nature communications

Article

https://doi.org/10.1038/s41467-025-59936-3

Tracing human genetic histories and natural selection with precise local ancestry inference

Received: 14 December 2023

Accepted: 6 May 2025

Published online: 16 May 2025

Check for updates

Jon Lerga-Jaso[1,2], Biljana Novković[1,2], Deepu Unnikrishnan[1], Varuna Bamunusinghe[1], Marcelinus R. Hatorangan[1], Charlie Manson[1], Haley Pedersen[1], Alex Osama[1], Andrew Terpolovsky[1], Sandra Bohn[1], Adriano De Marino[1], Abdallah A. Mahmoud[1], Karatuğ O. Bircan[1], Umar Khan[1], Manfred G. Grabherr[1] & Puya G. Yazdi[1] ✉

Local ancestry inference is crucial for unraveling demographic histories, discovering selection signals, and including admixed individuals in genomic studies for improved equity and portability. To date, the precision and resolution of local ancestry inference were limited by technical and dataset issues. To address them, we present Orchestra, a model we train on over 10,000 single-origin individuals from 35 worldwide populations that demonstrates superior accuracy in benchmarking analyzes. We employ Orchestra to shed light on the demographic history of Latin Americans, finding trace ancestries supported by historical records. We then deploy it to offer insight on the debated Ashkenazi Jewish origins, highlighting their South European heritage. Finally, Orchestra enables us to map selection signatures, identifying trace Scandinavian ancestry in British samples and unveiling an immune-rich region linked to respiratory infections passed down from the Viking conquests. Our work significantly advances the field of local ancestry inference, highlighting its use in admixed populations.

For decades, population geneticists have used DNA to infer the history of humankind[1–4]. Mating between individuals in geographical proximity, coupled with genetic drift and divergent demographic histories, helped shape our modern human genetic landscape, allowing us to reference any single individual to pre-defined reference populations[5,6]. However, as the movement of people across the globe has intensified in recent centuries, humans have become more admixed, and an increasing number of people cannot be traced back to a single reference population[7]. While global ancestry inference (GAI) allows us to infer an individual's overall admixture proportions, it fails to provide information about the fine-scale patterns across the genome. Despite similar global admixture proportions, two individuals may have very different ancestry compositions at any location within the genome[8]. That is why in admixed populations, local ancestry inference (LAI) becomes indispensable for various downstream applications.

Several LAI methods that infer the ancestry of different segments on each chromosome have been developed over the years[9–11]. RFmix has been the go-to method of many previous studies because of its reliability, which has withstood the test of time[9]. FLARE has been more recently designed for speed and efficiency, to tackle ever-increasing amounts of data[10]. Similarly, Gnomix has been recently developed to handle whole-genome data with the ability to detect deep historical admixture[11]. Various LAI methods have been successfully applied to admixed populations to boost power and resolution in GWAS[12], improve colocalization of GWAS and expression quantitative trait loci (eQTL)[13], and detect gene-gene and gene-environment interactions[14]. In addition, LAI models have been leveraged to improve polygenic risk

[1]Research & Development, Omicsedge, Miami, FL, USA. [2]These authors contributed equally: Jon Lerga-Jaso, Biljana Novković.
✉e-mail: pyazdi@omicsedge.com

scores (PRS) specifically for admixed individuals[15,16]. However, to date, LAI methods and all downstream efforts have mainly applied cross-continental resolution, looking at anywhere from two- to six-way admixture (e.g., European vs. African or East Asian).

We know that populations that are geographically close can exhibit significant genetic heterogeneity due to historical isolation and/or geographical or cultural barriers. Such regional variation can impact the genetic architecture of complex phenotypes. In Africa alone, the genomic diversity is vast, showing extreme allele frequency divergence in many medically relevant variants[17]. Similarly, genomic variability is extensive among Asians, who comprise nearly 60% of the total world population, with unequal genetic disorder burden and pharmacological susceptibility[18]. Subtle genetic clines can be observed even for Europeans[19,20]. Therefore, various genomic disciplines may have a lot to gain from broadening the scope of LAI to include within-continent diversity. Finding that currently available LAI methods struggle with closely-related reference populations, and especially when the number of reference populations is large, our aim was to increase the granularity of LAI and achieve a resolution that was previously possible only with GAI.

Here, we present Orchestra, a LAI method of high accuracy and resolution, and apply it to retrace the genetic history of Latin Americans, as a prime example of admixture. We next explore the relationship between 35 worldwide populations and show that Orchestra can be used to estimate genetic closeness between populations and shed light on their demographic history. Finally, we use Orchestra to detect natural selection signatures. We demonstrate that a more granular LAI method, such as Orchestra, can help delve into admixture between more closely-related populations and can shed light on evolutionary processes that we were not able to track with such resolution beforehand.

Results

Local ancestry deconvolution

Orchestra (Optimal [Re]Combination of Haplotypes to Establish Segmentation of a Target from Reference Ancestries), is a LAI algorithm that consists of a two-stage pipeline: a base layer and a smoothing module (Fig. 1a). The base layer classifies genomic windows of predetermined size by generating a distance measure between the target genome and each of the reference populations. This measure, we refer to as recombination distance, is the minimum number of segments needed to reconstruct a target sequence from the sequences present in each reference population. It approximates the number of crossover events needed to reconstruct a given sequence. The base layer uses a greedy

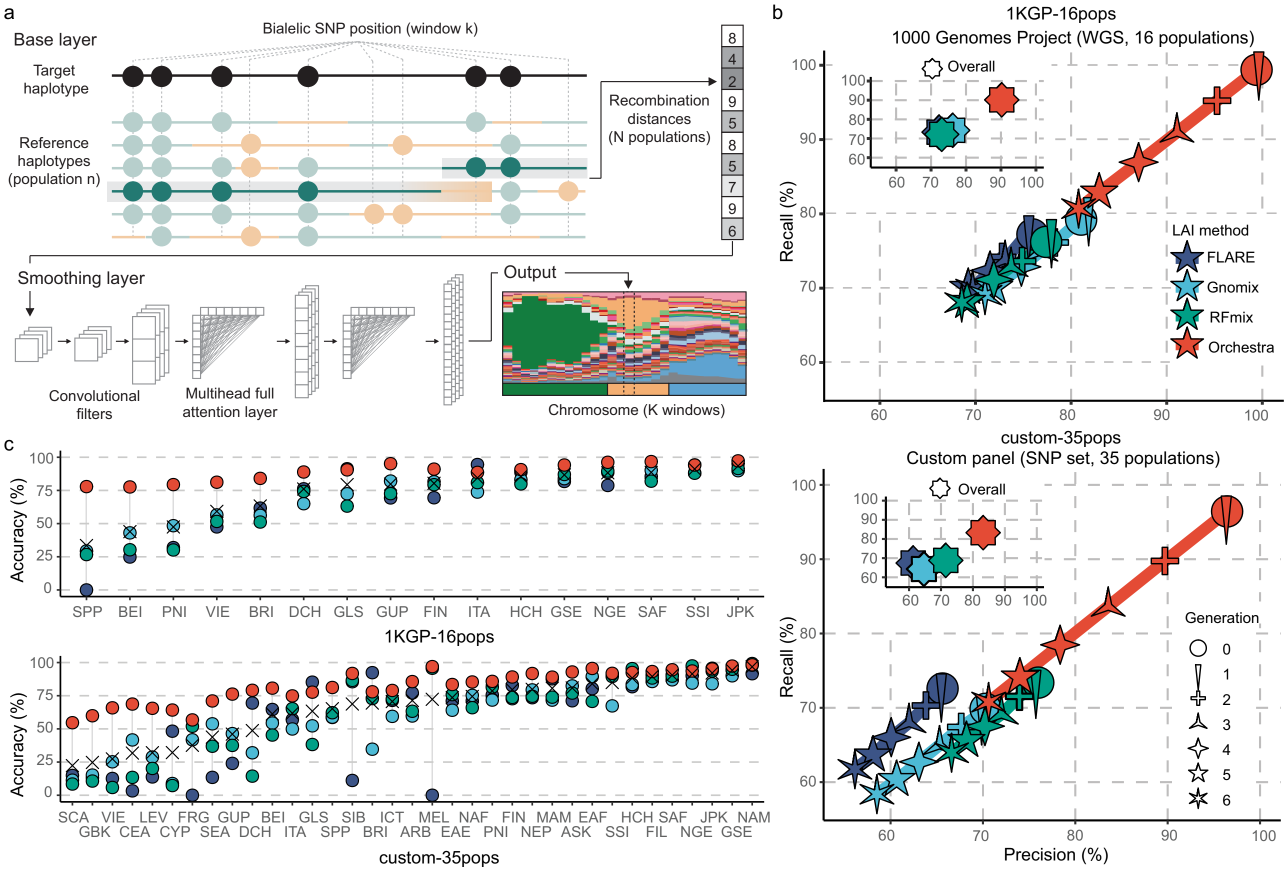

Fig. 1 | Local ancestry inference of Orchestra vs. other LAI methods. a Orchestra schematics. The input consists of a target genome of unknown ancestry and a reference set of single-origin individuals grouped into reference populations. For each genomic window, the base layer generates the recombination distance as the minimum number of segments needed to reconstruct the target sequence from the sequences of each reference population n. Filled circles represent alternative alleles. Green circles indicate a match between the reference genomes and the target, while orange circles indicate a mismatch. The smoothing layer processes these distance measures with a series of convolutional and attention layers using information from surrounding windows and broader regions of the genome, yielding the final probabilities for the N reference ancestries. This procedure is repeated for all windows in the genome. **b** Recall and precision of Orchestra (red), RFmix (green), Gnomix (light blue) and FLARE (navy) across 6 generations of simulated admixture in the 1KGP dataset (1KGP-16pops; N = 3276) and the larger custom dataset (custom-35pops; N = 24,408). The number of points for each star shape corresponds to the number of generations. **c** Accuracy (%) per population in the 1KGP-16pops and the custom-35pops datasets for Orchestra (red), RFmix (green), Gnomix (light blue) and FLARE (navy). Populations are ordered by mean accuracy across all methods (denoted by a cross). Accuracy is shown as synonymous with recall. See Fig. 3 legend and Supplementary Fig. 1 for a full list of populations with abbreviations.

approach in which a similarity matrix is calculated by an element-to-element comparison per position and per sample, to obtain a vector of recombination distances across all reference populations. The smoothing module is a deep learning model with convolutional and attention-based elements. The convolutional element processes the base layer insights generated for each window using the information from surrounding windows. The attention-based component provides a weak link to global ancestry. This is reflective of real world genomes, since the presence of a certain ancestry in one place of the genome increases the likelihood of finding that same ancestry in other genomic regions. Combining the recombination distance base layer with a deep learning smoothing module synergistically leads to a state-of-the-art technique for accurate ancestry deconvolution.

The accuracy of any ancestry model greatly depends on the quality of the reference panel. We assembled a set of reference populations by merging data from more than 30 published studies, combining both whole genome sequencing and array-based genotyping (Supplementary Table 1). A significant fraction of the total samples comes from non-UK ancestries captured by the UK Biobank (UKBB). With much shorter migratory distances just a few decades ago, we found that tracing ancestral origins by birth-place and self-reported ethnicity of UKBB participants was a sufficiently reliable initial proxy for ancestry. All retrieved samples underwent a series of quality filtering steps. We kept a composite set of directly genotyped variants obtained by combining all SNPs from array-based studies and filtered by a minor allele frequency (MAF) ≥ 5% to minimize imputation-related biases (see "Methods"). Next we conducted two GWASs to check if each SNP was associated with a genotyping platform or ancestry, and filtered out those that ranked in the top high and low end, respectively, to minimize batch effects and retain meaningful ancestry-informative differences. We then used two separate dimensionality reduction techniques to characterize relationships between samples and remove any samples that showed a disagreement between reported ancestry and inferred genetic origin: 1) Principal component analysis (PCA) followed by uniform manifold approximation and projection (UMAP)[21] and 2) t-distributed stochastic neighbor embedding (t- SNE)[22] used on genealogical nearest neighbor (GNN) statistics estimated with tsinfer[5]. This resulted in a high-quality reference panel of 10,169 non-admixed individuals from 35 world regions, which we used as our reference populations (Supplementary Fig. 1) for three-letter population abbreviations; see "Methods" for more details).

We benchmarked Orchestra against other leading LAI algorithms, including RFmix[9], FLARE[10] and Gnomix[11] and, using two reference panels: 1) 1KGP-16pops, a high-coverage WGS set of non-admixed and unrelated samples collected by the 1000 Genomes Project (1KGP) with 16 populations and 2) custom-35pops, our larger, more diverse curated panel with 35 populations. Both panels were split into test and training sets (20% and 80% of samples) and used to simulate 6 generations of random admixture using SLiM[23]. Precision and recall were reported as performance estimates on all chromosomes per generation and per population.

Orchestra substantially outperformed other LAI methods (Fig. 1b). When using the 1KGP-16pops reference panel, Orchestra's average recall and precision across generations was 90.17% and 90.22%, respectively; an improvement of +15.89% and +14.03% compared to the second best model, Gnomix. For the custom-35pops panel, the average recall and precision was 79.54% and 80.54%, respectively, an improvement of +15.04% and +13.99% compared to the next best model, RFmix. Orchestra was the most accurate across 6 generations of admixture. As expected, the accuracy decreased with an increasing number of generations. However, Orchestra's performance in the most admixed samples equaled or exceeded the best performance in the non-admixed generations by other LAI methods.

Orchestra retained high accuracy regardless of the reference population, with an ability to distinguish between closely related ancestries. Orchestra achieved accuracy greater than 75% for all populations within the 1KGP-16pops panel (Fig. 1c). For the custom-35pops panel, Orchestra achieved an accuracy of over 50% for all populations, and over 75% for 26 out of 35 populations. The other three LAI models struggled with a third of the populations, with accuracy below 50% (Fig. 1c). Orchestra's accuracy was superior at both region-wide and continental levels, the recall exceeding 93.43 and 98.90% for 1KGP-16pops and 87.73% and 94.03% for custom-35pops (Supplementary Fig. 2a, b). Substantial improvement in accuracy was also observed when using the r^2 metric, where Orchestra achieved an improvement of +24% for the custom-35pops and 23.9% for the 1KGP-16pops at the population level (Supplementary Fig. 3).

In addition to our two panels, we applied all LAI models to an independent set of over 10,000 UK biobank samples that were not included in the custom-35pops panel (Supplementary Fig. 2c). Orchestra outperformed the other LAI methods for over 90% of the 103 evaluated countries.

Retracing genetic histories

Latin Americans are a prime example of admixture, as their DNA can be traced to three broad sources, European, Sub-Saharan African and Native American. However, there is a wide fluctuation in the proportion of these ancestries throughout the continent. In addition, the genetic makeup of Latin Americans shows regional heterogeneity. For example, Colombians are more likely to have Senegalese, Gambian or Guinean African ancestry, while Brazilians are more likely to have ancestry from Angola and Congo. Similarly, while a lot of Latin Americans get their European ancestry from Spain or Portugal, many Argentinians also have Italian roots[24].

To assess the accuracy of our LAI model in these populations, we simulated Latin American individuals from Southern (SPP and ITA) and Northern (FRG and BRI) Europeans, Western (GSE, GLS and NGE) and Central and Southern (SAF) Africans and artificially-reconstructed Native Americans (NAM). Simulations were performed by emulating genetic intermixing for 12 generations using SLiM[23]. Native American genomes were created in silico using the Latin American samples from the 1KGP, keeping only the genomic segments identified as East Asian as a proxy for indigenous ancestry. Simulations were adapted to the genetic makeup that can be found today in three broad regions within Latin America: the Antilles, comprised of 55% European, 40% African (specifically NGE) and 5% Native American ancestry (NAM); Mexico and Central America, made up of 50% European, 10% African (GLS and GSE) and 40% Native American ancestry (NAM); and South America, composed of 65%, 15% and 20% of European, African (GLS, GSE and SAF) and Native American (NAM) ancestry, respectively[24]. For benchmarking purposes, we compared our results against FLARE, Gnomix and RFmix (Fig. 2a). Orchestra achieved an overall precision and recall of 77.17% and 76.73%, respectively, outperforming the other LAI models in all three aforementioned regions.

This gave us confidence to apply our model to real-life Latin American samples from the 1KGP and UKBB datasets (Fig. 2b), where Orchestra was able to successfully retrace major patterns in the genetic history of the Latin Americas. For example, the highest percentage of Native American ancestry (NAM) was found in Bolivia, Peru, Ecuador and Mexico, matching demographic and genetic reports from this region[24–26]. When we looked at the lengths of chromosome tracts for various ancestries (Supplementary Fig. 4), we found longer Native American ancestry fragments in Peru, consistent with the higher proportion of Native American ancestry in this population, which may suggest recent or ongoing gene flow from indigenous groups. The majority of African ancestry in the Caribbeans was assigned as Nigerian (NGE) and next Ghanaian, Ivorian, Liberian & Sierra Leonean (GLS). In contrast, a larger portion of the African ancestry in Brazil was assigned as Central, South & Southeast African (SAF), which captures populations of Bantu origin on the

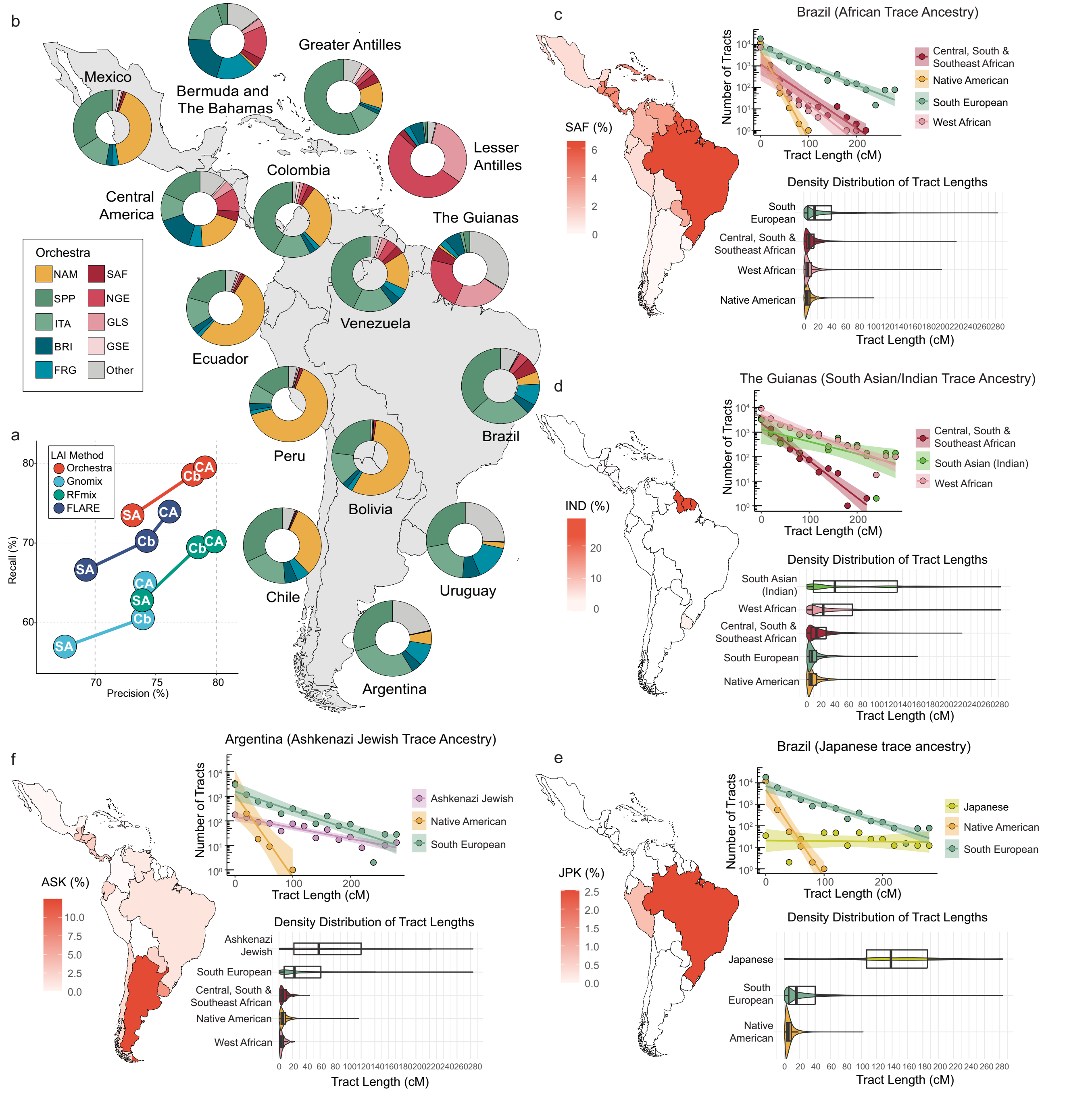

Fig. 2 | Ancestry inference in Latin Americans. Clockwise: **a** Percent recall and precision for ancestry deconvolution by FLARE (navy), Gnomix (light blue), RFmix (green), and Orchestra (red) in Latin Americans simulations equivalent to 12 generations of admixture (N = 4068); simulations were adjusted 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 Latin America according to Orchestra (N = 16,224). Proportions of Native American (NAM: yellow), Southern European (SPP, ITA: green), Northern European (BRI, FRG: blue) and African (SAF, NGE, GLS, GSE: red) ancestries are shown. NAM: Native American; SPP: Spanish & Portuguese; ITA: Italian; BRI: British & Irish; FRG: French & German; SAF: Central, South & Southeast African; NGE: Nigerian; GLS: Ghanaian, Ivorian, Liberian & Sierra Leonean; GSE: Gambian & Senegalese. **c**–**f** Orchestra was able to detect trace ancestries that reflect known historical population displacements and immigration events. **c** Central, South & Southeast African (SAF) ancestry in Brazil (N = 234), **d** Indian (IND) ancestries in the Guianas (N = 311), **e** Japanese & Korean (JPK) ancestry in Brazil (N = 240), and **f** Ashkenazi Jewish (ASK) ancestry in Argentina (N = 69). Graphs show the distribution of local ancestry tract lengths, as a function of tract length (in 20 cM bins). Error bands show 95% CI. Boxplots over violin plots show tract density distribution. The central line of the boxplot denotes the median, box boundaries represent the first and third quartiles, the whiskers range from minimum to maximum values.

African continent. This is in agreement with historical records of Africans in Brazil originating primarily from Angola, a former Portuguese colony[24]. A similar distribution of tract lengths across countries suggests parallel dynamics of African admixture, where relatively shorter tracts reflect the timeline of the transatlantic slave trade (1500 s to 1880s). Orchestra captured a higher percentage of Spanish & Portuguese (SPP) ancestry in Mexico, the Greater Antilles, Columbia and Venezuela. British & Irish ancestry (BRI) was more prevalent in Bermuda and the Bahamas, Lesser Antilles and the Guianas. Italian (ITA) ancestry was more prominent in Argentina,

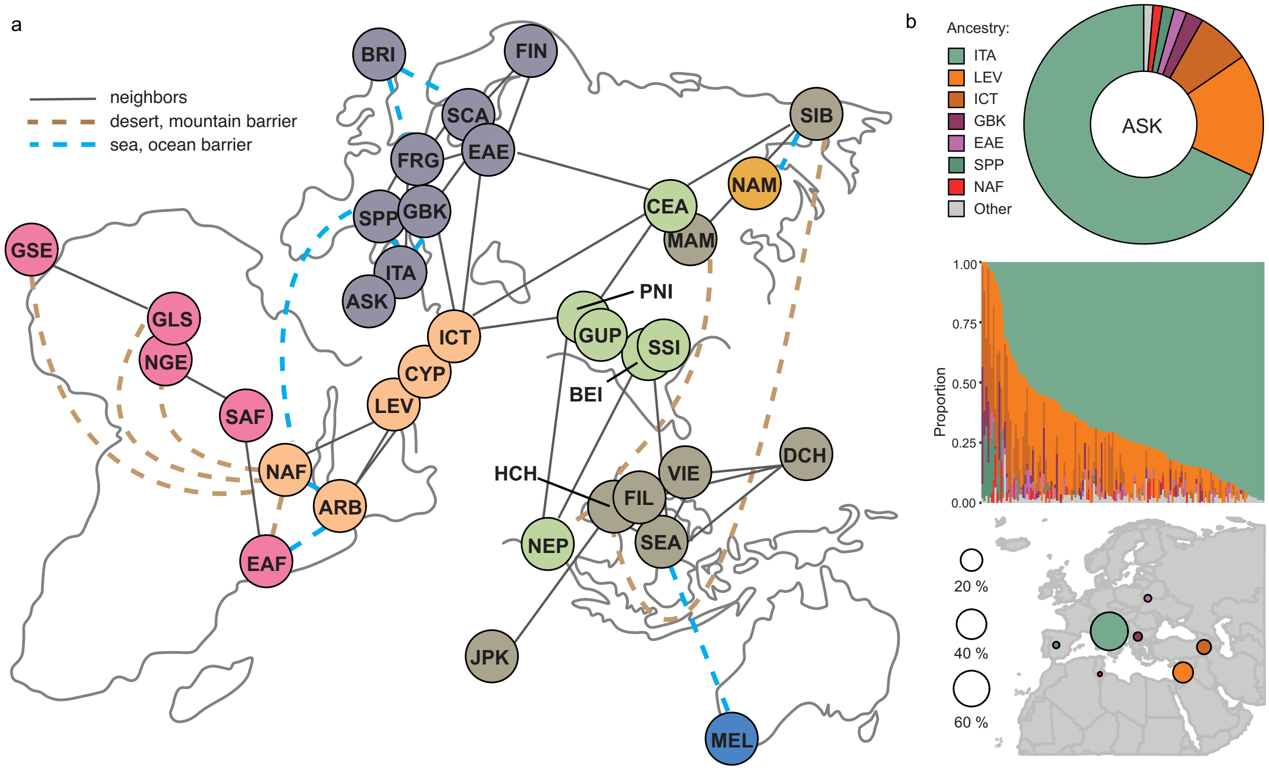

Fig. 3 | Ancestral mapping. a Ancestral map of 35 populations projected onto two-dimensional space using the SMACOF algorithm applied to a matrix of distances derived from the proportions obtained in the analysis. Solid lines connect neighboring populations. Dashed lines indicate populations separated by geographical barriers to dispersion. Continent contours were illustrated by hand. **b** Inferred ancestry of the Ashkenazi Jewish (ASK) population (N = 152) according to Orchestra when ASK was omitted from the custom-35-pops training set. Overall ancestry proportions (top); individual ancestry proportions with each bar representing a single individual (middle), and ancestry proportions shown on a map (bottom). ARB Arab, ASK Ashkenazi Jewish, BEI Bengali & East Indian, BRI British & Irish, CEA Central Asian, CYP Cypriot, DCH Dai Chinese, EAE Eastern European, EAF Northeast African, FIL Filipino, FIN Finish, FRG French & German, GBK Greek & Balkan, GLS Ghanaian, Ivorian, Liberian & Sierra Leonean, GSE Gambian & Senegalese, GUP Gujarati Patels, HCH Han Chinese, ICT Iraqi, Iranian, Caucasian & Turkish, ITA Italian, JPK Japanese & Korean, LEV Levantine, MAM: Manchurian & Mongolian, MEL Melanesian & Aboriginal Australian, NAF North African, NAM Native American, NEP Nepalese, NGE Nigerian, PNI Pakistani & North Indian, SAF Central, South & Southeast African, SCA Scandinavian, SEA Southeast Asian, SIB Siberian, SPP Spanish & Portuguese, SSI Sri Lankan & South Indian, VIE Vietnamese.

Brazil and Uruguay. These findings match known demographic and historical evidence[27].

Interestingly, Orchestra was able to detect several notable trace ancestries (Fig. 2c–f). In addition to the aforementioned SAF in Brazil, these also include a high percentage of Indian ancestries (IND = PNI, SSI, BEI, GUP) in the Guianas, reflecting the indenture system used in former British colonies[28], a high percentage of Ashkenazi Jewish (ASK) ancestry in Argentina, which hosts the largest Ashkenazi Jewish community in South America, as well as the Japanese ancestry detected in Brazil and Peru (JPK), which experienced a large wave of Japanese immigration over the first half of the 20th century[29]. The chromosome tract lengths for each of these ancestries match our expectations for the respective time periods for each admixture event (Fig. 2c–f, and Supplementary Fig. 4).

In addition to Latin America, we applied this method to all UKBB samples not used in our reference panel and to samples belonging to ethnicities found in various datasets that were not included in the custom-35pops panel (Supplementary Figs. 5–7).

Ancestral mapping

Seeing that we were able to identify 35 different populations with superior accuracy (Fig. 1), next we explored the relationships among these populations. We created 35 distinct reference panels for each target population, where that target population was omitted from its own reference panel. Orchestra was run to obtain admixture proportions, which were then converted into a matrix of distances that were projected onto two-dimensional space using the SMACOF algorithm. The resulting network (Fig. 3a) largely reflects geographical proximity and replicates known relationships between various populations[19,30].

Ancestral mapping results for individual populations are shown in Supplementary Figs. 8–14. For example, when we removed our French & German (FRG) population from the reference panel, the FRG samples were mapped as mostly British & Irish (BRI, 58.9%), Scandinavian (SCA, 13.1%), Italian (ITA, 9.5%) and Eastern European (EAE, 9.4%), with the ITA ancestry more prevalent in the French and the Swiss, while EAE ancestry was more common in Austrians and Germans (Supplementary Fig. 10). Our Iraqi, Iranian, Caucasian & Turkish population (ICT) was reconstructed as mostly Levantine (LEV, 65.2%), Pakistani & North Indian (PNI, 11.9%), Cypriot (CYP, 9.6%), Central Asian (CEA, 6.4%) and Greek & Balkan (GBK, 3.8%). PNI ancestry was more common in the East, in Iraqis and Iranians, while GBK was more present in the West, especially in the Turks (Supplementary Fig. 9). In these and many other populations, we observed a genetic cline, indicating there is genetic heterogeneity within most of our 35 populations.

The ancestry and origin of the Askenazi Jewish have been subject to heated debate over the last two decades. Here we mapped our Ashkenazi Jewish as primarily Italian (ITA, 68%), followed by Levantine (LEV, 16.6%), Iraqi, Iranian, Caucasian & Turkish (ICT, 7.2%), Greek & Balkan (GBK, 2.4%) and Eastern European (EAE, 1.7%) (Fig. 3b). This largely agrees with several reports based on both modern and medieval Ashkenazi Jewish DNA[31–33].

Detecting natural selection signatures

To check if we could leverage our LAI model to detect signatures of natural selection, we first aimed to replicate previously identified signals. We followed the methods described in Cuadros−Espinosa et al. 2022[34], where combined statistics based on admixture proportions (F_{adm} and LAD) was used to scan genomes of admixed populations for selection signals (see Supplementary Methods). Where possible, we retrained Orchestra using the same or approximated reference populations to those that were used in the original study. Out of the seven admixed populations tested, we were able to completely replicate signals in six populations and partially in one population (Supplementary Fig. 15). This suggests that these signals are robust to discovery with different methods, and that Orchestra may be used to recover signals of natural selection at a local level.

We then proceeded to apply Orchestra to British samples (N = 415,859) in the UK biobank dataset. Figure 4a shows the distribution of Scandinavian (SCA) ancestry in this population. SCA ancestry was particularly enriched in the East of England and East Midlands, where we also found the highest density of former tentative Viking settlements, inferred as settlement names ending in -by, -thorpe or -toft, confirming previous reports of Scandinavian hotspots in Eastern England[35,36]. Next we aimed to identify potential adaptive signals using the F_{adm} and LAD framework. We found a significant enrichment of SCA ancestry on chromosome 10, region 10q11.21-22 (Fig. 4b, c, and Supplementary Fig. 16). We identified significant variants in this region that have been functionally linked to several immune-related genes and potential targets for natural selection, including *MAPK8*, *WASHC2C* and *MARCH8*. Interestingly, both *MAPK8* and *WASHC2C* have been linked to smallpox virus infection and replication rates[37,38], which is of note considering that the Vikings were reported to be carriers of smallpox-like viruses[39]. Apart from smallpox, these genes have also been linked to influenza, bacterial pneumonia, tuberculosis[37], and other infectious diseases prominent in Middle Age Britain[40]. Furthermore, the region displays an enrichment of GWAS hits where SCA ancestry is associated with elevated erythrocyte and hemoglobin levels, and there is a higher prevalence of SCA ancestry among UKBB participants reporting lower incidences of "respiratory infection" and "influenza with pneumonia" (Supplementary Fig. 17).

Discussion

While our world is becoming increasingly admixed, genomic studies have largely focused on non-admixed populations with a pronounced European bias[8,41,42]. This is an issue when we consider that genomic models developed and trained in one population have poor portability outside that population[42–44]. Recently, strides have been made to address this gap, but most have been limited to cross-continental resolution, due to limitations in LAI accuracy[12–16].

To address this, we have developed Orchestra, a LAI model that can account for the genetic heterogeneity in recently admixed populations. Orchestra can work with reference panels made of many combined datasets. It is able to accurately retrace demographic histories of complex admixed populations, such as Latin Americans, and does so on a fine-grained regional scale. Further, Orchestra can be used to elucidate relationships between different populations. We offer a new lens into the ongoing debate about the origins of the Ashkenazi Jewish, supporting strong genetic ties to the Italian Peninsula[31–33]. Finally, Orchestra's local aspect enables us to apply it to downstream applications, such as detecting signatures of selection. We trace Scandinavian ancestry in British UKBB samples, which allows us to detect a potential immune-related signal on chromosome 10. This possibly Viking-derived region may have provided an edge against respiratory infections, such as smallpox, influenza or pneumonia. This region is, to this day, linked to a lower rate of respiratory infections and influenza in UKBB participants.

There is potential for improving the accuracy of Orchestra by creating more sophisticated reference panels. Some of the issues we had to overcome in this study were batch effects due to diverse sequencing technologies, insufficient coverage in more dated datasets, insufficient data from certain parts of the world and sample size imbalance between different reference populations. No reference panel can be perfect, as we are by definition breaking up genomic continuums into discrete populations. Still, we expect that improving the reference panel will generate further finer-scale insights into recent admixture around the globe.

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.

A limitation of Orchestra in its current iteration is its use of windows to infer local ancestry. With admixture, genomic segments become increasingly smaller with each generation due to cross-over and recombination. While Orchestra is modeled on recombination to reconstruct ancestry within a window, this results in a trade-off between accuracy and the ability to detect signals from further back in time between closely-related populations. Orchestra is relatively accurate at least up to around 12 generations of admixture and can detect trace ancestries from further back in time. This means it is suited to reconstruct events of relatively recent admixture, within Modern and potentially Medieval history. However, for reconstruction of Ancient history, other non-window based LAI models would have a clear advantage[45,46].

It is salient to note that while Orchestra provides a quality metric for ancestry assignments for each window, this metric is derived from the deep learning model (see Pseudo-probability vector in Supplementary Methods; Supplementary Fig. 18). It has not yet been calibrated to reflect real ancestral probabilities and has not been considered in this study. The deep learning quality metric is affected by a number of factors, including the size and quality of a reference population and how genetically close that population is to others in the reference panel. Ancestries not present in the reference panel will tend to classify as the phylogenetically closest available reference population, and the quality metric will not recognize those cases (Supplementary Fig. 19). These are important to keep in mind when using Orchestra. Considering the strengths, weaknesses and the coverage of the reference panel used for ancestry inference is key to interpreting Orchestra's results. We recognize the need to further develop and calibrate the confidence metrics, and we plan to address this in the future to enhance Orchestra's applicability.

An important limitation of our method is that it is very computationally intensive. While other methods we benchmarked 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. While it may

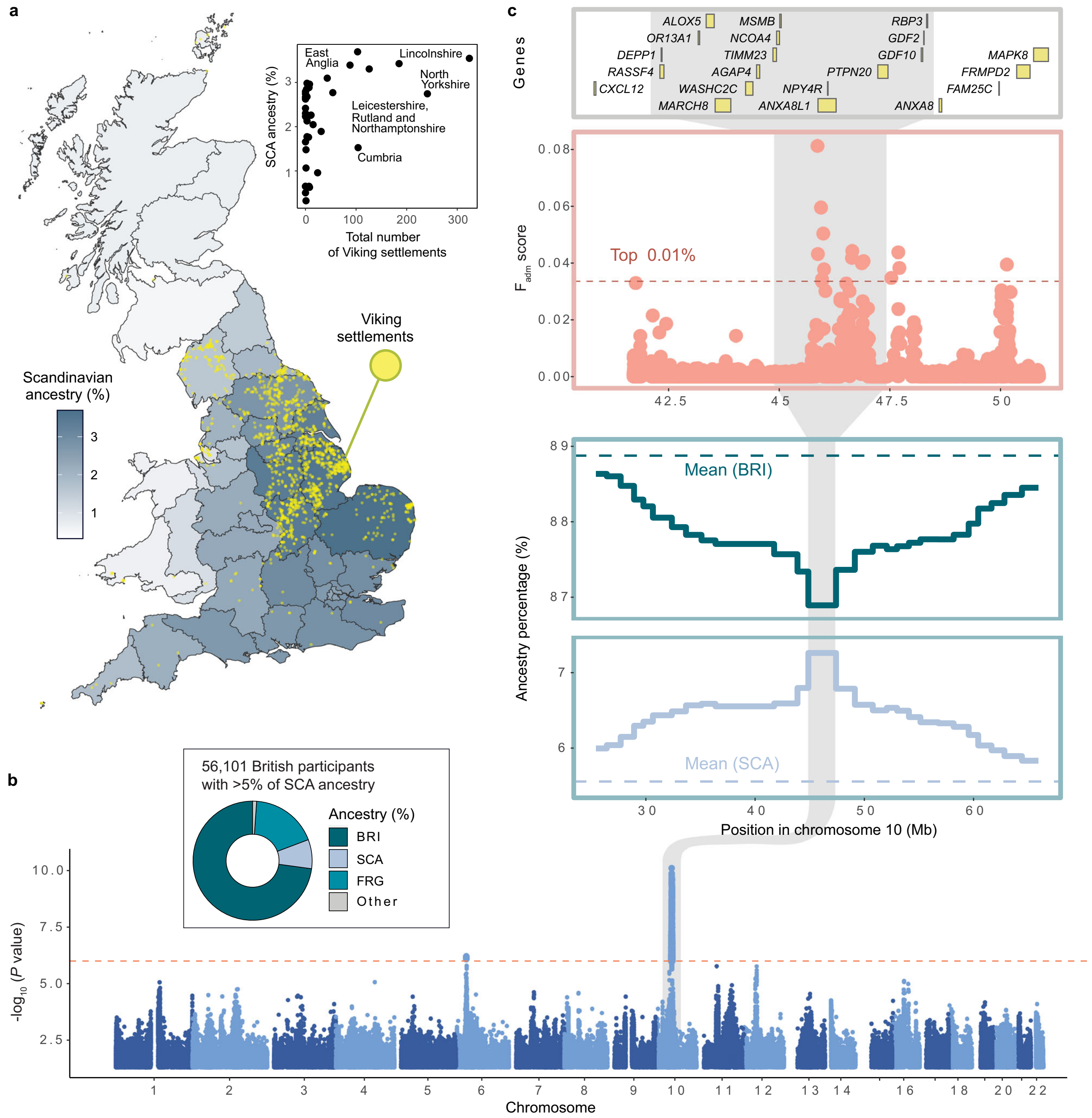

Fig. 4 | Signatures of Scandinavian ancestry and natural selection in British UK biobank participants. a Percentage of Scandinavian ancestry in the UK (blue gradient) against presumed Viking settlements (yellow dots), inferred from settlement names ending in: -by, -thorpe, or -toft. Inset shows the correlation between the total number of settlements with Viking-derived names and average percent of Scandinavian (SCA) ancestry per county. **b** Genome-wide adaptive admixture signals in 56,101 British individuals from the UK biobank dataset with over 5% SCA ancestry. Larger points indicate variants surpassing the significance threshold of $P < 10^{-6}$, denoted by a horizontal dotted line. The signal detected on chr10 is marked by a gray-shaded area. *P* values were derived from F_{adm} and LAD scores based on SNP rank and integrated using Fisher's method and evaluated with a two-sided chi-squared test. Inset, estimated ancestral composition for this sample set. BRI British & Irish; SCA Scandinavian; FRG French & German. **c** Local percentages of BRI and SCA ancestries on chromosome 10 windows. Average percentage levels are depicted by horizontal lines. F_{adm} score per variant and genes within the region are shown above.

allow the model to capture finer details than shorter and less memory-intensive training processes, it also makes our model more expensive and time-consuming to train. To mitigate this, we are making pre-trained models available for end-users, trained on a range of different populations and regional levels, which are ready for immediate use and can be used for rapid inference. However, for users who may need to retrain models for specific analyzes, but do not have the computational resources required, we recommend considering alternative software like RFmix, which has consistently demonstrated robust and reliable results across diverse populations at a region-wide level.

In conclusion, Orchestra advances the field of LAI, enabling accurate detection of chromosomal segments at regional levels. It also takes an important step towards a more equitable genomics, promising to improve a range of downstream applications, such as long-range phasing and GWAS, and by extent genomic and personalized medicine in admixed populations.

Methods

Orchestra

Orchestra is a LAI algorithm that consists of a two-stage pipeline: an original deterministic base layer and a smoothing module based on a deep learning architecture that combines elements of convolutional neural networks[47] and attention-based (transformer) mechanisms[48]. Orchestra's base layer is a deterministic algorithm modeled on recombination, built on the assumption that shared homologous haplotypes are identical by descent (IBD). We expect that a person would share longer DNA segments with individuals from the same population and shorter segments with individuals from remote populations. The base layer looks at each window on a chromatid and finds the minimum number of segments that would be needed to reconstruct that window when those segments are sampled from specific, carefully selected reference populations. We refer to this value as recombination distance. In this study, each chromosome was divided into windows spanning 600 SNPs.

The base layer works as a greedy search algorithm. In each window, the algorithm starts at the first position, looking for the longest continuous matching haplotype in a reference population. Where the match stops, the algorithm starts again from that position to find the longest local match, and so on. This is carried out using a NumPy array of the (boolean) matches at any given position between the sample and the reference sequences (as rows) and using the product along the rows. This allows to accelerate the computation in a trade-off of some use of memory for extra speed, plus it allows the calculation to be parallelized easily across the data. The procedure is repeated until it produces a vector of recombination distances against all reference populations.

The smoothing layer uses the vector of recombination distances produced by the base layer as an ancestry fingerprint that gets converted to a measure of ancestry in terms of probabilities. This layer is designed to give higher weights to low-frequency classes (populations) in the loss function to handle class imbalance effectively. Then, the information from surrounding and more remote windows is factored in, to output a final ancestry label. To do this effectively, the smoothing layer consists of two different types of layers: convolutional and attention-based. There are five convolutional and two attention layers. The attention layers are sandwiched between the third and fourth, and the fourth and fifth convolutional layers. The convolutional layers are moving filters that generate different insights from the base layer output and retain the information in parallel. Whereas in a normal convolutional neural network we would have pooling layers in between the convolutional layers[47], we use attention layers that process the result of the parallel filters as a (similarity) vector space that would typically be the output of an embedding into an n-dimensional vector space in a transformer architecture[48] to provide global information flow using proximity in this convolutional vector space. The purpose of the convolutional layers is to process information at the window level in the case of the first two convolutional layers and to bring local information concerning other nearby windows in the case of the third, fourth and fifth convolutional layers. The closest windows tend to have a larger impact due to the fact that windows tend to form blocks of a given ancestry, but the attention mechanism allows the use of a comprehensive context to weight the base layer outputs from all other windows. In this regard, the attention augments the local convolutional layers with global information flow. The simplification of the attention layer relative to regular transformer architecture allows the use of a context long enough to span the entirety of the windowed data for a chromosome pack. The final convolutional layer provides the output in terms of probabilities. The population that is assigned the maximum probability is given as the final ancestry.

Due to computation limitation, the smoothing layer was not trained on all chromosomes at a time, but was instead trained on chromosome packs (1/2, 3/4,..., 17/18, 19/20/21/22). Training on a larger set of chromosomes at a time, or even the entirety of the genome, is expected to further increase the accuracy of the smoothing layer.

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.

Reference panel

Despite the increasing number of publicly available datasets, obtaining a comprehensive and balanced reference panel remains a challenging step in ancestry deconvolution. Here, we used 1KGP-16pops (N = 1365), composed of unrelated non-admixed individuals collected by 1KGP, as the gold standard dataset, which enables easy accuracy comparisons with previously published studies. Next we created the custom-35pops dataset (N = 10,169), where we aimed to assemble a genome-wide dataset of non-admixed modern samples from diverse populations around the globe[49–83]. A detailed list of all data mined for this study can be found in the Supplementary Table 1. The granularity we were able to achieve on each continent was largely dependent on the number and diversity of samples we had 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. Precise information about dataset preparation and merging can be found in Supplementary Methods.

Our custom-35pops dataset is composed of many studies, including genotyping arrays and WGS, which gave rise to artifacts that interfered with the detection of biological patterns due to batch effects[84]. As to our knowledge, there are no effective and systematic algorithms to remove them, therefore we applied conventionally recommended quality control measures[85] (see Supplementary Methods).

After batch effect removal, we used two complementary strategies to identify admixed individuals. The first was PCA-UMAP, an existing protocol for dimensionality reduction[86] that allowed us to visualize relatedness among individuals. The second was GNN-tSNE, where we used tsinfer[5] to infer the tree sequences for our dataset and compute the genealogical nearest neighbors (GNN), to which we applied t-distributed stochastic neighbor embedding (t-SNE), another non-linear dimensionality reduction technique. Outliers (i.e., admixed genomes) were removed automatically using a K-Nearest Neighbor (KNN) algorithm (see Supplementary Methods).

To detect Native American ancestry in target genomes, we generate a reference panel of 100 pure in silico Native American genomes (see Supplementary Methods).

Simulated admixed individuals

SLiM v3.7[23] was used to generate admixed genomes based on single-ancestry populations from the reference panel. We forward simulated 1–6 generations of admixture. True local ancestry of every position in every simulated individual was tracked across generations using the tree-sequence recording function, and browsed with tskit and pyslim packages[87]. HapMap recombination map was supplied for modeling the non-uniform recombination events across the genome.

Given that SLiM does not support simulations with more than one chromosome, we ran each chromosome independently but followed the same mating scheme over generations to obtain whole-genome simulations with the same ancestors for all chromosomes. We tracked the pedigree obtained in the first chromosome run by tagging each simulated individual and each pair's offspring, and reproduced the same genealogy for the others.

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 ~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. We found that the benchmarking results were nearly identical. Thus, we opted to train Orchestra in the fully intermixed random-mating scenario to ensure robustness.

We ran a non-Wright-Fisher model of evolution, since we implemented a couple of modifications. Despite choosing parents via random sampling, we avoided inbreeding by recording pedigree information and thus identifying the degree of relatedness of each pair selected. Only individuals that were not close relatives were allowed to mate (at least a coefficient of relationship <6.25%). Besides, to avoid family relationships among output generations, we ran independent simulations for each degree of admixture, so the parents of the simulated individuals in a certain generation are not the simulated individuals in the previous generation. Finally, to avoid bottlenecks due to the small sample sizes of some of the populations in the reference panel, we generated a large number of individuals in the first generation directly (twice the initial number of the training cohort and four times the testing set). This resulted in 3276 simulated individuals for the 1KGP-16pops panel (546/generation), and 24,408 simulated individuals for the custom-35pops panel (4068/generation).

Benchmarking

For benchmarking, we inferred local ancestry with RFMix v2.03[9], FLARE v0.1.0[10] and Gnomix[11]. We split the samples randomly into training (80%) and testing (20%) cohorts, and measured performance globally, per generation, and per population. We report precision and recall as accuracy measures with scikit-learn[88], which are computed per population; i.e., the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) is calculated with respect to the selected population as the positive class. In the case of TN and FN, the prediction can be anything other than the population under consideration. We additionally report a commonly used alternative measure for accuracy by assessing each method with Pearson's r^2. We calculated the coefficient of determination by comparing both the inferred and the true local ancestry label per haplotype, as well as the diploid ancestry dose by counting the number of ancestry labels (0, 1 or 2) per site. Values of r^2 were estimated for each ancestry separately and the weighted mean r^2 across all ancestries was reported per generation.

All programs were supplied with the same training (as reference panel) and test (as target) genotype data, and the HapMap recombination map. Statistical parameters were kept at default values with the following exceptions: the Gnomix model was trained from scratch and its accuracy was optimized in our scenario using Large mode, smooth_size: 100, context_ratio: 0.10 and window_size_cM: 0.5; whereas the number of generations for RFmix was specified with -G 6. For the 1KGP-16pops WGS panel, we processed variants using a MAF filter (< 0.5%) to make RFmix computationally viable, a procedure similarly adopted by FLARE.

We assembled a new panel using UKBB samples for additional benchmarking, providing a distinct and independent validation dataset to assess Orchestra's ability to generalize beyond the original test set. We picked samples that (1) closely aligned with their respective ancestral groups based on both PCA + UMAP and GNN+t-sne dimensionality reduction approaches but were not incorporated into the final reference panel due to sample size limits, and/or (2) they were close to their population cluster but overlapped with other/s and therefore were excluded at the quality control stage. For regions like the Republic of Ireland, Wales, Scotland, and England, which had a larger sample base in this panel, we considered a maximum of 1000 samples per region. Altogether, we collected 10,241 samples spanning 103 countries. Accuracy was evaluated by attributing the appropriate ancestry label from our ancestry reference panel to each country and then comparing it with the LAI results from each method (Supplementary Fig. 2c).

Retracing genetic histories

We simulated Latin American individuals from Southern (SPP and ITA) and Northern (FRG and BRI) Europeans, Western (GSE, GLS and NGE) and Central and Southern (SAF) Africans and artificially-reconstructed Native Americans (NAM) from our custom ancestry panel. Simulations were performed by emulating genetic intermixing for 12 generations using SLiM[23]. Details about the simulations can be found in the Supplementary Methods section. For our benchmarking evaluation, we partitioned the samples randomly into training and testing groups, allocating 80% and 20% of the simulations, respectively, to assess Orchestra against other LAI tools. We ensured that training and testing samples never coincided across the three simulated regional datasets, guaranteeing consistent comparison metrics. We evaluated the performance in terms of precision and recall for each Latin American region using scikit-learn.

For real-world data analysis, we identified UKBB participants born in the Americas using birthplace codes (data-field f.20115), specifically from South America (>C.600) and North America (C.400-C.500). Additionally, we incorporated admixed American 1KGP samples (codes: MXL, PUR, CLM and PEL). We then extracted the composite SNP set from both the imputed UKBB and whole-genome 1KGP sets and ran Orchestra for LAI assessment. Samples with ancestry patterns closely matching British ancestry (BRI + FRG + SCA > 80%) were discarded (Supplementary Fig. 23).

Finally, we applied Orchestra to all UKBB samples not selected for our reference panel, and samples obtained from various datasets, especially from Reich lab, that were not included in our reference panel because they belonged to ethnic groups other than those found in our 35 reference populations (Supplementary Figs. 5-7).

Ancestral mapping

We created 35 different reference panels for each target population, by excluding the target population from its own reference panel. Orchestra was then trained on each reference panel and applied to the target population. Due to computational limitations, training on these 35 datasets was limited to chromosomes 17–22. The admixture proportions obtained for each population were converted into a matrix of distances that were projected onto two-dimensional space using the SMACOF (Scaling by MAjorizing a COmplicated Function) algorithm implementation of scikit-learn using the sklearn.manifold.MDS function with the option metric=True to convert a non-Euclidean symmetric dissimilarity matrix to coordinates in 2 dimensional space. The SMACOF algorithm uses a random initialization followed by stress majorization to iteratively minimize a stress function given by the squared difference between the dissimilarity matrix entries and the Euclidean distances between the points in 2 dimensional space. To create the matrix of distances, the obtained series of ancestry proportions were first averaged over the samples for each population. These averages were then assembled inside a two dimensional matrix with each dimension being the number of the populations, with 0 on the diagonal. 0.00000000000000001 was then added to all the entries of the matrix to remove any 0 entries and allow the reciprocal to be taken. The matrix was then added to its transpose, to make it symmetric, which was a necessity to apply the algorithm. The reciprocal was taken to convert the similarity measures of proportions into dissimilarity, before being input into the sklearn.manifold.MDS function to produce the coordinates.

Detecting natural selection signatures

To detect natural selection in admixed populations we focused on the F_{adm} and LAD statistics described in Cuadros-Espinoza et al.[34]. We explain the steps in detail in the Supplementary Methods section. We first scanned the genomes of seven admixed populations already analyzed in Cuadros-Espinoza et al.[34], leveraging either publicly available genotypes[89–93] or an appropriate proxy from the UKBB or Reich lab. These served as positive controls to gauge the accuracy in replicating previously identified signals. Detailed information on the datasets and the respective references for these populations is provided in Supplementary Fig. 15. External datasets were lifted over to hg38 and imputed, followed by the extraction of the composite SNP set for LAI assessment with Orchestra. In the absence of a direct proxy for Malagasy, we considered African, Black or mixed participants from the UKBB with a significant proportion of Austronesian (SEA + FIL > 1%) and South African (SCA > 20%) ancestry and low Indian (PNI + BEI + SSI < 20%) heritage (89% of these participants being from Southeast African countries).

We then extended this methodology to the White British from the UK, where we treated the British population as admixed, and looked at the Scandinavian component in British genomes. We analyzed 415,859 British participants from the UKBB dataset with detailed birth location information within the UK (using north and east coordinates or data-fields f.129 and f.130). Birth locations were categorized according to the 2018 NUTS Level 2 boundaries (counties) from the Office for National Statistics (https://data.gov.uk/). Shapefiles were loaded and processed using rgdal R package. Scandinavian (SCA) average percentage was next computed based on individuals mapped to each specific geographic area in Britain. Additionally, we retrieved the index of place names in Great Britain (July 2016) to pinpoint those towns or cities with Viking-origin names, suggesting past Viking settlements (evidenced by suffixes such as -by, -thorpe, or -toft). Next we took 287,346 British samples that showed traces of SCA ancestry. Due to the emergence of a signal in chromosome 10, we wanted to ensure the validity of this observation by analyzing three additional sets of British samples, each varying in their Scandinavian ancestry enrichment. Specifically, we assessed British samples bearing >1%, >5%, and >20% Scandinavian ancestry (Supplementary Fig. 16). Given that SCA is the population with the lowest accuracy in our panel (Fig. 1c), we removed 10 windows from the telomeric regions. We further refined the signal by adjusting the averaged SCA ancestry percentage present in every window against the standard deviation observed in its chromosome pack. Chr10 signal remained consistent across sets as illustrated in Supplementary Fig. 16.

OpenTargets. We selected variants that surpassed the established *P* value significance threshold and explored the Open Targets database (https://www.opentargets.org/) for potential target genes. These genes were then ranked based on the aggregate score obtained from all variants.

GWAS enrichment. To investigate if the number of GWAS Catalog (http://www.ebi.ac.uk/gwas/) [release 2023-07-20] hits in the selected region is higher than expected by chance, we crossed GWAS Catalog signals ($P < 10^{-8}$) with 1000 Genomes Project variants and grouped together those in high LD ($r^2 \geq 0.8$) and associated to the same phenotype. Enrichment *P* values were calculated by comparison with a null distribution from random genomic regions as a background model, controlling by region size and excluding gaps, sexual chromosomes and the major histocompatibility complex region, known to harbor a vast number of associations. Since the size of the region to be studied is ~2.5 Mb, the number of simulations cannot be too large without covering the entire genome and having overlapping simulations. Thus, we also tested GWAS enrichment with a Fisher's exact test, which showed very similar results (odds ratio correlation: $R^2 = 0.97$; *P* value correlation: $R^2 = 0.69$). We selected this latter statistic for the analyzes due to its higher statistical power. GWAS Catalog reported traits were grouped by parent categories according to EFO terms from the ontologyIndex R package. *P* values were adjusted by Bonferroni correction.

To determine whether this region in chr10 of Scandinavian ancestry, as opposed to British, would result in an increase or decrease for each GWAS phenotype, we computed the frequency difference of the GWAS variants between these two populations using our custom ancestry panel, providing insight into the potential direction of the effect resulting from this shift in ancestry.

Phenotype mapping. Following the association strategy of admixture mapping, we contrasted the proportion of SCA ancestry among cases and controls to evaluate the influence of the chr10 locus on phenotypes gathered by the UKBB. For that, we took main and secondary ICD10 codes and self-reported illness codes (data-fields f.41202, f.41204 and f.20002, respectively). Phenotypes represented by fewer than 100 cases were excluded. Given the absence of significant phenotypes using the Fisher exact test and subsequent *P* value correction, we filtered phenotypes based on their nominal *P* value ($P < 0.05$). Additionally, we only considered those phenotypes with a minimum of two Scandinavian haplotypes in the cases set to mitigate biases stemming from the infrequent Scandinavian ancestry. Phenotypes were then ranked according to their impact or odds ratio.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The 1KG-16pops simulated dataset generated in this study has been deposited in the zenodo database under https://doi.org/10.5281/zenodo.15147095. This dataset includes both original 1KGP samples and simulated samples representing six generations of random admixture generated using SLiM and contains both genotype data and local ancestry annotations. The custom-35pops reference and simulated dataset are unavailable because parts of this dataset are under restricted access, for reasons of patient confidentiality. However, this panel can be recreated after obtaining access to individual datasets listed below, and models trained on this dataset are freely available (see Code availability). The UK Biobank data are available under restricted access; access can be obtained by application to the UK Biobank (https://www.ukbiobank.ac.uk/using-the-resource/). The POPRES data are available in the dbGaP database under accession code phs000145.v4.p2 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000145.v4.p2) under restricted access; access can be obtained by following the dbGaP application process. The Ashkenazi Jewish and Gambian Genome Variation Project data are available in the European Genome-phenome Archive (https://ega-archive.org) under accession codes EGAS00001000664 and EGAS00001001311, respectively, under restricted access; access can be obtained by request to the EGA Data Access Committee. The 1000 Genomes Project data (https://www.internationalgenome.org/data-portal/data-collection/30x-grch38), Human Genome Diversity Project data (https://www.internationalgenome.org/data-portal/data-collection/hgdp), Simons Genome Diversity Project data (https://www.internationalgenome.org/data-portal/data-collection/sgdp), and Korean Personal Genome Project data (http://opengenome.net/index.php/Korean) are publicly available. The modern-day data from Dr David Reich's laboratory are publicly available via the Allen Ancient DNA Resource (https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data)[94]. The Middle Eastern WGS data from Almarri et al.[57] can be found at ftp://ngs.sanger.ac.uk/production/appg/. The Berber and Arab datasets from Arauna et al.[71] and Anagnostou et al.[69] are available at https://figshare.com/articles/North_African_Berber_dataset/3501761 and https://zenodo.org/records/3546051, respectively. The Basque data from Flores-Bello et al.[78] are available at https://figshare.com/s/61a54472b63fd0101859. The data from Botigué et al.[77] and Henn et al.[70] can be accessed through the Human Genome Diversity Panel at the Institut de Biologia Evolutiva https://www.biologiaevolutiva.org/dcomas/software-data/ and https://figshare.com/s/61a54472b63fd0101859. The Sudanese data from Hollfelder et al.[72] and Dobon et al.[73] are available via Dryad (https://datadryad.org/dataset/doi:10.5061/dryad.bs06h) and at https://www.upf.edu/web/evolutionary-systems-biology. Data from Yunusbayev et al.[75,76], Pathak et al.[79], Mörseburg et al.[83], Tätte et al.[82], Tambets et al.[51], and Behar et al.[50,74] can be accessed through the Estonian Biocentre (https://evolbio.ut.ee/). The East Indonesian data from Hudjashov et al.[90] can be accessed at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE80534. The Fulani data from Vicente et al.[92] can be accessed at https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-8434. Other publicly available resources used in this study include the HapMap recombination map (https://github.com/odelaneau/shapeit4/tree/master/maps), the GWAS Catalog (https://www.ebi.ac.uk/gwas), the Open Targets database (https://www.opentargets.org/), and the 2018 NUTS Level 2 boundaries (https://data.gov.uk/).

Code availability

Orchestra is available from GitHub: https://github.com/omicsedge/orchestra-paper; https://doi.org/10.5281/zenodo.14946430. Pre-trained models are available at https://doi.org/10.5281/zenodo.14946712. A toy example is available on CodeOcean. https://doi.org/10.24433/CO.2416186.v1. A more complex toy example that explores selection signals for adaptive admixture in admixed Mexicans is available at https://doi.org/10.5281/zenodo.14949924.

References

1. Cann, R. L., Stoneking, M. & Wilson, A. C. Mitochondrial DNA and human evolution. *Nature* **325**, 31–36 (1987).
2. Rosenberg, N. A. et al. Genetic structure of human populations. *Science* **298**, 2381–2385 (2002).
3. Harcourt, A. H. Human phylogeography and diversity. *Proc. Natl Acad. Sci. USA.* **113**, 8072–8078 (2016).
4. Nielsen, R. et al. Tracing the peopling of the world through genomics. *Nature* **541**, 302–310 (2017).
5. Kelleher, J. et al. Inferring whole-genome histories in large population datasets. *Nat. Genet.* **51**, 1330–1338 (2019).
6. Bergström, A. et al. Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**, eaay5012 (2020).
7. Atkin, A. L., Christophe, N. K., Stein, G. L., Gabriel, A. K. & Lee, R. M. Race terminology in the field of psychology: Acknowledging the growing multiracial population in the U.S. *Am. Psychol.* **77**, 381–393 (2022).
8. Tan, T. & Atkinson, E. G. Strategies for the genomic analysis of admixed populations. *Annu. Rev. Biomed. Data Sci.* **6**, 105–127 (2023).
9. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013).
10. Browning, S. R., Waples, R. K. & Browning, B. L. Fast, accurate local ancestry inference with FLARE. *Am. J. Hum. Genet.* **110**, 326–335 (2023).
11. Hilmarsson, H. et al. High resolution ancestry deconvolution for next generation genomic data. Preprint at bioRxiv 2021.09.19.460980 https://doi.org/10.1101/2021.09.19.460980 (2021).
12. Atkinson, E. G. et al. Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power. *Nat. Genet.* **53**, 195–204 (2021).
13. Gay, N. R. et al. Impact of admixture and ancestry on eQTL analysis and GWAS colocalization in GTEx. *Genome Biol.* **21**, 233 (2020).
14. Patel, R. A. et al. Genetic interactions drive heterogeneity in causal variant effect sizes for gene expression and complex traits. *Am. J. Hum. Genet.* **109**, 1286–1297 (2022).
15. Sun, Q. et al. Improving polygenic risk prediction in admixed populations by explicitly modeling ancestral-differential effects via GAUDI. *Nat. Commun.* **15**, 1016 (2024).
16. Marnetto, D. et al. Ancestry deconvolution and partial polygenic score can improve susceptibility predictions in recently admixed individuals. *Nat. Commun.* **11**, 1628 (2020).
17. Choudhury, A. et al. High-depth African genomes inform human migration and health. *Nature* **586**, 741–748 (2020).
18. Chan, S. H. et al. Analysis of clinically relevant variants from ancestrally diverse Asian genomes. *Nat. Commun.* **13**, 6694 (2022).
19. Novembre, J. et al. Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008).
20. Irving-Pease, E. K. et al. The selection landscape and genetic legacy of ancient Eurasians. *Nature* **625**, 312–320 (2024).
21. McInnes, L. et al. UMAP: uniform manifold approximation and projection. *J. Open Source Softw.* **3**, 861 (2018).
22. van der Maaten, L. & Hinton, G. Visualizing data using t-sne. *J. Mach. Learn. Res.* **9**, 2579–2605, https://www.jmlr.org/papers/v9/vander7maaten08a.html (2008).
23. Haller, B. C. & Messer, P. W. SLiM 3: Forward Genetic Simulations Beyond the Wright-Fisher Model. *Mol. Biol. Evol.* **36**, 632–637 (2019).
24. Ongaro, L. et al. The genomic impact of European colonization of the Americas. *Curr. Biol.* **29**, 3974–3986.e4 (2019).
25. Chacón-Duque, J. C. et al. Latin Americans show wide-spread Converso ancestry and imprint of local Native ancestry on physical appearance. *Nat. Commun.* **9**, 5388 (2018).
26. Ruiz-Linares, A. et al. Admixture in Latin America: geographic structure, phenotypic diversity and self-perception of ancestry based on 7342 individuals. *PLoS Genet* **10**, e1004572 (2014).
27. Homburger, J. R. et al. Genomic insights into the ancestry and demographic history of South America. *PLoS Genet* **11**, e1005602 (2015).
28. Roopnarine, L. East Indian indentured emigration to the Caribbean: beyond the push and pull model. *Caribb. Stud.* **31**, 97–134 (2003).
29. Durand, J. & Massey, D. S. New world orders: continuities and changes in Latin American Migration. *Ann. Am. Acad. Pol. Soc. Sci.* **630**, 20–52 (2010).
30. Peter, B. M., Petkova, D. & Novembre, J. Genetic landscapes reveal how human genetic diversity aligns with geography. *Mol. Biol. Evol.* **37**, 943–951 (2020).
31. Costa, M. D. et al. A substantial prehistoric European ancestry amongst Ashkenazi maternal lineages. *Nat. Commun.* **4**, 2543 (2013).
32. Waldman, S. et al. Genome-wide data from medieval German Jews show that the Ashkenazi founder event pre-dated the 14th century. *Cell* **185**, 4703–4716.e16 (2022).

33. Xue, J., Lencz, T., Darvasi, A., Pe'er, I. & Carmi, S. The time and place of European admixture in Ashkenazi Jewish history. *PLoS Genet* **13**, e1006644 (2017).
34. Cuadros-Espinoza, S., Laval, G., Quintana-Murci, L. & Patin, E. The genomic signatures of natural selection in admixed human populations. *Am. J. Hum. Genet.* **109**, 710–726 (2022).
35. Gretzinger, J. et al. The Anglo-Saxon migration and the formation of the early English gene pool. *Nature* **610**, 112–119 (2022).
36. Leslie, S. et al. The fine-scale genetic structure of the British population. *Nature* **519**, 309–314 (2015).
37. Chen, J. et al. The roles of c-Jun N-terminal kinase (JNK) in Infectious Diseases. *Int. J. Mol. Sci.* **22**, 9640 (2021).
38. Huang, C. Y. et al. A novel cellular protein, VPEF, facilitates vaccinia virus penetration into HeLa cells through fluid phase endocytosis. *J. Virol.* **82**, 7988–7999 (2008).
39. Mühlemann, B. et al. Diverse variola virus (smallpox) strains were widespread in northern Europe in the Viking Age. *Science* **369**, eaaw8977 (2020).
40. Robb, J., Cessford, C., Dittmar, J., Inskip, S. A. & Mitchell, P. D. The greatest health problem of the Middle Ages? Estimating the burden of disease in medieval England. *Int. J. Paleopathol.* **34**, 101–112 (2021).
41. Korunes, K. L. & Goldberg, A. Human genetic admixture. *PLoS Genet* **17**, e1009374 (2021).
42. Martin, A. R. et al. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
43. Belbin, G. M. et al. Toward a fine-scale population health monitoring system. *Cell* **184**, 2068–2083.e11 (2021).
44. Privé, F. et al. Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. *Am. J. Hum. Genet.* **109**, 12–23 (2022).
45. Chintalapati, M., Patterson, N. & Moorjani, P. The spatiotemporal patterns of major human admixture events during the European Holocene. *Elife* **11**, e77625 (2022).
46. Wangkumhang, P., Greenfield, M. & Hellenthal, G. An efficient method to identify, date, and describe admixture events using haplotype information. *Genome Res* **32**, 1553–1564 (2022).
47. Krizhevsky, A., Sutskever, I. & Hinton, G. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2017).
48. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 5998–6008 (2017).
49. Byrska-Bishop, M. et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426–3440(2022).
50. Behar, D. M. et al. The genome-wide structure of the Jewish people. *Nature* **466**, 238–242 (2010).
51. Tambets, K. et al. Genes reveal traces of common recent demographic history for most of the Uralic-speaking populations. *Genome Biol.* **19**, 139 (2018).
52. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
53. Carmi, S. et al. Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins. *Nat. Commun.* **5**, 4835 (2014).
54. Kim, J. et al. KoVariome: Korean national standard reference Variome database of whole genomes with comprehensive SNV, indel, CNV, and SV analyses. *Sci. Rep.* **8**, 5677 (2018).
55. Lowy-Gallego, E. et al. Variant calling on the GRCh38 assembly with the data from phase three of the 1000 Genomes Project. *Wellcome Open Res* **4**, 50 (2019).
56. Mallick, S. et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
57. Almarri, M. A. et al. The genomic history of the Middle East. *Cell* **184**, 4612–4625.e14 (2021).
58. Malaria Genomic Epidemiology Network. Insights into malaria susceptibility using genome-wide data on 17,000 individuals from Africa, Asia and Oceania. *Nat. Commun.* **10**, 5732 (2019).
59. Zhang, W. et al. Whole genome sequencing of 35 individuals provides insights into the genetic architecture of Korean population. *BMC Bioinforma.* **15**, S6 (2014).
60. Wang, C. C. et al. Genomic insights into the formation of human populations in East Asia. *Nature* **591**, 413–419 (2021).
61. Jeong, C. et al. The genetic history of admixture across inner Eurasia. *Nat. Ecol. Evol.* **3**, 966–976 (2019).
62. Biagini, S. A. et al. People from Ibiza: an unexpected isolate in the Western Mediterranean. *Eur. J. Hum. Genet.* **27**, 941–951 (2019).
63. Vyas, D. N., Al-Meeri, A. & Mulligan, C. J. Testing support for the northern and southern dispersal routes out of Africa: an analysis of Levantine and southern Arabian populations. *Am. J. Phys. Anthropol.* **164**, 736–749 (2017).
64. Skoglund, P. et al. Reconstructing prehistoric African population structure. *Cell* **171**, 59–71 (2017).
65. Skoglund, P. et al. Genomic insights into the peopling of the Southwest Pacific. *Nature* **538**, 510–513 (2016).
66. Lazaridis, I. et al. Genomic insights into the origin of farming in the ancient Near East. *Nature* **536**, 419–424 (2016).
67. Lazaridis, I. et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409–413 (2014).
68. Pickrell, J. K. et al. The genetic prehistory of southern Africa. *Nat. Commun.* **3**, 1143 (2012).
69. Anagnostou, P. et al. Berbers and Arabs: tracing the genetic diversity and history of Southern Tunisia through genome wide analysis. *Am. J. Phys. Anthropol.* **173**, 697–708 (2020).
70. Henn, B. M. et al. Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genet* **8**, e1002397 (2012).
71. Arauna, L. R. et al. Recent Historical Migrations Have Shaped the Gene Pool of Arabs and Berbers in North Africa. *Mol. Biol. Evol.* **34**, 318–329 (2017).
72. Hollfelder, N. et al. Northeast African genomic variation shaped by the continuity of indigenous groups and Eurasian migrations. *PLoS Genet* **13**, e1006976 (2017).
73. Dobon, B. et al. The genetics of East African populations: a Nilo-Saharan component in the African genetic landscape. *Sci. Rep.* **5**, 9996 (2015).
74. Behar, D. M. et al. No evidence from genome-wide data of a Khazar origin for the Ashkenazi Jews. *Hum. Biol.* **85**, 859–900 (2013).
75. Yunusbayev, B. et al. The Caucasus as an asymmetric semipermeable barrier to ancient human migrations. *Mol. Biol. Evol.* **29**, 359–365 (2012).
76. Yunusbayev, B. et al. The genetic legacy of the expansion of Turkic-speaking nomads across Eurasia. *PLoS Genet* **11**, e1005068 (2015).
77. Botigué, L. R. et al. Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. *Proc. Natl Acad. Sci. USA.* **110**, 11791–11796 (2013).
78. Flores-Bello, A. et al. Genetic origins, singularity, and heterogeneity of Basques. *Curr. Biol.* **31**, 2167–2177 (2021).
79. Pathak, A. K. et al. The genetic ancestry of modern Indus Valley populations from Northwest India. *Am. J. Hum. Genet.* **103**, 918–929 (2018).
80. Nelson, M. R. et al. The population reference sample, POPRES: a resource for population, disease, and pharmacological genetics research. *Am. J. Hum. Genet.* **83**, 347–358 (2008).
81. Changmai, P. et al. Indian genetic heritage in Southeast Asian populations. *PLoS Genet* **18**, e1010036 (2022).
82. Tätte, K. et al. The genetic legacy of continental scale admixture in Indian Austroasiatic speakers. *Sci. Rep.* **9**, 3818 (2019).
83. Mörseburg, A. et al. Multi-layered population structure in Island Southeast Asians. *Eur. J. Hum. Genet.* **24**, 1605–1611 (2016).

84. Leek, J. T. et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11**, 733–739 (2010).
85. Laurie, C. C. et al. Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet. Epidemiol.* **34**, 591–602 (2010).
86. Diaz-Papkovich, A., Anderson-Trocmé, L., Ben-Eghan, C. & Gravel, S. UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLoS Genet* **15**, e1008432 (2019).
87. Kelleher, J., Etheridge, A. M. & McVean, G. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput. Biol.* **12**, e1004842 (2016).
88. Pedregosa, F. et al. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830(2011).
89. Pierron, D. et al. Strong selection during the last millennium for African ancestry in the admixed population of Madagascar. *Nat. Commun.* **9**, 932 (2018).
90. Hudjashov, G. et al. Complex patterns of admixture across the Indonesian archipelago. *Mol. Biol. Evol.* **34**, 2439–2452 (2017).
91. Moreno-Estrada, A. et al. Human genetics. the genetics of Mexico recapitulates Native American substructure and affects biomedical traits. *Science* **344**, 1280–1285 (2014).
92. Vicente, M. et al. Population history and genetic adaptation of the Fulani nomads: inferences from genome-wide data and the lactase persistence trait. *BMC Genomics* **20**, 915 (2019).
93. Laso-Jadart, R. et al. The genetic legacy of the indian ocean slave trade: recent admixture and post-admixture selection in the Makranis of Pakistan. *Am. J. Hum. Genet.* **101**, 977–984 (2017).
94. Mallick, S. et al. The Allen Ancient DNA Resource (AADR) a curated compendium of ancient human genomes. *Sci. Data* **11**, 182 (2024).

Acknowledgements

We thank the team and our colleagues at Omics Edge and Genius Labs. This research was conducted by using the UK Biobank Resource under application number 84038. All work was funded by a commercial source, Omics Edge, a subsidiary of Genius Labs Company. Omics Edge provided only funding for the study, but had no additional role in study design, data collection and analysis, decision to publish or preparation of the manuscript beyond the funding of the contributors' salaries.

Author contributions

Conceptualization: J.L.J., B.N. and P.G.Y.; Data curation: J.L.J., B.N., V.B., M.R.H., H.P., S.B., A.D.M., A.A.M. and K.O.B.; Formal analysis: J.L.J., B.N., D.U., V.B., M.R.H., C.M., H.P., A.O., A.T., and M.G.G.; Methodology: J.L.J., B.N., D.U., V.B., C.M., U.K., M.G.G., P.G.Y.; Investigation: J.L.J., B.N., D.U., V.B., M.R.H., C.M., H.P., A.O., M.G.G.; Software: D.U., V.B, C.M., A.O., A.T., S.B., A.D.M., A.A.M., K.O.B., U.K., M.G.G., Visualization: J.L.J. and B.N.; Writing - original draft: J.L.J. and B.N.; Writing - review & editing: J.L.J,.B.N. and P.G.Y. Funding acquisition, Project administration, Resources & Supervision: P.G.Y.

Competing interests

All authors are either employed by and/or hold stock or stock options in Omics Edge, a subsidiary of Genius Labs. In addition, P.G.Y. has equity in Systomic Health LLC and Ethobiotics LLC. This work has been used to file a provisional patent application. There are no other relevant activities or financial relationships which have influenced this work.

Inclusion and diversity

We support inclusive, diverse, and equitable conduct of research.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-59936-3.

Correspondence and requests for materials should be addressed to Puya G. Yazdi.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.