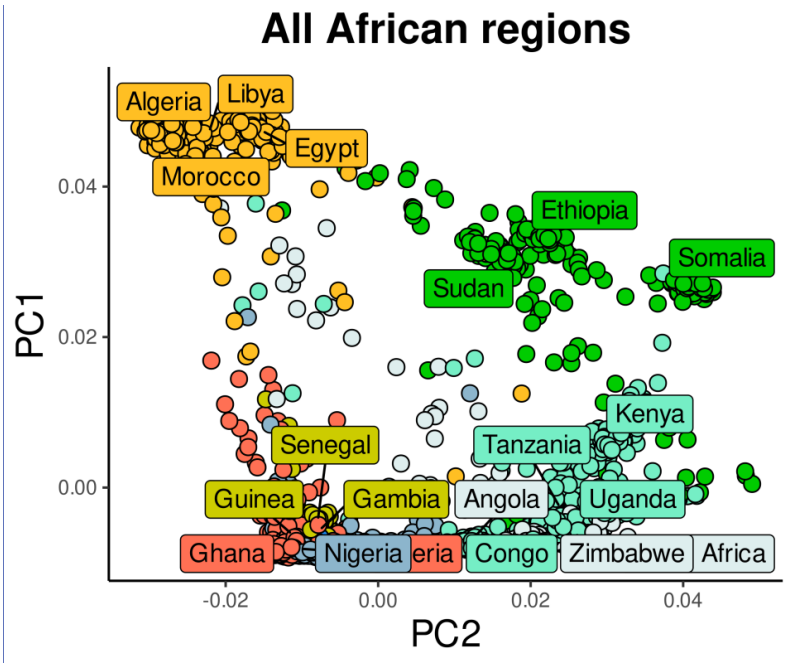


Boxplot illustrating the distribution of INFO scores for imputed variants in the UKBB dataset. The left box represents all variants from the UKBB dataset, while the right box shows the variants included in our ancestry panel.

We extended this analysis to other datasets that were imputed and included in our reference ancestry panel, as detailed in Supplementary Table 1 of the manuscript. Since we imputed these datasets using Beagle software, we utilized the DR2 metric as a measure of imputation quality. The median DR2 value for the variants included in our ancestry panel exceeded 0.96 (see next figure below) across different datasets, demonstrating consistently high imputation accuracy. There were two exceptions to this: (1) Berber and Arab samples from Arauna et al. (2017) and Anagnostou et al. (2020) had a median DR2 of 0.82; and (2) Sudanese populations from Dobon et al. (2015) and Hollfelder et al. (2017) datasets exhibited a much lower median DR2 of 0.14. To mitigate the impact of this lower imputation quality in Sudanese populations, we supplemented them with a substantial number of individuals from the UKBB who clustered appropriately in our principal component analysis (check PCA below). In fact, the Northeast African population demonstrated high accuracy in local ancestry inference (Figure 1c of the manuscript), suggesting that this group was not as challenging to analyze as anticipated. This likely stems from the distinct genetic differentiation of these populations compared to other African regions.



Boxplots depicting the distribution of DR2 scores for imputed variants across various datasets included in our ancestry reference panel.



PCA of UKBB African samples across the first two principal components, with distinct color coding for regional groups: Yellow dots represent North African samples; green dots indicate the Northeast African group; light greenish-blue dots denote Central and South African samples; and red, blue and yellow-greenish dots correspond to samples from West African coastal countries. Country names are labeled at the centroid of samples from each country within this two-dimensional space.

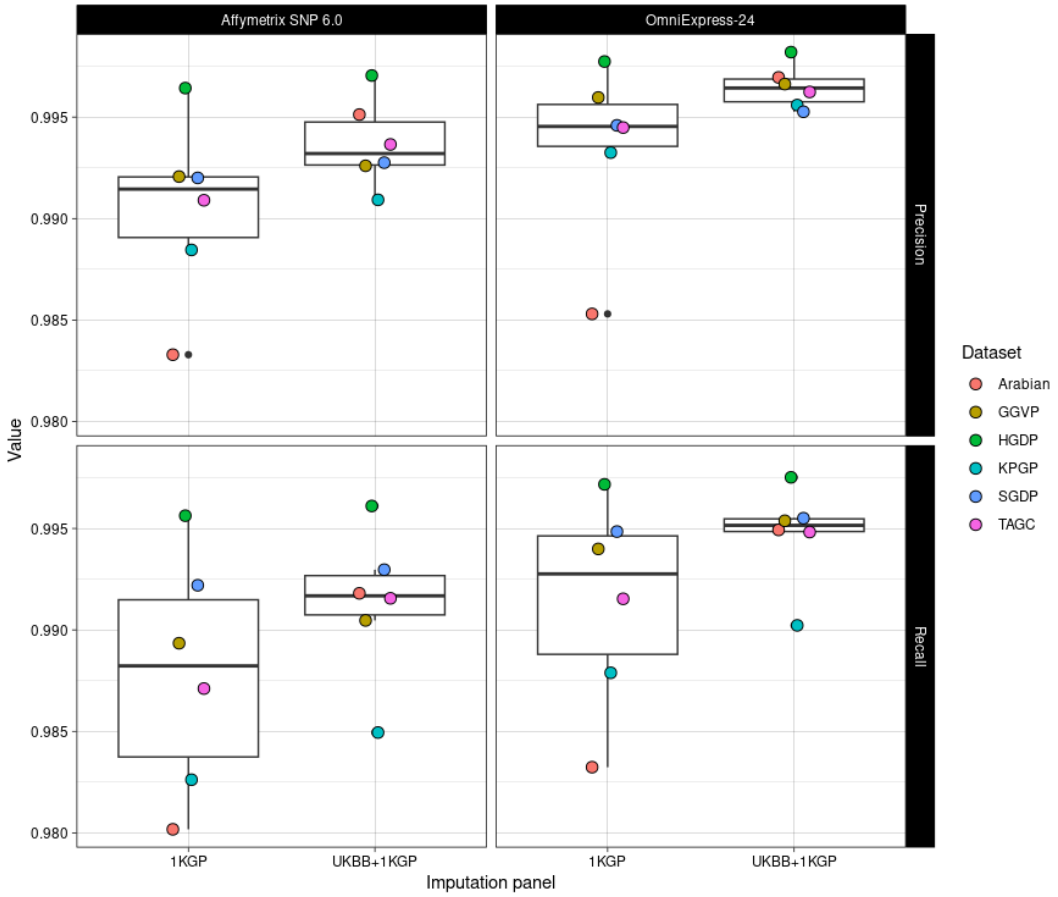
By integrating the union of all genotyped SNP sets from the array-based studies, we substantially increased the proportion of directly genotyped variants in our final panel. This approach, combined with the high imputation quality indicated by robust INFO and DR2 scores, effectively minimized the risk of batch effects due to imputation discrepancies. Our comprehensive quality control measures, which include stringent filtering (MAF > 5%) and statistical associations to remove problematic variants (as described in the 'Batch Effect Removal' section of Supplementary Methods), further mitigated potential batch effects. While the typical GWAS consortia procedure involves separate imputation across arrays followed by intersection after INFO score filtering, our method, although slightly different, achieves a similar goal by ensuring the inclusion of highly imputable variants, as demonstrated by our analysis. We appreciate your feedback and are ready to include these details in the manuscript to clarify our methodology and address any concerns.

a. Related, data were imputed with 1kG, but there are larger and better performing reference panels out now; e.g. TOPmed, the joint 1kG/HGDP callset from Koenig et al. If this reference is to serve as the training data on which all LAI calls are built having the best-in-class imputation done would be vital.

We appreciate your suggestions regarding the use of more recent and diverse reference panels such as TOPmed and the joint 1kG/HGDP callset. We acknowledge the critical importance of utilizing the best available imputation reference to ensure the accuracy and robustness of our LAI predictions. At the time our study was conducted, TOPmed was not accessible to our team, and our choice of reference panel needed to balance diversity, sample size and population representation. Based on our previous analyses (De Marino et al. 2022), we chose to impute with Beagle software using the 1000 Genomes Project (1KG) 30x as a reference panel due to its comprehensive diversity and balance across different populations. However, we acknowledge that panels like HGDP offer broader geographic and genetic

diversity, albeit with fewer samples per population, which is crucial for filling geographical representation gaps not covered by the 1KG.

To address potential concerns about the impact of using the 1KG panel alone, we conducted an additional experiment where we imputed masked non-array variants in whole genome sequencing datasets (see figure below). These were then imputed using both our original 1KG panel and a new reference panel composed of UK Biobank samples (originally imputed using the Haplotype Reference Consortium and UK10K reference panels). The results demonstrated high precision and recall (near 1) for both panels, though as expected, the UKB+1KG panel showed slightly superior performance. This confirmed that our genotype imputation calls maintain high accuracy, particularly for the common variants (>5% MAF) that are the focus of our analysis.



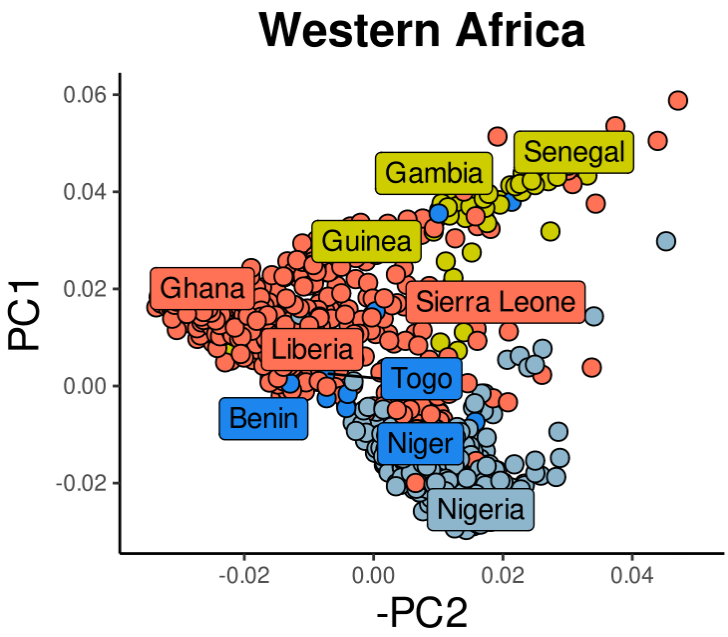
This figure displays the accuracy of imputed genotypes compared to actual WGS calls across various datasets, using two different reference panels: 1KG and UKBB+1KG. Precision and recall were used to assess the quality of imputation when non-array variants are masked and re-imputed. The left and right columns distinguish the performance metrics obtained using Affymetrix SNP 6.0 and OmniExpress-24 platforms, respectively, for simulating array data and masking non-array variants in WGS datasets. Given that our analysis primarily focused on common variants (MAF > 5%), we expect to achieve even higher accuracy. Datasets used: Arabian (Almarri et al. 2021), GGVP (Gambian Genome Variation Project), HGDP (Human Genome Diversity Project), KPGP (Korean Personal Genome Project), SGDP (Simons Genome Diversity Project) and TAGC (The Ashkenazi Genome Consortium).

Moving forward, we are committed to incorporating these insights and exploring broader and more diverse reference panels in order to ensure the highest quality of our genetic analyses in future projects.

7. Some discussion of grouping for ancestry labels is warranted. For example, why are multiple diverse African countries given a single ancestry label (e.g. GLS = Ghanaian, Ivorian, Liberian, & Sierra Leonan)?

The granularity we were able to achieve on each continent was largely dependent on the number and diversity of samples we had available. Africa, in particular, we believe has the greatest potential for increased granularity, once more samples become available. Where samples were limited or not divergent enough based on initial experiments, we grouped populations into meaningful broader geographic regions with shared genetic ancestry.

For example, our PCA analysis of Western African samples did reveal subtle genetic distinctions within the GLS group that could potentially separate these populations (see figure below), but unfortunately we faced major constraints due to sample size. Specifically, after filtering for potential admixture, notably with the Nigerian cluster, we were left with a limited number of samples -only 6 samples from Ivory Coast and 2 from Liberia- which restricted our ability to define more granular ancestry groups.



This figure displays a PCA of UKBB Western African samples across the first two PCs. Red dots represent samples from the GLS group (Ghana, Liberia, Sierra Leone and Ivory Coast), green points indicate Senegambian and Guinean samples, light blue points denote Nigerian samples, and dark blue points correspond to samples from other countries situated between the GLS group and Nigeria. The labels with country names are placed at the mean position of all samples from that country in this two-dimensional space.

On the other hand, we had a larger sample size for the French & German population – which also includes Dutch, Belgian, Swiss and Austrian samples. However, in this case, the diversity in these populations was not enough to tell them apart successfully on a local level.

We have added the PCA figures into the supplementary section, and a brief explanation to the methods section.

8. The authors reference historical events explaining the presence of some cryptic ancestries in modern populations. Do the local ancestry tract lengths match the expectations for the timing based on

these historical records? This would be most convincing; as it is one could conceivably explain the presence of most ancestries based on some past event, so demonstrating the timing of relevant genetic admixture events is concordant with records would be more direct.

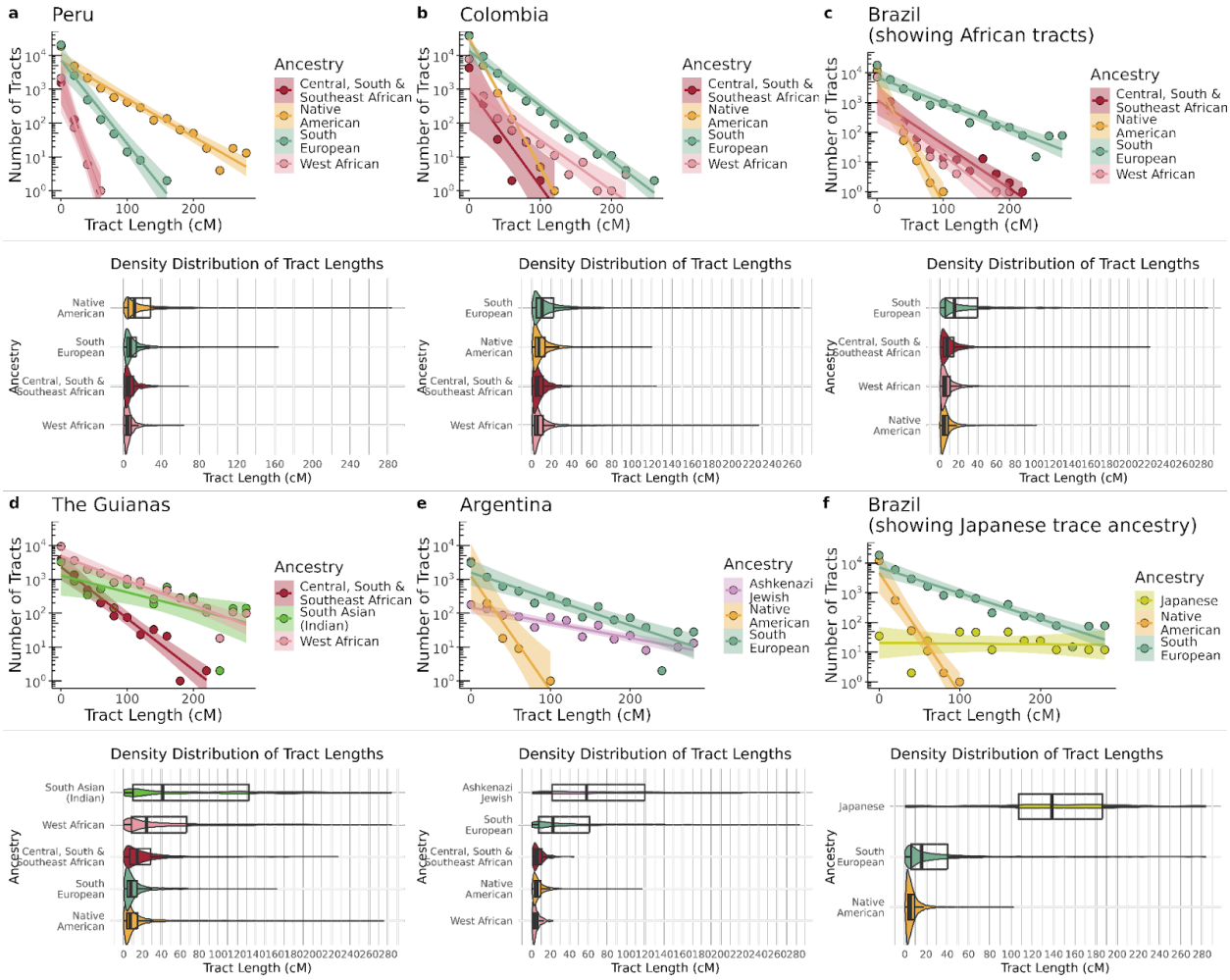
Thank you for your valuable feedback. We agree that aligning local ancestry tract lengths with historical records would strongly validate our findings, particularly regarding cryptic ancestries in modern populations. In response, we have generated the attached plot, which will be added as a supplementary figure in the manuscript. This plot provides a detailed comparison of local ancestry tract lengths across various populations and aligns well with expected historical events.

It is well established that the length of chromosome segments from distinct ancestries reflects the number of generations since admixture, with longer segments indicating more recent admixture events. Our analysis of tract lengths in three countries with known differences in ancestry proportions -Peru, Colombia and Brazil (panels a, b and c; see below)- shows high concordance with previous studies by Moreno-Estrada et al. (2013) and Homburger et al. (2015), further supporting the validity of our approach.

Key observations include the unique presence of large Native American ancestry fragments in Peru, consistent with the higher proportion of Native American ancestry in this population and may suggest recent or ongoing gene flow from indigenous groups. Similar distribution of African ancestry tract lengths across countries suggests parallel dynamics of African admixture, with shorter tracts reflecting the timeline of the transatlantic slave trade (1500s to 1880s). In Brazil, the presence of larger Central African ancestry tracts aligns with historical records of the later stages of the slave trade, which increasingly involved Central African populations.

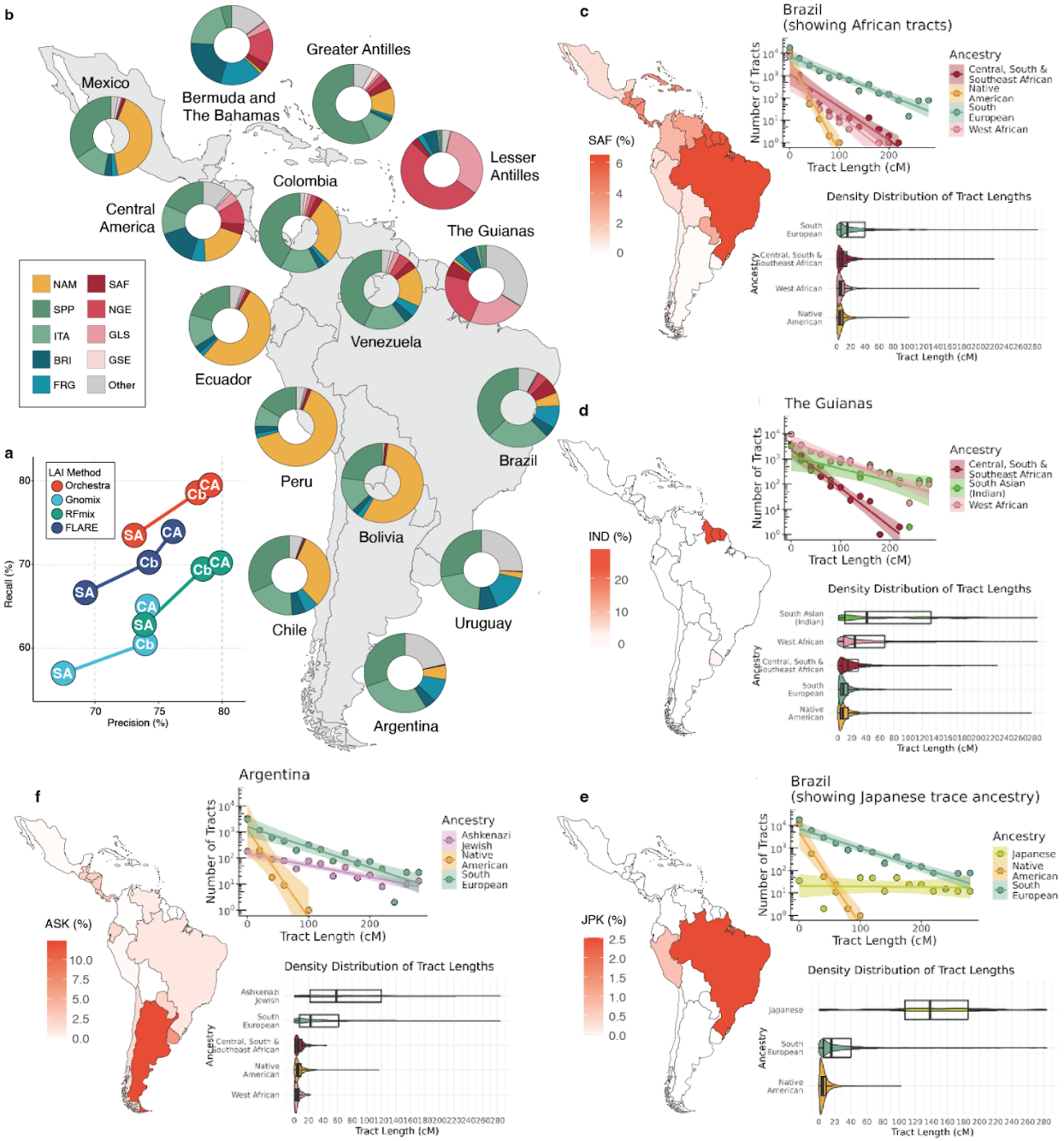
Panels d, e, and f highlight cryptic ancestries in South America: South Asian (Indian) in The Guianas, Ashkenazi Jewish in Argentina, and Japanese in Brazil. These ancestries correspond to specific historical migrations. The South Asian (Indian) tracts in The Guianas align with indentured labor migration from India between 1838 and 1917, reflecting moderate tract lengths that suggest admixture during this period. In Argentina, Ashkenazi Jewish ancestry is evident in tract lengths that correspond to Jewish immigration waves between 1880 and 1920. The Japanese ancestry in Brazil, with the longest tract lengths, reflects the more recent immigration beginning in 1908, particularly during the late 1920s and early 1930s. These tract lengths provide a timeline for the introduction and integration of these ancestries into the populations, directly aligning with historical records.

In Argentina, the predominance of longer European tracts reflects the significant European immigration during the Spanish colonial period and subsequent influx from Southern Europe in the 19th and 20th centuries.



This figure shows the distribution of local ancestry tract lengths across South American populations, with each panel displaying the number of tracts per ancestry as a function of tract length (in 20 cM bins) and their density distribution. Panels a (Peru), b (Colombia) and c (Brazil) illustrate Native American, South European and African admixture, reflecting European colonization and the transatlantic slave trade. Panels d (The Guianas), e (Argentina) and f (Brazil) depict trace ancestries from specific historical migrations: South Asian (Indian) tracts from 19th and early 20th-century indentured labor migration; Ashkenazi Jewish tracts from late 19th and early 20th-century Jewish immigration; and long Japanese tracts indicating immigration starting in 1908. The tract lengths correspond with the timing of these historical admixture events, with longer tracts indicating more recent admixture.

These findings demonstrate that the local ancestry tract lengths match historical expectations, providing direct evidence for the timing and dynamics of genetic admixture events that have shaped modern populations. We have incorporated this analysis into the revised manuscript to strengthen our results and better link our genetic findings with historical context. We appreciate the reviewer's suggestion, which has enhanced our work and we welcome further feedback. Based on these, we have also created a new alternative Fig 2. (see below).

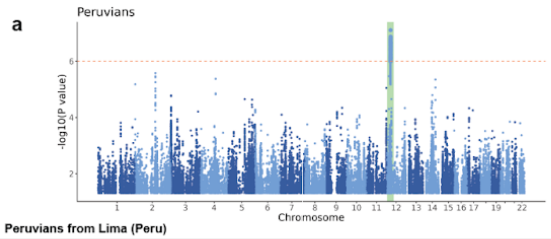


Alternative Fig. 2 | Ancestry Inference in Latino Americans. Clockwise: (**a**) Percent recall and precision for ancestry deconvolution by FLARE (navy), Gnomix (light blue), RFmix (green), and Orchestra (red) on Latino simulations (equivalent to 12 generations of admixture; we adjusted the simulations to mimic the actual genetic composition of different regions within the continent: CA = Central America, Cb = Caribbean, SA = South America). (**b**) Ancestral composition of 1KGP admixed American populations and UKBB participants that were born in the Americas revealed by Orchestra. Proportions of Native American (yellow), Southern European (green), Northern European (blue) and African (red) ancestries are shown. (**c-e**) Orchestra was also able to detect trace ancestries that reflect known historical population displacements and immigration events. **(c)** Central, South & Southeast African (SAF) ancestry in Brazil, **(d)** various Indian ancestries in the Guianas (IND), **(e)** Japanese & Korean (JPK) ancestry in Brazil and Peru, and **(f)** Ashkenazi Jewish (ASK) ancestry in Argentina. Graphs show the distribution of local ancestry tract lengths, as a function of tract length (in 20 cM bins) and their density distribution.

9. The authors note a somewhat surprising lack of concordance in detection of selection signals between their work and a prior publication. They explain their failure to replicate some peaks as due to different reference panels, but is that really enough to completely change signals? Also they don't just fail to detect some peaks that were previously found, they also detected new peaks as well. Are these real or false positives? The presence of multiple new signals isn't discussed in the text at all.

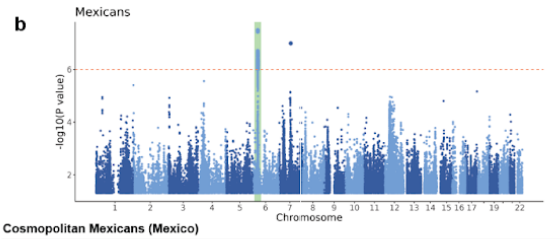
Regarding the lack of replication of some selection signals, we believe the following factors contributed: (i) our reference panel included broad populations that may have overlapped with the target admixed populations or closely related groups (e.g., Fula samples were already included under Gambian and Senegalese population in our custom-35pops panel); (ii) we used reconstructed Native American genomes combining North and South Native American fragments, which may not have been optimal for detecting signals in specific populations like admixed Peruvians or Mexicans, where region-specific Native American (from South or North, respectively) references were more appropriate; (iii) target populations sometimes differed, such as using Peruvians from UKBB instead of 1KGP, or lacking suitable proxies for certain groups like the Malagasy; (iv) our 35-population panel introduced greater diversity compared to the 2-3 curated and more focused panels used in Cuadros-Espinoza (2022), making it more challenging for Orchestra to distinguish signals; (v) differences in sample sizes between our study and the prior work; and (vi) potential artifacts or false positives cannot be entirely ruled out. Additionally, our model was trained with six generations of admixture, which may not align with target populations whose admixture occurred much earlier (e.g., Fulani admixture estimated at ~63 generations ago).

To address these issues, we reanalyzed our data using revised models and reference panels that closely align with those used in Cuadros-Espinoza et al. (2022), or appropriate proxies where exact matches were unavailable. As a result, we successfully replicated the previously reported selection signals, with the exception of two signals in the Sahelian Arabs. This demonstrates that Orchestra is robust and capable of detecting local selection signals when appropriate reference data is used.



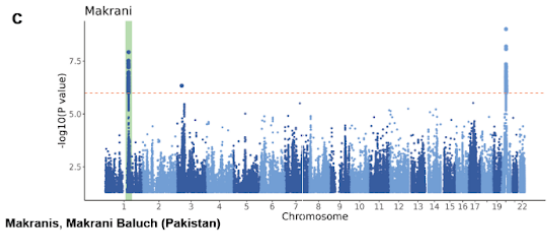
Peruvians from Lima (Peru)

Study	Dataset	N	Source populations vs. main inferred ancestries [admixture proportions]
Cuadros-Espinosa et al. 2022	1KGP (The 1000 Genomes Project Consortium, Nature 2015)	85	East Peruvian Native Americans (Quechua, Aymara) [0.77] / Iberian populations from Spain (IBS) [0.20] / Sub-Saharan Africans (GWD, LWK, YRI) [0.03]south Bantu speakers (Luhya, Sotho, Zulu) [0.60] / Mandarese [0.40]
This study	Peruvians from 1KGP	87	South Native Americans [0.60] / Iberians [0.38] / Sub-Saharan Africans [0.2]



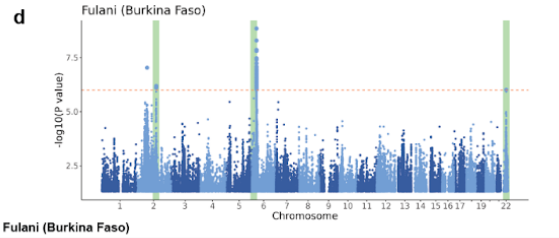
Cosmopolitan Mexicans (Mexico)

Study	Dataset	N	Source populations vs. main inferred ancestries [admixture proportions]
Cuadros-Espinosa et al. 2022	Moreno-Estrada et al., 2014	418	Mexican Native Americans (Maya Campeche, Tepehuano, north Zapotec) [0.54] / Iberian populations from Spain (IBS) [0.42] / Sub-Saharan Africans (YRI) [0.04]
This study	Mexicans from 1KGP	65	North Native Americans [0.54] / Iberians [0.44] / SubSaharanAfrican [0.02]



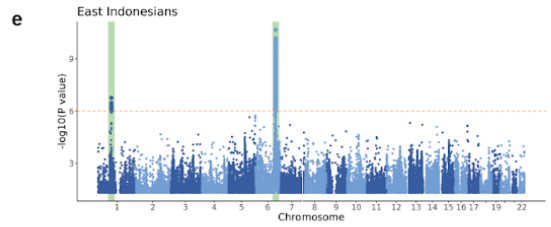
Makranis, Makrani Baluch (Pakistan)

Study	Dataset	N	Source populations vs. main inferred ancestries [admixture proportions]
Cuadros-Espinosa et al. 2022	EGAS00001002558 (Laso-Jadart et al. 2017)	46	Baluch [0.85] /east and south Bantu speakers (Luhya, Sotho) [0.15]
This study	Makrani samples from Reich lab	24	Proxy for Baluch (individuals from Pakistan) [0.98] / Bantu and Luhya people, individuals from Lesotho [0.02]



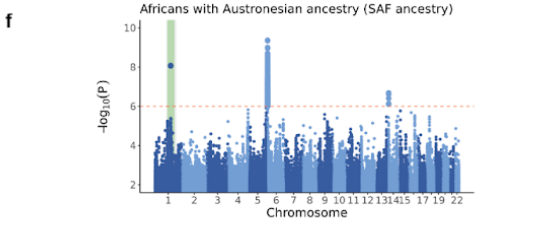
Fulani (Burkina Faso)

Study	Dataset	N	Source populations vs. main inferred ancestries [admixture proportions]
Cuadros-Espinosa et al. 2022	E-MTAB-8434 (Vicente et al., 2019)	53	West Africans (Jola, Igbo) [0.77] / north Africans (Mozabite) + Europeans (CEU, Czech) [0.23]
This study	E-MTAB-8434 (Vicente et al., 2019)	53	Jola people [0.87] / North Africans (Mozabite) + Europeans (Czech, French and British individuals) [0.13]



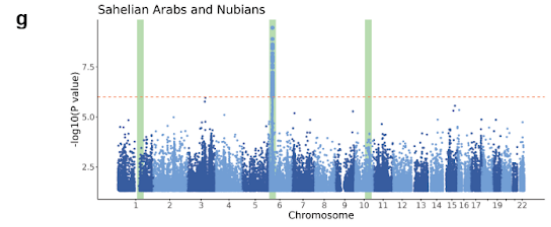
East Indonesians: Flores Bena & Bama, Sumba, Timor, Lembata, Alor, Pantar (Indonesia)

Study	Dataset	N	Source populations vs. main inferred ancestries [admixture proportions]
Cuadros-Espinosa et al. 2022	GSE80534 (Hudjashov et al., 2017)	201	Filipinos [0.58] / Papuans [0.42]
This study	GSE80534 (Hudjashov et al., 2017)	201	Custom-35pops panel: SEA [0.92] / FIL [0.04] / MEL [0.03]



Malagasy (Madagascar)

Study	Dataset	N	Source populations vs. main inferred ancestries [admixture proportions]
Cuadros-Espinosa et al. 2022	EGAS00001002549 (Pierron et al., 2017)	700	East and south Bantu speakers (Luhya, Sotho, Zulu) [0.60] / Mandarese [0.40]
This study	Southeast African UKBB participants with >1% Austronesian heritage	221	BEI [0.31] / SSI [0.25] / SAF [0.17] / PNI [0.08] /FRG [0.03] / BRI [0.03] / SEA [0.03]



Sahelian Arabs: Bataheen, Gaalien, Shaigia, Messiria & Nubians: Danagla, Mahas, Halfawieen (Sudan)

Study	Dataset	N	Source populations vs. main inferred ancestries [admixture proportions]
Cuadros-Espinosa et al. 2022	http://jakobsson-lab.iob.uu.se/data/ (Hollfelder et al., 2017)	97	East Africans (Gumuz, Somali) and Nile peoples (Baria, Dinka, Nuer, Shilluk) [0.50] / South Europeans (Tuscans, Sardinians) and Druze and Bedouins [0.50
This study	http://jakobsson-lab.iob.uu.se/data/ (Hollfelder et al., 2017)	97	Nile peoples [0.56] / South Europeans (Tuscans, Sardinians) and Druze and Bedouins [0.44]

We have updated the Results and Methods sections of our manuscript to reflect these findings. This reanalysis has significantly enhanced the reliability of our study, confirming that Orchestra can effectively

replicate known signals identified by other software. We have also removed earlier results that may have been misleading to prevent any misunderstandings. Your feedback has been invaluable in improving our work, and we appreciate the opportunity to clarify these points.

10. There is currently no benchmarking of computational efficiency across this and competing methods. This is an additional important consideration for users given how computationally intensive LAI is, and some discussion of the runtime/computational load required for various methods would be valuable.

You raise a critical point regarding the computational demands of LAI, which indeed vary significantly among the methods we assessed. Our method, in particular, is substantially more computationally intensive than the others we compared it against. While other methods can be executed on standard servers, our approach requires high-performance computational resources. This is due to our models undergoing extended training periods, ranging from several days to weeks, and demanding significant memory resources.

Given this considerable disparity, a direct comparison of computational efficiency across these methods is not feasible, as our method operates on an entirely different scale. However, despite the heavy computational demands during training, we have made pre-trained models available for end-users, which can be used for rapid inference. We have also provided a range of models trained on different populations and regional levels, ready for immediate use. For users who may need to retrain models for specific analyses, we have shared the necessary code and framework, though it is important to note that the computational resources required for this will greatly exceed those needed for other LAI software.

In response to your feedback, we have updated the discussion section of our manuscript to explicitly address these computational considerations. We believe that the superior performance of our method may be linked to the extended training times, which allow the model to capture finer details that shorter and less memory-intensive training processes might miss. We have outlined the availability of our pre-trained models and provided guidance on when new model training might be necessary. Additionally, for users seeking quick LAI results with a model not yet developed by our team, we recommend considering alternative software like RFmix, which has consistently demonstrated robust and reliable results across diverse populations at a region-wide level.

We have incorporated these details into the revised manuscript to ensure clarity regarding the computational requirements of our method. We appreciate the reviewer bringing this important aspect to our attention and for helping us improve our discussion. We remain open to training new models upon request and welcome collaborations that leverage our computational resources.

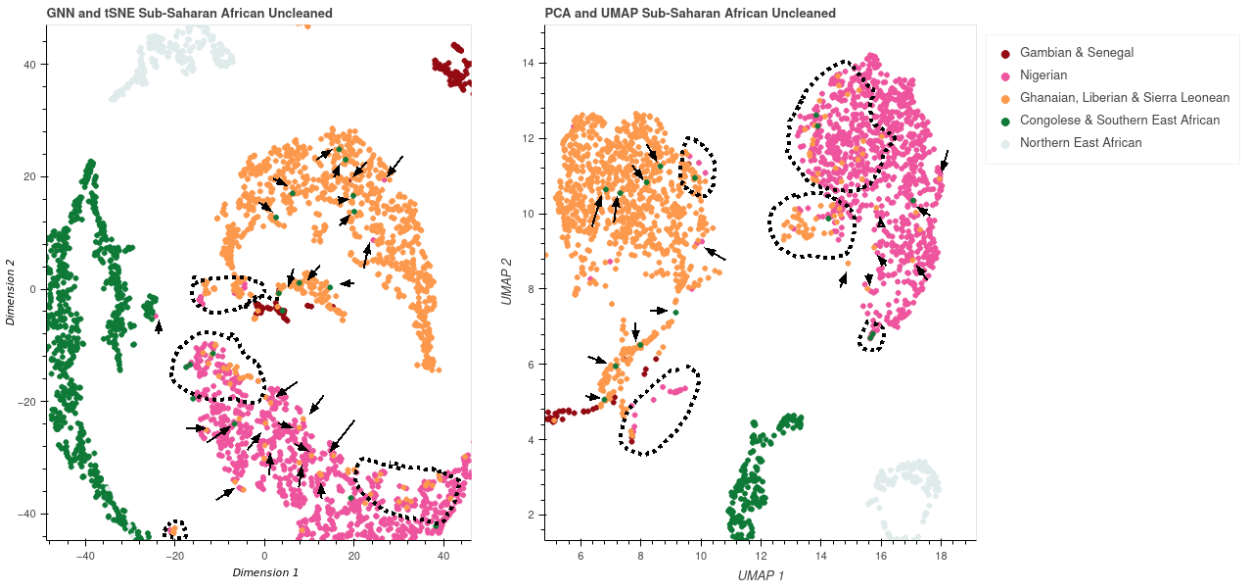
Other comments:

a. *How did you decide on these dimensionality reduction techniques? As has recently come to the fore, UMAP is typically not the best choice for representing ancestry, as it will discretize populations more than they are in actuality (ancestry is continuous). UMAP is not featured here for anything other than reference panel curation, but the authors may be cautioned about pushback for its use in an ancestry context.*

Thank you for highlighting the concerns regarding our use of dimensionality reduction techniques, particularly UMAP. We acknowledge that UMAP can over-discretize the continuous nature of ancestry into discrete clusters, creating artificial boundaries. However, this characteristic was strategically employed in our study to enhance the curation of our reference panel. Specifically, we applied UMAP to

the first principal components of PCA to accentuate sample relationships that might be obscured by PCA alone, as UMAP tends to amplify denser data regions. This combined approach leveraged PCA's ability to capture large-scale genetic structures and UMAP's capacity to elucidate finer-scale interactions, aiding in the identification and exclusion of admixed samples. Indeed, we did not use this technique for sample selection but rather to remove potentially admixed individuals. Following their removal, and if a sufficient number of samples remained, we randomly selected individuals from the residual pool as references for specific populations. Given our focus on excluding individuals who clearly clustered with different population groups, we believe that using UMAP for assembling the reference ancestry panel poses no harm in this context.

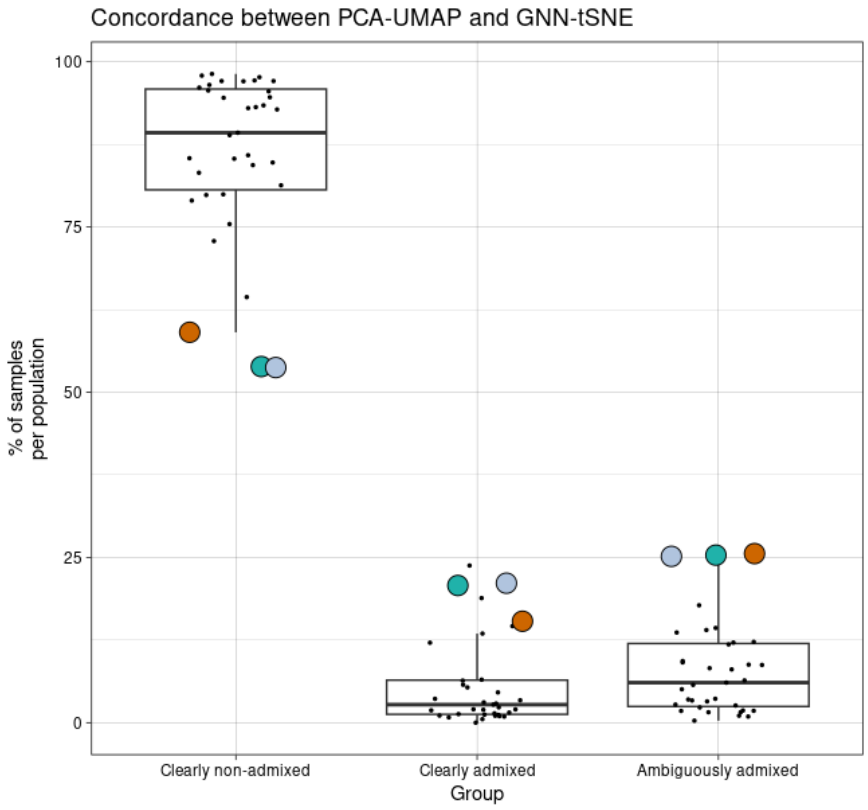
Moreover, to enhance the robustness of our selection process, we complemented the PCA-UMAP analysis with t-SNE applied to genealogical nearest neighbor (GNN) statistics from tsinfer results. This 'dual-filtering' strategy enabled us to cross-validate the ancestry consistency of each sample and reduce the biases inherent in each dimensionality reduction method. The figure below provides an example of a 2D visualization using both techniques, where mislabeled or admixed individuals associated with populations of opposing ancestry are evident. These individuals were excluded from further analyses due to their potential admixture.



This figure presents a 2D visualization of Sub-Saharan UKBB samples using the dimensionality reduction techniques described earlier. Arrows and circles highlight samples that are located in different population clusters than expected, indicating potential admixture or incorrectly assigned/reported ethnicity.

Additionally, for populations exceeding 300 individuals, given the ample availability of samples, we included samples only if they met the filtering criteria established by both PCA-UMAP and GNN-tSNE analyses, meaning all samples identified as admixed by either method were excluded. We then employed random selection to ensure a broad representation of the reference population, choosing samples from various positions within the cluster rather than focusing solely on those at the center. For smaller populations, individuals were removed only if flagged by both methods concurrently. This approach minimized the exclusion of potentially representative samples due to the peculiarities of a single method, ensuring adequate sample sizes for accurate population representation.

The figure below shows that the concordance rates between these strategies were generally high, though there were notable exceptions in a couple of Asian populations and the Scandinavian population. This careful evaluation of methodological overlaps ensured that samples consistently clustered with their reported ancestry across different analyses, focusing primarily on removing admixed individuals rather than selecting artificially 'pure' samples.



This figure illustrates the concordance between PCA-UMAP and GNN-tSNE techniques. The "Clearly non-admixed" group encompasses samples from the reported population correctly clustered according to expectations. The "Clearly admixed" group contains individuals identified as admixed by both techniques, showing clustering in unexpected populations. The "Ambiguously admixed" group includes samples flagged as admixed by one technique but not the other. Samples in the "Ambiguously admixed" group were included in the reference panel if fewer than 300 were available for a population; otherwise, they were excluded. Highlighted populations are represented by distinct colors: light blue for Scandinavian, light green for Southeast Asian, and brown for Turkish, Iraqi, Iranian and Caucasian populations. These highlighted populations exhibit significant discrepancies in admixture classification with over 25% of their samples categorized as "Ambiguously admixed."

In conclusion, we decided to utilize UMAP after careful consideration of its strengths and limitations. While we recognize the ongoing debate within the scientific community about the best practices for using dimensionality reduction techniques in representing ancestry, we believe that integrating UMAP with other established methods and rigorously validating our approach through various parametrizations has allowed us to conduct a nuanced analysis of the intricate patterns of human ancestry. Moreover, in response to the feedback you have given us, we have expanded our supplementary methods in the revised manuscript. This expansion is intended to provide readers with a clearer insight into our strategic objectives and the reasoning behind our methodological choices.

b. *Nice strategy to leverage global ancestry in the smoothing to refine LAI predictions; this is similar to what is done in the EM iteration in RFmix.*

Thank you for your comment. Indeed, similar to RFmix, our approach utilizes global ancestry information to refine LAI predictions. We believe that such strategies are crucial for improving the precision of ancestry deconvolution at the local level.

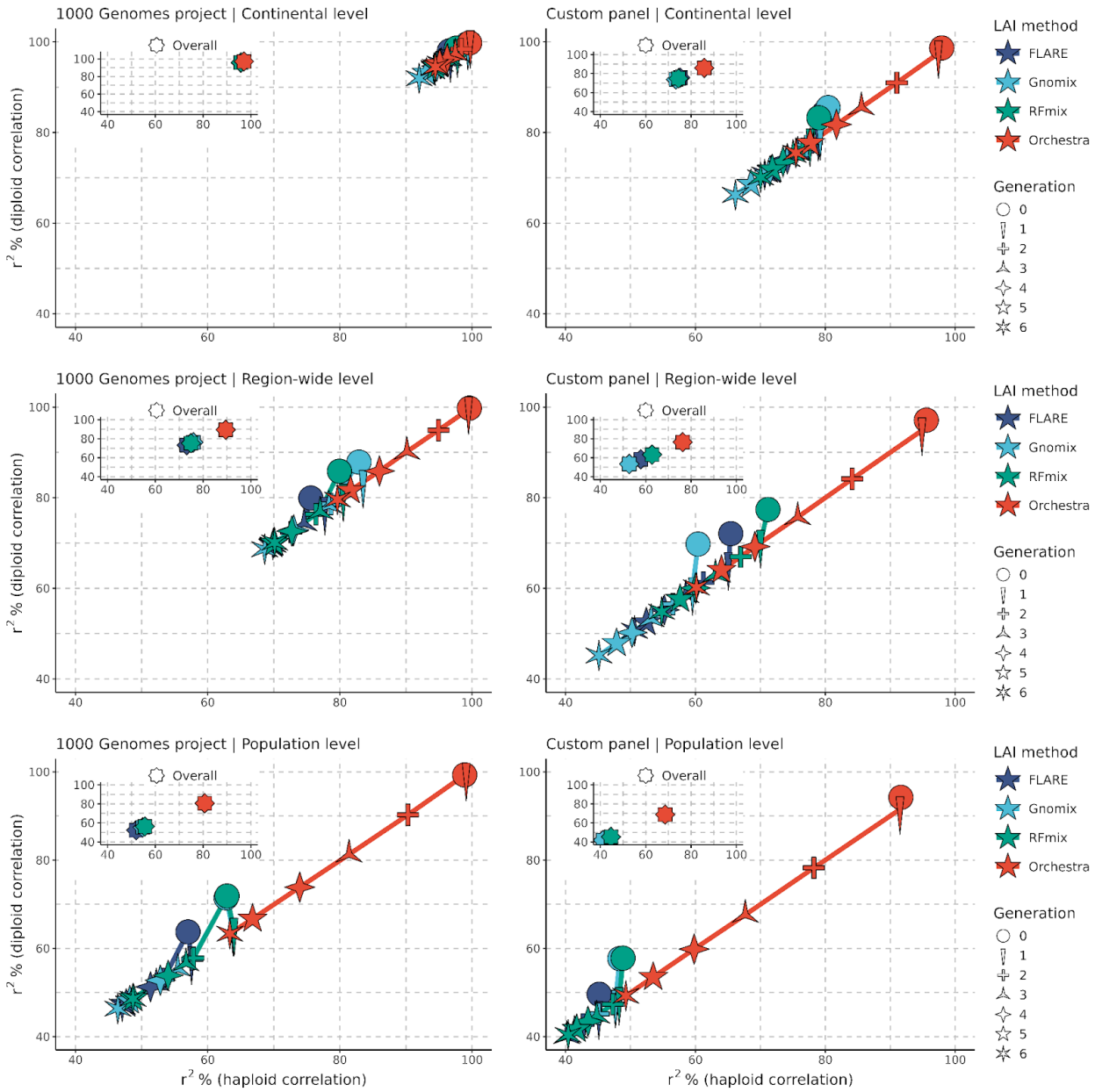
c. *Figure 1 Panel b has a ton of shapes going on - some other way to show this might be clearer…*

We explored several alternatives for presenting the data in Figure 1 Panel b, including numerical annotations and dividing the plot into separate subplots for each accuracy metric. However, these approaches did not enhance the clarity or informativeness compared to the current star-shaped markers, which we found to effectively demonstrate the progression across generations in an aesthetically pleasing manner.

While we believe that the star shapes provide a clear and coherent representation, we are open to further improvements and would welcome any specific suggestions you or the editorial team might have to enhance the figure's clarity. We are committed to ensuring that our figures meet the highest standards of clarity and readability.

d. *Panels are blocking points in Extended Data Fig. 3.*

This issue has been addressed. The updated figure is presented below:



e. Some abbreviations for geography are non intuitive and/or conflict with field standard ones. For example CSA is typically Central and South Asian (per HGDP) but here is used to indicate Central, South, and Southeast Africa.

Thank you for bringing this up. We have updated the following labels in text, tables in figures. We hope they are more intuitive and to our knowledge they don't conflict with standard labels:

- Central, South & Southeast African: CSA -> SAF
- Northeast African: NEA -> EAF
- Nigerian: NGA -> NGE
- North Indian & Pakistani (NIP) -> Pakistani & North Indian (PNI)

- South Indian & Sri Lankan (ISL) -> Sri Lankan & South Indian: (SSI)
- Bengali & East Indian: BNI -> BEI
- Greek & Balkan: GBA -> GBK
- Turkish, Iraqi, Iranian & Caucasian (ICM) -> Iraqi, Iranian, Caucasian & Turkish (ICT)
- Chinese Dai (CHD) -> Dai Chinese (DCH)
- Han Chinese: CHI -> HCH

f. *Methods of the algorithm are written very clearly.*

Thank you for your positive feedback on the clarity of our algorithm's methodology section. We aimed to ensure that the processes and rationale were transparent and accessible to our readers.

g. *Please add page and line numbers.*

We apologize for any inconvenience caused by the absence of page and line numbers in the submitted manuscript. It seems there may have been an inconsistency during the editorial process, as Reviewer 1 received a version with these inclusions. We will try to ensure that the next version of our article includes both page and line numbers throughout to facilitate easier referencing and reviewing. Thank you for bringing this to our attention.

Reviewer #2 (Remarks on code availability):

I believe a more open data sharing strategy is warranted. For the empirical data, the aggregated reference dataset should be somehow made available rather than simply directing the user to the component studies. Aggregating this resource was a lot of work, as the authors indeed highlight in their main text, so is an unreasonable ask to task each subsequent user to replicate the procedure. I appreciate that some of the component studies have access restrictions/applications, but perhaps a streamlined data application for reuse could be implemented such that the aggregated data file could be directly shared.

A significant portion of our data originates from dbGaP, requiring individual access requests for each component study. Additionally, our panel includes a considerable number of individuals from the UK Biobank, which not only requires a separate application process but also enforces more stringent access controls compared to other datasets. These complexities present substantial barriers to creating a unified data application.

To mitigate these challenges, we provide pre-trained models based on the aggregated dataset. These models are readily accessible and allow researchers to benefit from our comprehensive analysis without the cumbersome process of individual data access requests. Providing these models serves as a practical alternative for the scientific community to utilize our findings effectively.

On the software front, Orchestra itself should be distributed in a more scalable way than by emailing the corresponding author.

We agree that a more scalable and accessible distribution method is crucial for broader usability. In response, we have made Orchestra available on our GitHub repository (https://github.com/omicsedge/orchestra-paper). Additionally, we have uploaded various pre-trained models to AWS that can be directly applied to new datasets, facilitating immediate use:

- https://orchestra-paper-public.s3.us-east-1.amazonaws.com/1kg-16pops.tar.gz
- https://orchestra-paper-public.s3.us-east-1.amazonaws.com/custom-35pops.tar.gz
- https://orchestra-paper-public.s3.us-east-1.amazonaws.com/custom-35pops-latino.tar.gz
- https://orchestra-paper-public.s3.us-east-1.amazonaws.com/cosmopolitan-mexicans.tar.gz
- https://orchestra-paper-public.s3.us-east-1.amazonaws.com/fulani.tar.gz
- https://orchestra-paper-public.s3.us-east-1.amazonaws.com/makrani.tar.gz
- https://orchestra-paper-public.s3.us-east-1.amazonaws.com/peruvians.tar.gz
- https://orchestra-paper-public.s3.us-east-1.amazonaws.com/sahelian-arabs.tar.gz

Furthermore, we have provided a toy example on CodeOcean, which can be run interactively. This example serves to demonstrate the software's capabilities and offers users a hands-on approach to understanding how Orchestra operates in a controlled environment.

We have also uploaded scripts and data for a more complex toy example to the AWS link, where you can explore selection signals for adaptive admixture in Admixed Mexicans by estimating LAD and Fadm scores:

- https://orchestra-paper-public.s3.us-east-1.amazonaws.com/canonical_results.tar.gz
- https://orchestra-paper-public.s3.us-east-1.amazonaws.com/prepared_ref_and_target_panels.tar.gz
- https://orchestra-paper-public.s3.us-east-1.amazonaws.com/toy_example_scripts.tar.gz
- https://orchestra-paper-public.s3.us-east-1.amazonaws.com/observed_and_expected_frequencies_for_Fadm.tar.gz

We believe these steps will greatly enhance the accessibility and usability of Orchestra.

RESPONSE TO REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

This paper presents Orchestra, a machine learning-based method for Local ancestry inference (LAI). The results shown here represent a substantial improvement in LAI accuracy, which not really been seen in the last decade. The authors have comprehensively replied to the reviewers comments. They have made their code and method available, edited the text for clarity, and reworked some analyses. They show that their method is a clear improvement in local ancestry inference, improving accuracy and demonstrate that fine-scale analyses are possible

The authors have posted code on GitHub and Code Ocean, posted pre-trained models on AWS, and included two ready-made example analyses. This represents a substantial improvement in the availability and accessibility to the science and code presented in the manuscript.

I appreciate the author's clarification of their approach to train-test data splitting. It is my understanding that they have segmented the training and test data in a reasonable way. I remain slightly skeptical that some of the headline results will fully generalize because the increased flexibility of the model in Orchestra (vs other methods) may be better able to capture (nearly unavoidable) data artifacts such as batch effects, compared to previous methods. However, it seems the authors have taken reasonable steps to try to address this.

I appreciate the added details on tract length distributions in the Alternative figure 2, and suggest that this is the better figure. But I find that the figures could be more clearly labeled and laid out.

Thank you for your notes. We have kept the alternative figure 2. We have also made an effort to make the figure labels and legends more clear and detailed.

Line 19 - "SNPs, which account for only 0.1% of our DNA" The authors have revised this sentence in this version, but I still find it unclear. Do the authors mean that approx. 1 in 1000 basepairs of DNA contains a SNP? This very much depends on the set of samples you look at and the frequency of SNP sites that are considered. There are very many SNPs with rare alleles. I still think the authors should consider what idea they want to convey here. Perhaps they are making a point about pairwise differences between individuals at SNP variant sites, but I'm not sure.

We tried to convey that although SNPs represent a small portion of the genome, we can use that small portion to reveal a plethora of information about humans. Just something we thought was fascinating. However, we recognize that this is not very important to the overall narrative so we have decided to remove the sentence. We feel the introduction reads more straightforward and to the point now.

Reviewer #1 (Remarks on code availability):

The authors have made the analysis code publicly available, and have provided a small example analysis .

There are details installation instructions and code to facilitate installation and setting up the software environment and data.

Reviewer #2 (Remarks to the Author):

I am broadly satisfied with the response from the authors. They have provided a thorough and detailed comparison of tract lengths with historical records, which significantly enhances the credibility of their findings. Additionally, the revised selection scanning they employed offers a more robust validation of their results, further reinforcing the reliability of their outcomes. I also appreciate the improvements made to the language in the introduction and discussion sections. The authors have addressed my critiques to that effect effectively, particularly in their discussion of existing competing LAI tools and with respect to the computational requirements for Orchestra. These sections are now clearer and more informative, providing a better context for understanding the advantages and limitations of their method. Moreover, the authors have taken significant steps to enhance the accessibility of their work by openly distributing their code and pre-trained models based on the curated reference panels. This is vital, facilitates reproducibility and will also encourage broader use and validation of their method within the scientific community.

My only remaining gripe is their inclination to defer the consideration of accuracy estimates to a future publication. Accuracy is highly important for LAI, so it would be fantastic to incorporate those analyses into this work if that effort is ongoing. At the very least, some discussion of the considerations for use with ancestry components not in the reference should be added to the text, since misclassification will occur in that instance and apparently cannot be easily filtered out.

Thank you for highlighting this point. Upon revisiting our answer to the first round of revisions, we realized we had misunderstood your question pertaining to accuracy estimates and have not addressed it adequately. Please allow us to rectify that here.

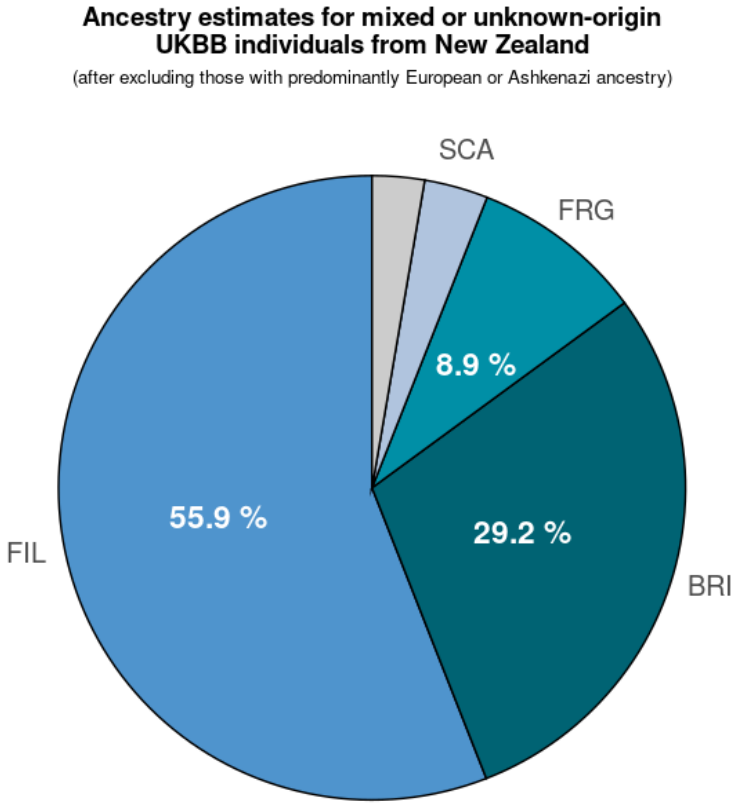
If we understand it correctly now, the main question is: what happens if there is a gap in the reference populations, and there is a target population we do not have? How does a person of that population, or a genomic region of that population in an admixed person, get classified? Do we have a metric that would tag this as a low confidence region?

As of now, gaps in reference data are something that has to be considered in the study design and also when interpreting the data. This is why we tried to design our custom reference panel to be comprehensive, and designed our 35 reference populations to cover as much on the world map as feasible – in order to have any person or DNA segment classified to a category as accurately as possible. Due to unavailability of data from certain regions, gaps do exist in our data. We argue that, in such cases, such a person or a DNA region will get assigned to the phylogenetically closest group. Our confidence metric, which is a function of the deep learning model, will still assign these calls with a relatively high confidence.

Let’s look at an example. An unfortunate gap in our data is the polynesian population – we have not been able to acquire any Polynesian samples from any of the databases available to us. What would happen with a person from New Zealand who has Māori (Polynesian) ancestry? We would assume that any Māori fragments would be classified as the phylogenetically nearest branch among our 35 reference populations, which in this case would be Filipino (FIL) – as the Polynesians originated from this region (Taiwan, Philippines) about 3000 years ago and native Taiwanese and Negrito tribes from the Philippines are phylogenetically closest to Polynesians

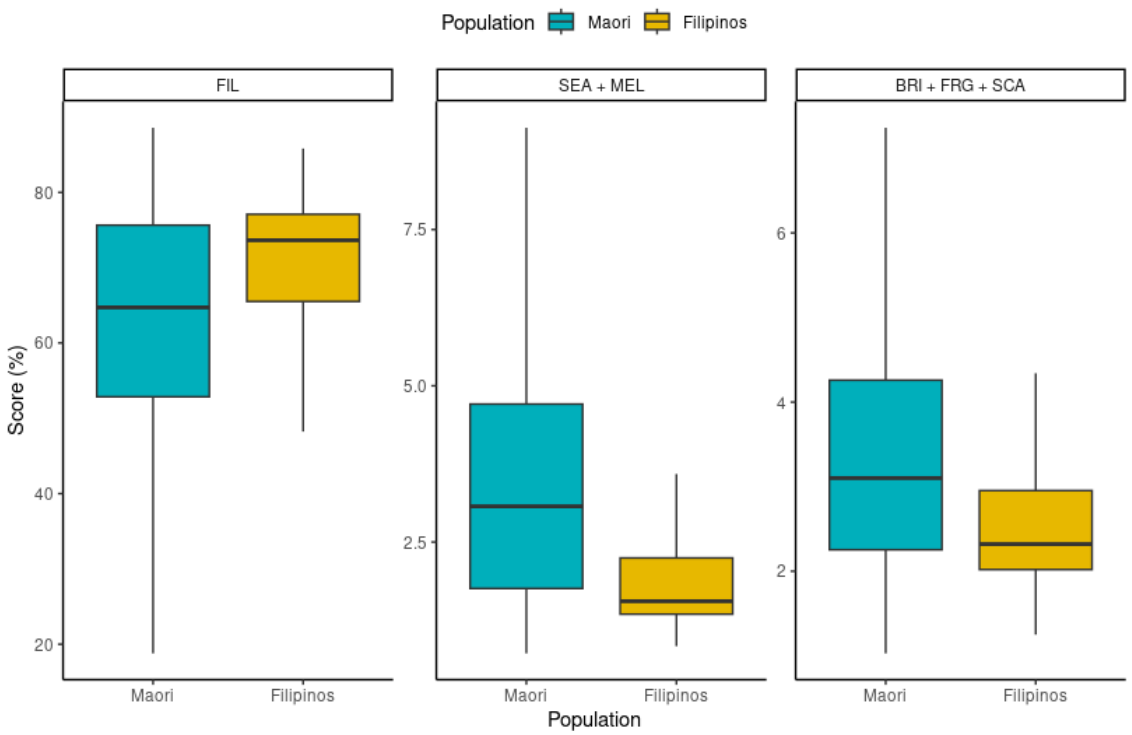
(https://pmc.ncbi.nlm.nih.gov/articles/PMC5515717/; https://www.nature.com/articles/srep29890).

We can test this because in the UK Biobank, we have several participants born in New Zealand (Supplementary Fig. 7). From this dataset, we removed individuals self-reporting as "British", "Irish", "White", "Any other white background" or "Chinese", and any other samples of predominantly European ancestry (≥80% BRI + FRG + SCA + SPP + GBA + ITA + FIN + EAE + ASK). After filtering, the final dataset comprised 10 individuals, including 3 identified as "Any other mixed background" and 7 with no reported origin. We used this subset to explore ancestry patterns potentially linked to the Māori indigenous population. The pie chart below shows a major portion classified as Filipino (FIL), as expected.



Distribution of ancestry estimated for mixed and unknown origin UKBB individuals from New Zealand. Significant portion of windows is classified as Filipino (FIL).

To see what happens with the accuracy metric, we compared accuracy scores for windows classified as FIL, between these 10 individuals from New Zealand – of likely Māori descent –, with 10 Filipinos of ≥80% FIL ancestry also found in the UK Biobank. Scores for Southeast Asian and Melanesian (SEA + MEL) and British or North West European (BRI + FRG + SCA) ancestry probabilities in these windows were also assessed.



Distribution of Orchestra's quality scores for windows that classify as FIL in individuals of Maori (N=10) and Filipino (N=10) descent in UKBB. The data indicates that FIL scores in the Maori are relatively lower, but still high compared to those in Filipinos.

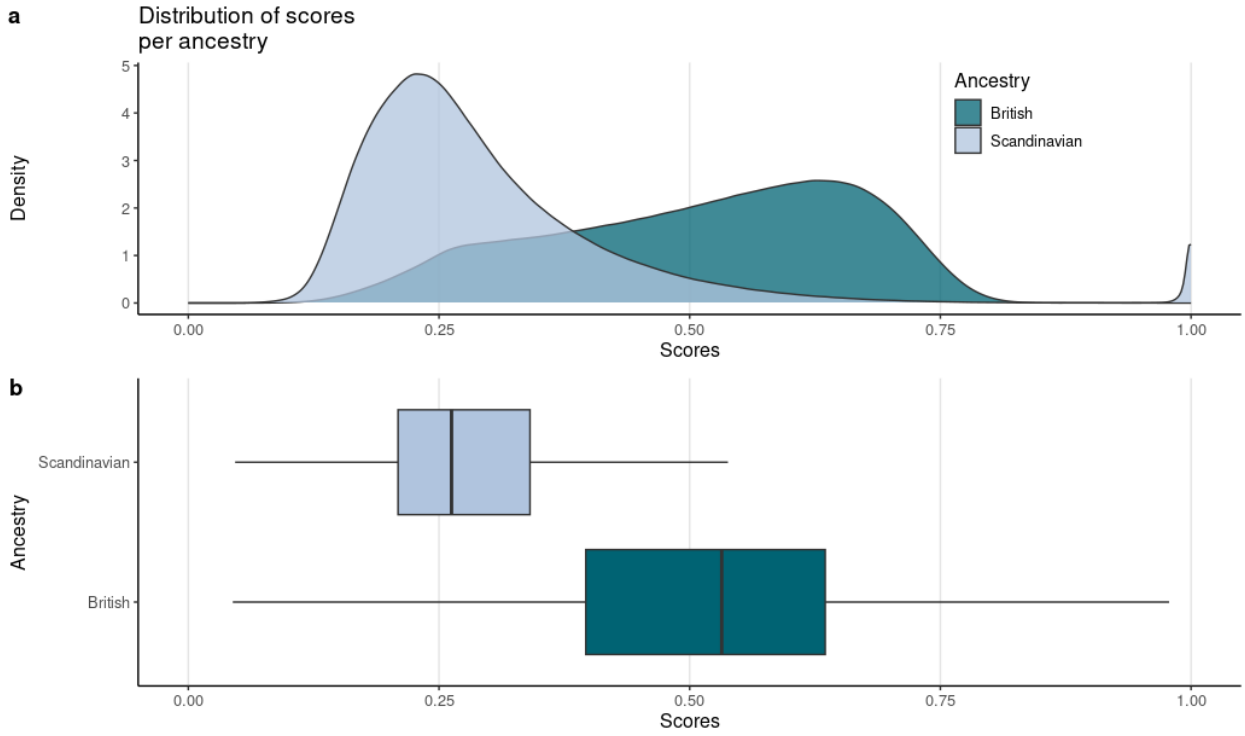
As seen above, FIL ancestry scores for Māori individuals were slightly lower than those of Filipinos but remained relatively high in both groups. For those same windows, those of potential Māori descent showed slightly higher BRI probabilities, likely due to admixture or adjacency to BRI-called windows. SEA and MEL probabilities were also slightly elevated in Māori, potentially reflecting the contribution of these ancestries to the polynesian ancestry. This supports our expectation that a missing population would classify as the phylogenetically closest population in the reference panel, in the case of Māori that would be FIL.

Therefore the accuracy metric we use would not be able to alert to a missing population in the reference panel, especially when there are phylogenetically close populations. If we were to try and design such a metric, we are not sure how we would go about it. That would require some serious thought and lots of experimentation.

We use the fact that missing populations classify to phylogenetically close populations as a feature, in several instances in our current study. One is ancestral mapping, where we remove a population out of the reference panel on purpose, to see which populations our algorithm would recognise as the next closest.

Another example where we use this assumption is when we extract Native American fragments from the Latin American Native populations in the 1K Genome Project. There we assume that all native fragments would classify as East Asian, when we offer only European, East Asian and African as the 3 reference populations.

The confidence metric shows how uncertain some calls are, based on a number of factors, including the size and quality of each reference population, or how close that population is genetically to another one. That is what we tried to explain with our example of the BRI (large, high quality reference population) and SCA (lower quality small reference population) chromosome 10 example we provided for the first round of reviews (Figs below).



Distribution of Orchestra's quality scores for British and Scandinavian ancestry calls within the UKBB British samples. The data indicates that British ancestry calls are generally more reliable, with higher quality scores, whereas Scandinavian-inferred windows tend to have lower scores, reflecting greater uncertainty.

The calibration of this metric for ancestry assignment is still expected to take quite an involved effort – as we would need to find a way to meaningfully convert a deep learning metric to real ancestral probabilities. We are unfortunately unable to perform that at this timepoint. We have updated the discussion and the supplementary methods to highlight all these points and better inform the users about these considerations when using this metric.

As for the concerns pertaining to misclassifications that cannot easily be filtered out – we stress that knowing and understanding one's reference panel is key to interpreting the data, and we don't see a way around that at the moment.

Reviewer #2 (Remarks on code availability):

I have perused the github page for the release of this code and am satisfied with the availability of code and the usability of the documentation.