Genome Biology

METHODOLOGY

Open Access

PANE: fast and reliable ancestral reconstruction on ancient genotype data with non-negative least square and principal component analysis

Luciana de Gennaro[1*], Ludovica Molinaro[2*], Alessandro Raveane[3], Federica Santonastaso[3], Sandro Sublimi Saponetti[1], Michela Carlotta Massi[3], Luca Pagani[4,5], Mait Metspalu[5], Garrett Hellenthal[6], Toomas Kivisild[2,5], Mario Ventura[1*] and Francesco Montinaro[1,5*]

*Correspondence: luciana.degennaro@uniba.it; ludovica.molinaro@kuleuven.be; mario.ventura@uniba.it; francesco.montinaro@gmail.com

[1] Department of Biosciences, Biotechnology and Environment, University of Bari, Bari, Italy
[2] Department of Human Genetics, KU Leuven, Leuven, Belgium
[3] Human Technopole, Milan, Italy
[4] Department of Biology, University of Padova, Padua, Italy
[5] Institute of Genomics, University of Tartu, Tartu, Estonia
[6] Department of Genetics, Evolution & Environment, University College of London, London, UK

Abstract

The history of human populations has been strongly shaped by admixture events, contributing to patterns of observed genetic diversity across populations. In this study, we introduce the Principal component Ancestry proportions using NNLS Estimation (PANE) method that leverages principal component analysis and non-negative least squares to assess the ancestral compositions of admixed individuals given a large set of populations. Our results show its ability to reliably estimate ancestry across several scenarios, even those with a significant proportion of missing genotypes, in a fraction of the time required when using other tools.

Background

The history of human populations has been strongly shaped by past admixture events that cumulatively have contributed to patterns of genetic variation observed today [1, 2]. Several interdisciplinary studies proved that virtually all human populations have interacted throughout their history in complex demographic scenarios, including migration and admixture [3–6]. These interactions resulted in a sudden or gradual transfer of genetic material, generating new groups different from their sources [1]. Given its significance for evolutionary and medical studies, many algorithms focusing on the inference of the genetic composition of admixed populations have been developed. In this context, it has been shown that using phased genotype data can offer a higher resolution description of genetic population structure compared to unphased data [1, 2, 4, 7–9].

However, existing methods often present limitations when dealing with low-coverage ancient DNA (aDNA) data. Algorithms using haploid-called genotypes to estimate allele

frequencies and allele-sharing probabilities at limited numbers of overlapping variant positions have been designed to meet these challenges.

Imputation is a commonly used method to overcome the limitations of low-coverage ancient DNA (aDNA), as it can increase the information content in a sample by inferring missing SNPs. This approach has been successfully used for ancient DNA data, allowing a better overall resolution [10–12]. However, the accuracy of genotype imputation relies on several factors, such as the quality of the sample data and the properties of the phased reference panel.

Among others, qpAdm [13, 14] is one of the most widely used approaches to ancient data [9, 15–21], given its ability to deal with pseudo-haploid genotypes typically used in aDNA analysis and model admixture events involving multiple sources [5, 13, 14]. This tool takes advantage of the fact that genetic variation within a specific population can be summarized by comparing its allele frequencies to those of three additional groups using a "treeness test" belonging to the *F*-statistics family, the *f4* metric [14, 22].

For a given target population T, a set of putative sources of admixture P_i, and a set of "right populations" R_i with different relationships to P_i, qpAdm builds a matrix A of *f4* in the form (T, X, R_1, R_i), in which X can alternatively be T or a P_i population. Given that any *f4* in the state (T, T, R_1, R_i) is 0, qpAdm solves the equation $w \bullet A = 0$, where w are the admixture coefficients (weights), assuming that their sum is equal to 1 [13].

QpAdm framework can be iterated multiple times to test several scenarios, allowing the evaluation of the models based on their p-values. However, sifting through all possible proxy sources and the right populations for an admixture event can be overwhelming. In addition, a recent survey has shown that, depending on the approach and the quality of the genetic data analyzed, qpAdm may suffer from high false discovery rates, adding substantial uncertainty to the interpretation of the results of admixture inference [23].

A similar approach, introduced by Haak et al. [13], but less frequently employed, uses a non-negative least squares (NNLS) approach on a matrix of *f4*s in the form *f4*(X, R_1, R_i, R_j), where X is either T or any P population [13, 24].

F-statistics results broadly recapitulate genetic relationships emerging from principal component analysis (PCA) [25], widely used in population genetics to quantify genetic affinity between populations or individuals, including ancient ones.

There is indeed a geometric relationship between the two metrics, although they are based on different statistical principles: the *F*-statistic is based on the measurement of the branch lengths of a hypothetical tree in which the analyzed populations are related, while PCA reduces the dimensionality of the data while maintaining the maximum variance present among individuals. In detail, considering four populations A, B, C, and D projected in a PC space, the *f2*(A, B) is correlated with the Euclidean distance between A and B computed in PC coordinates, while the f3(A; B, C) will be proportional to the orthogonal projection of A–B on A–C. Similarly, the *f4*(A, B; C, D) will be related to the orthogonal projection of A–B onto C–D [25]. Moorjani et al. showed that *f4* ratios can be used to estimate the rate of admixture [26].

Considering these results, it is, in principle, possible to use PC coordinates to infer admixture proportions of a target population using a set of putative sources. Different attempts and approaches have recently been proposed using principal components [27].

In this study, we present PANE (Principal component Ancestry using NNLS Estimation), in which we aim to leverage PCA and NNLS to assess the ancestral compositions of admixed individuals given a large set of populations.

Similarly to tools like qpAdm, which leverages a set of reference putative sources to describe the target admixed samples, we expect the assignment to be more accurate if proxy sources are genetically close to the true source of the admixture. Any post-admixture event (i.e., strong drift, gene flow) that increases the genetic distances between true sources, proxy sources, and target groups might cause a decrease in the accuracy of the assignments. We test PANE on different simulated models, incorporating high levels of missingness. We show its ability to reliably estimate ancestry across numerous scenarios, even those with a significant proportion of missing genotypes, in a fraction of the time required when using other tools.

Results

PANE workflow and datasets

Here we provide an overview of the methodology implemented in PANE (Fig. 1). We simulated a set of 20 unadmixed and 16 admixed populations (Additional file 1: Fig. S1, Additional file 2: Table S1). For each admixed group, we simulated an admixture event involving two or three sources [28] with minor source contributions ranging from 5 to 40%, to test PANE performance in various conditions and settings, accommodating a wide range of routinely performed approaches. For each of the true sources, we also simulated a sister group that split 3 thousand years ago (KYA) to mimic a proxy source: a group related to the real admixing source but not the direct contributor to the admixture event. These proxy populations allowed us to test whether PANE could infer the closest proxy sources to the admixing populations.

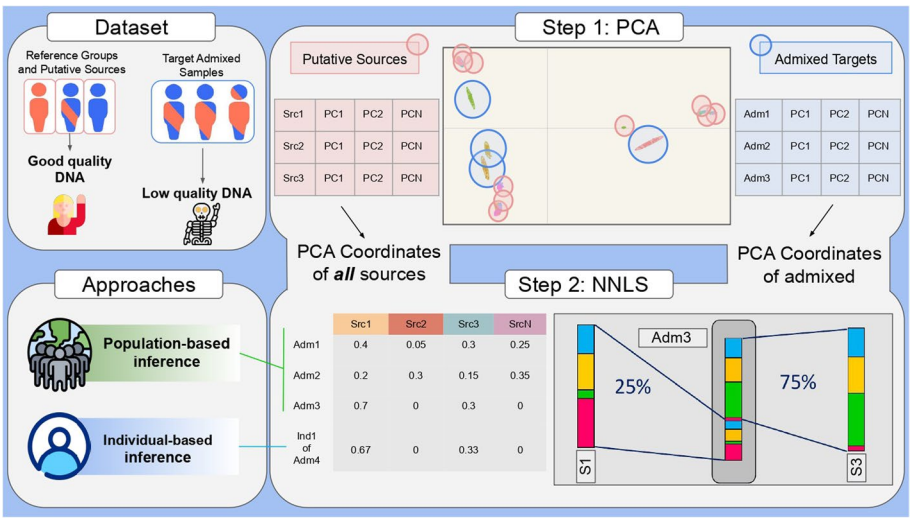

Fig. 1 Schematic representation of PANE workflow. PANE harnesses non-negative least squares using individual or population principal component vectors. PC analysis can be performed using both high-quality genomic data and datasets with missing data, or to accommodate varying degrees of missingness, such as projection or probabilistic PCA

Specifically, we simulated admixture events between groups with different degrees of affinity, from highly divergent to closely related populations, with pairwise F_{st} between populations ranging from 0.01 to 0.23, including bottleneck events, expecting a lower assignment accuracy in cases where the source groups are genetically closer [28]. For each scenario, we tested our approach on the average principal component (PC) coordinates from each admixed group (population-wise approach) and on each admixed individual separately (individual-wise approach).

We initially tested PANE performance considering as putative admixture sources the entire panel of the true sources or their sister groups ("proxy sources"), even though only two (two-way admixture) or three (three-way admixture) sources ("true sources") were used to simulate the admixture event. We then applied the same framework by projecting the admixed groups onto the PC space constructed from the first 10 components, using all the true sources and their respective sister groups (Fig. 1). The number of components was selected after running a preliminary assessment of PANE performance as described in Text S1. Subsequently, using NNLS, we modeled the average PC coordinates across individuals of each admixed group as a mixture of those of all the available sources, considering as sources either the true or the sister groups panel. Standard errors (SE) were estimated using a jackknife approach [2, 29], as described in the "Methods" section.

In this framework, PCA space should be built on high-quality data with a low missingness degree, such as modern sequence data, high coverage, or imputed ancient genotypes [10, 11].

Data and source availability

Our approach to assessing ancestral composition using principal component analysis and NNLS (PANE) is available as an R package at github.com/lm-ut/PANE and in Zenodo https://doi.org/https://doi.org/10.5281/zenodo.14016612 [30, 31]. This tool is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0). To view a copy of this license, visit https://creativecommons.org/licenses/by-nc-nd/4.0/deed.en.

Requests for accessing previously published data used in this work should be directed to the corresponding authors of the publications where they were originally presented [19, 20, 32, 32–42]. The Allen Ancient Data Resource is available at [43, 44] and Raveane et al. [34] at https://capelligroup.wordpress.com/data/.

Results on simulated genotypes with no missingness

PANE performance with true sources

In the population-based approach, for all the 16 simulated admixed populations, PANE successfully assigned the main ancestry components to the true sources that contributed to the admixture event despite the large panel of potential source groups available. The ancestry proportions of the true source groups (Additional file 2: Table S2A) yielded a maximum error of 0.014 and a maximum jackknife standard error (SE) of 0.012 when two sources contributed to the target population (Fig. 2A–C). Minor additional contributions were assigned to other groups but never exceeding 0.004 (Pop 8 in Fig. 2A). In these cases, the additional ancestral component was assigned to groups closely related

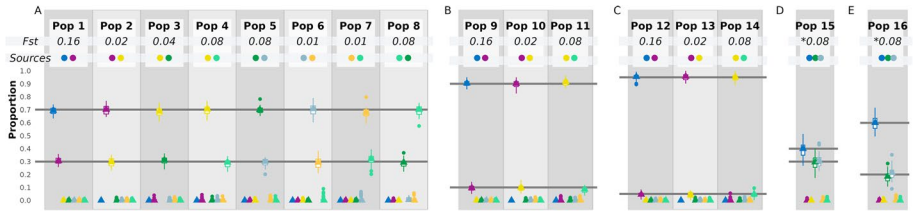

Fig. 2 PANE assignment using true sources for each admixed group; triangles and boxplots show the population and individual ancestry estimation, respectively. On the upper part of the panel the Fst values (minimum values are marked with *) and the Sources that contribute to the admixture event in the simulated populations are shown: A) populations obtained from the combination of two sources with proportions 70-30%; B) admixed populations obtained from the combination of two sources with proportions 90-10% and C) 95-5%. D) Three-way admixed populations generated by combining three sources with 40-30-30% and E) 60-20-20% proportions

to the true source ($F_{st} \leq 0.01$, [28]). In three-way admixed populations (Pop 15 and Pop 16 in Fig. 2D–E), the true sources are always recognized with a maximum error of 0.010 (SE < 0.01).

The ancestry assignment estimation obtained per individual is in Additional file 2: Table S2. 70-30% admixed populations (Pop 1–8) show an average error lower than 0.029 when the admixing sources that split more than 9 KYA (kilo years ago), which increases (0.038) when the simulated population split was less than or equal to 9 KYA (Pop 6 and 7).

Admixed populations with lower source contributions (Pop 9–14) record an average error of a maximum of 0.022. In this case, lower error values are observed for populations with a less recent split (for example Pop 9 and Pop 12). The highest average error in the individual-based analysis is observed in the three-way admixed populations (Additional file 2: Table S2B). An over/underestimation exceeding 0.05 of the assigned contribution to the main sources in the 32% of individuals is recorded. Only one individual has an error larger than 0.1.

PANE performance with proxy sources

We evaluated PANE performance on a PCA where the admixed groups were projected onto the PC space built on all the remaining populations. We then modeled the admixed groups as a mixture of all the proxy sources only. As we knew which simulated proxy group was indeed the sister group of the real admixing source, we calculated the assignment error by considering differences in observed and expected ancestry proportions and whether PANE could indeed select the closest sister group.

In this scenario, for all the 16 tested populations, the proxies of the true sources were recognized without ever assigning even minimal contributions to other populations (Additional file 2: Table S2).

For two-way admixed populations with proportions of 70–30%, the average error is 0.033 (SE < 0.009, Additional file 2: Table S2C). Generally, the error estimates tend to be larger when the admixing source populations (Pop 3, 4, 6, and 7 in Fig. 3) are characterized by a higher genetic similarity due to recent split times and bottleneck events. However, the error never exceeds 0.057 (Pop 4).

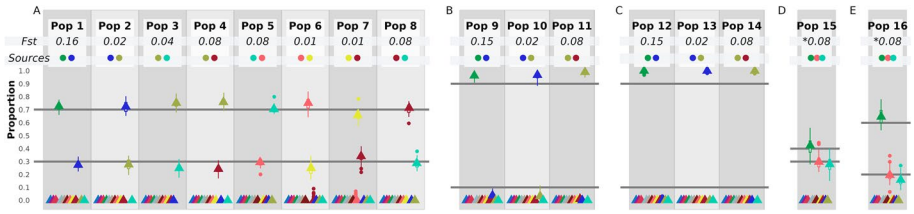

Fig. 3 PANE assignment using proxy sources for each admixed group. Triangles and boxplots show the population and individual ancestry estimation, respectively. On the upper part of the panel the Fst values (minimum values are marked with *) and the proxy Sources that contribute to the admixture event in the simulated populations are shown. A) populations obtained from the combination of two sources with proportions 70-30%; B) admixed populations obtained from the combination of two sources with proportions 90-10% and C) 95-5%. D) Three-admixed populations generated by combining three sources with 40-30-30% and E) 60-20-20% proportions

The PANE accuracy is also robust in the case of three-way admixed populations, with a maximum error of 0.031 (SE < 0.0087).

The overestimation of the major component becomes more important in strongly imbalanced contribution cases. When the contribution of the minor source is 10% (Fig. 3B), the minor contribution is underestimated (Pop 10, 11 in Fig. 3B). For populations where the minor source contributed 5%, PANE completely misses the minor source contribution and assigns the total of the ancestral component to the main source (SE < 0.007) (Fig. 3C).

In individual-based inferences (Additional file 2: Table S2D), PANE correctly assigns ancestral proportions in the two-ways 70–30% admixed individuals. The estimates obtained for the 50 individuals within each group are characterized by a maximum average error of 0.0582 (Pop 6, Fig. 3A).

For the admixed populations with minor source contributions of 10% and 5%, the contributions of the minor sources are underestimated or completely missed. In detail, for populations 9 and 10, the average estimated minor contribution is 0.038 and 0.034, with 29% of individuals showing less than 2%. On the other hand, in population 11, the average minor contribution is 0.016, with 64% of individuals showing less than 2%. No relevant contribution from other sources is recorded. For populations 12, 13, and 14, 85% of individuals are modeled as unadmixed, with the remaining individuals showing an average minor contribution of 1.3%.

For the three-way admixed populations, PANE is always able to recognize the correct sources and assigns them the right proportions with a maximum error of 0.05 (Pop 16, Fig. 3E) since for some individuals there is a slight overestimation of the main source at the expense of one of the other two minor sources.

Results on pseudo-haploid simulated data

We tested PANE using pseudo-haploid samples, simulated by introducing different degrees of missing genotypes (up to 50%) and pseudo-haplodized (see the "Methods" section) mirroring the fragmentary nature of data commonly adopted in aDNA studies. We tested PANE on a PCA where the PC space is built by the diploid genomes of the proxy sources, onto which we projected the pseudo-haploid genotypes of the admixed

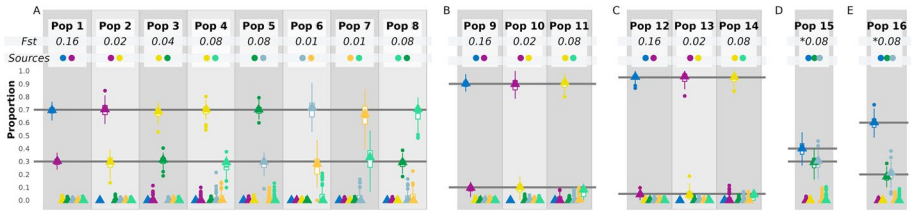

Fig. 4 PANE assignment using pseudo-haploid simulated data and modeling each admixed group as a mixture of all the available proxy sources. Triangles and boxplots show the population and individual ancestry estimation, respectively. n the upper part of the panel the Fst values (minimum values are marked with *) and the proxy Sources that contribute to the admixture event in the simulated populations are shown. A) populations obtained from the combination of two sources with proportions 70-30%; B) admixed populations obtained from the combination of two sources with proportions 90-10% and C) 95-5%. D) Three-admixed populations generated by combining three sources with 40-30-30% and E) 60-20-20% proportions

groups and all possible true sources. In this scenario, we tested whether ASAP could model the admixed groups with the pseudo-haploid true sources.

As shown by Additional file 2: Table S3, PANE correctly detects the closest admixture sources even in a large panel of putative donors, despite the target and source samples being pseudo-haploid and containing missing genotypes (Fig. 4). Indeed, the average assignment error is at a maximum of 0.033. In this case, PANE always identifies the true sources and assigns a marginal additional component to other sources (maximum 0.004). Furthermore, the jackknife SE is also generally low, with a maximum of 0.017 (Additional file 2: Table S3) seen in the admixed population whose sources split more recently (7.5 KYA). Even when single samples are targeted, the true sources are generally recognized and the major source ancestry assignments show an average error of 0.039. Despite the low average error, the maximum per sample error can reach 0.248, caused by the misassignment to the most closely related group to the sources ($F_{st}=0.01$, [28]).

PANE performance with limited reference genetic variation availability

We tested PANE in a scenario where only the proxy, but not the true sources of the admixture, were available. The rationale behind this analysis is to mimic the lack of true mixing sources when exploring aDNA datasets while leveraging the availability of diploid genomes to build the PC space. In this scenario, we subsetted the source panel to only two putative proxy sources to model the admixed groups that underwent a two-way admixture event, or three putative sources for the three-way admixture targets. PANE can be used with either a large or a small source panel. However, in a real-case scenario, selecting a limited number of the source groups requires an initial hypothesis of the demographic history of the admixed group to select the optimal proxy sources.

In this test, we projected onto the PC space the pseudo-haploid genotypes of (i) the target admixed group and (ii) the closest proxy sources of each true source, two proxy sources in case of a two-way admixture, and three for the three-way admixture. The PC space was built with the diploid genomes of the remaining proxy sources. We modeled the target admixed group's relative admixture proportions given the projected proxy sources, relying on a limited donor panel of two or three groups.

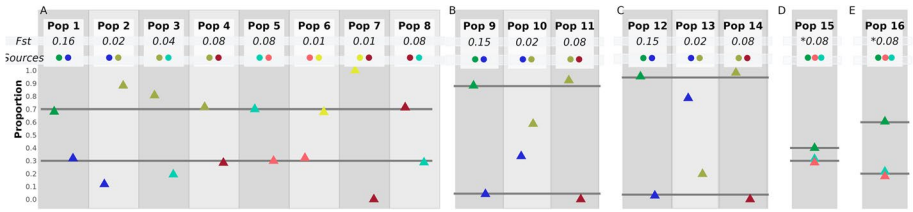

Fig. 5 PANE performance with limited reference genetic variation availability. Only population-based inferences are shown.n the upper part of the panel the Fst values (minimum values are marked with *) and the proxy Sources that contribute to the admixture event in the simulated populations are shown. A) populations obtained from the combination of two sources with proportions 70-30%; B) admixed populations obtained from the combination of two sources with proportions 90-10% and C) 95-5%. D) Three-admixed populations generated by combining three sources with 40-30-30% and E) 60-20-20% proportions

Given the large error in individual analysis, mostly due to the lack of a proper reference dataset, we focused on the population-based approach (Fig. 5). In such a scenario (complete results available in Additional file 2: Table S4), error estimates are lower than 0.043 for all groups whose sources diverged more than 24 KYA (Pop 1, 4, 5, 8, 9, 11, 12, 14). For the only group whose sources split 24 KYA (Pop 3), the error increases to 0.11. In contrast, for all the other groups with closer sources, the error estimates range between 0.16 and 0.58, with jackknife SE estimation following the same pattern (Additional file 2: Table S4).

Assessing the effect of strong genetic drift on PANE inference

We assessed the effect of strong genetic drift on admixture proportions estimated by PANE, simulating diploid genotypes of admixed populations that witnessed a strong reduction immediately after the admixture and continued to evolve for different time periods. Our results show that, on average, genetic drift reduces the reliability of inference with a magnitude proportional to the time that occurred after the split (Additional file 1: Fig. S2). In detail, PANE carried a maximum 0.15 error in individual-based analyses, and a 0.12 error in the population average when harnessing a population admixed 500 generations ago.

Benchmarking PANE versus existing global ancestry inference tools

We compared PANE with qpAdm, Rye, and Unlinked-ChromoPainter NNLS, which harness *f4*-statistics, PCA, and a modified Li and Stephens model with infinite recombination between SNPs for the ancestry composition inference, respectively [13, 14, 27, 45]. We compared the accuracy in estimating the ancestral proportions of the four approaches using the pseudo-haploid genotypes of both the target admixed samples and the true sources of the admixture. Our method behaves similarly to the others (Fig. 6A–C); the correlation of ancestry assignments (Fig. 6D) of PANE, qpAdm, and Rye is higher than 0.95 (PANE vs qpAdm $R^2 = 0.968$, p-value $< 10e-6$; PANE vs Rye $R^2 = 0.998$, p-value $< 10e-6$, PANE vs CP $R^2 = 0.985$). Among the four harnessed algorithms, qpAdm is characterized by the highest average error, and all four approaches show a lower accuracy for the admixed populations characterized by a subcontinental admixture, in which the two admixing sources are generically close ($F_{st} = 0.01$) [28]; see Additional file 2: Table S5).

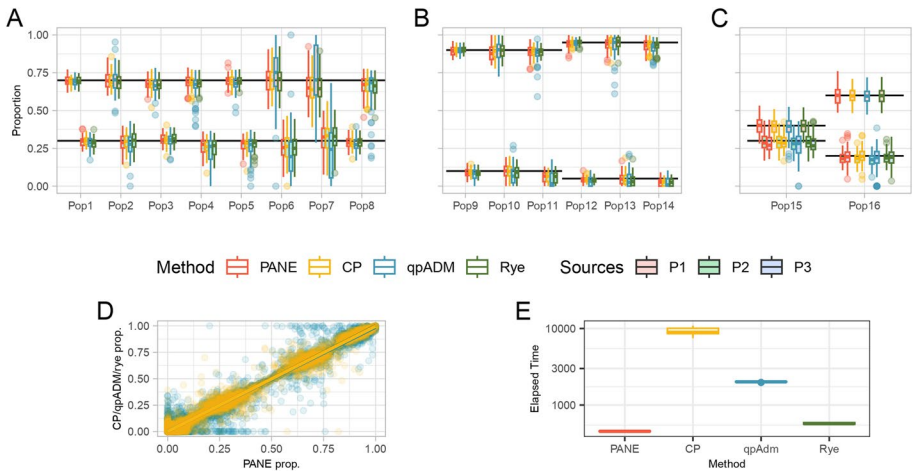

Fig. 6 Comparison between PANE (red), ChromoPainter NNLS (CP, yellow), qpAdm (blue), and Rye (green) modeling the ancestral proportions of pseudo-haploid admixed populations given a set of pseudo-haploid sources. **A** Ancestry proportions for simulated populations with 70–30% sources' contribution. **B** Ancestry proportions for simulated populations with 90–10% (Pop 9–11) and 95–5% (Pop 12–14) sources' contribution. **C** Ancestry proportions for three-way admixture simulated populations with 40–30–30% (Pop 15) and 60–20–20% (Pop 16) sources' contribution. **D** Correlation of the ancestry proportion assignment between CP, Rye, and qpAdm on the *y*-axis and PANE on the *x*-axis, and **E** computational time for a subset of 100 individuals

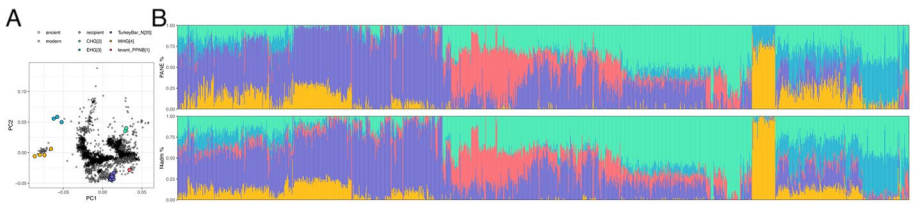

Fig. 7 Ancestry inference using PANE on the aDNA dataset. **A** The PCA used as input by PANE. **B** Ancestry proportions for 1350 ancient individuals (*x*-axis ordered by *k*-mean cluster numbers computed on PANE inferred proportions): the upper panel has been estimated using PANE, while the lower panel shows estimates reported in the original publication using F4admix [46]

We also compared the computational speed of each framework (Fig. 6E) replicating (10 iterations) the ancestry inference for the same set of 100 individuals. For PANE and Rye, we included the PC time (10 analysis), while for qpAdm we took into consideration the estimation of the *f2*. ASAP outperforms the other methods: PANE computational time stands at 454 s (SD = 5 s), while Rye reaches 575 s (SD = 14.1 s), qpAdm 2011s (SD = 22.399 s), and ChromoPainter 156 min and 23 s (9383 s, SD = 1148 s).

PANE performance on real data

We tested PANE on real data using a dataset of different ancient Eurasian populations [32, 46]. We projected 1380 ancient individuals into the first 10 principal components inferred using 1668 present-day individuals. Following Lazaridis et al. [32, 46], we applied PANE on 1350 target individuals, using five putative sources: Western Hunter-Gatherers (WHG), Caucasus Hunter-Gatherers (CHG), Eastern Hunter-Gatherers (EHG), Anatolia Neolithic (AN), and Levant Neolithic.

The ancestry compositions captured by PANE on real data show a significant correlation ($R = 0.92$, $p < 0.0001$) with F4admix results obtained in the original paper [46], confirming the reliability of PANE in real-world scenarios (Fig. 7, Suppl. Table 6).

Moreover, we explored the individual ancestral composition of specific geographic locations in different time transects, as in Lazaridis [46]. This enabled us to pinpoint the emergence of ancestral influences across different geographical regions and prehistoric periods. First, we examined the Anatolian region and confirmed an increase in Caucasus/Levantine ancestries of around 3000 KYA, accompanied by a subsequent reduction in local Anatolian ancestry (Additional file 1: Fig. S3). Then, we confirmed the introduction of CHG-related ancestry into Steppe populations around 5000 KYA, alongside the absence of Anatolian ancestry in this region prior to 3000 KYA. We did not observe an increase in Levantine PPN ancestry, suggesting that most Eastern influence is associated with AN ancestry. Our approach corroborates again the complex genetic composition observed within the Yamnaya cluster, characterized by consistent CHG admixture (Additional file 1: Fig. S4).

PANE analysis further identified a less pronounced overrepresentation of CHG ancestry than EHG ancestry in Aegean Bronze Age populations. This observation suggests significant gene flow occurring after the Neolithic period, particularly during the Early Bronze Age, across the Aegean and Balkan Peninsula regions [33–35, 47] (Additional file 1: Fig. S5). Similar trends were also observed in Italy, where Iron Age Southern Italian samples exhibited the highest frequency of Caucasus hunter-gatherer ancestry, found almost absent in Central Italian Etruscans (Additional file 1: Fig. S6) [48].

Although overall there is a high correlation between the two inferences, we observed 273 (out of 6750) highly discordant estimates (HDE), in which the ancestral proportion difference exceeds 0.2.

1. When considering Western Hunter-Gatherer ancestry, we observed a correlation $R = 0.9$ (p-value $< 1e^{-4}$) and 53 HDEs. Many of them include hunter-gatherers from Serbia and Romania, which are modeled by PANE as approximately 70% WHG with the remaining ancestry mainly assigned to EHG, while 90% WHG and 10% EHG were estimated by Lazaridis et al. [46]. These samples were first published by [42], who described them as a combination of WHG and EHG using qpAdm (although the estimates are associated with very low p-values) and D-statistics.
2. Concerning the ancestry of Turkey Barcin Neolithic individuals, commonly known as AN [46], we observed a correlation $R = 0.95$ ($p < 1e-4$) and 59 HDEs. Nevertheless, a few individuals exhibit a substantial discrepancy in AN ancestry proportion between the two compared methods, making it challenging to determine which of the two approaches has the highest performance. For example, PANE estimates higher AN ancestry for some Mycenaean individuals [19] while F4admix gives higher Anatolian ancestry for an Iron Age individual from Lebanon [37] (Additional file 2: Table S6). Both methods can be inaccurate in some cases, as shown by comparisons with previous studies.

3. Concerning the Iran Neolithic/CHG ancestry, we observed a correlation R of 0.95 (p-value $< 1e-4$) and 47 HDEs. Most (20) are related to populations from Chalcolithic and Bronze/Iron Age Near East (Iran and Lebanon) individuals. For example, seven Bronze Age individuals from Shahr I Sokhta are modeled as having a substantially smaller Iran Neolithic/CHG ancestry for PANE estimations (mean = 0.65) compared to F4admix (mean = 0.94). In Narasimhan et al. [20], when Shahr I Sokhta individuals are modeled using qpAdm, they show an average of 0.66 IN/CHG (SD = 0.05).
4. In the case of EHG, we noted 40 HDEs and an R-value of 0.86 (p-value $< 1e-4$). As for WHG, most of the HDEs are related to Hunter-Gatherers from the Iron Gates regions of Serbia and Romania, for which PANE estimates a higher proportion of EHG when compared to Lazaridis et al. [46]. Furthermore, in four Bell Beaker individuals from Germany, France, and England, PANE estimates a very low proportion of such ancestry.
5. We then observed 74 HDEs and a correlation R of 0.88 (p-value $< 1^{e-4}$) for Levant Neolithic ancestry. Most of the HDEs are related to ancient individuals from the Near East, for which estimates of Levant PPN are always higher than those inferred by Lazaridis et al. [46]. These results align with previous estimates on the same samples. For example, for the individual I3832, which was modeled as 0.58 Levant PPN and 0.42 Iran Chalcolithic using LINADMIX in its original publication [38], PANE estimated the Levant PPN proportion at 0.77, which was 0.38 when using F4admix [46]. A possible explanation for this discrepancy is related to the fact that in [46], the same individual is modeled to have approximately ~ 0.2 related to AN. Similarly, PANE ancestral composition for individuals from Roman and Iron Age from Lebanon are in line with previous DyStruct inferences [37]. Furthermore, F4admix [19] estimated a substantial proportion of Levant PPN ancestry in two Greek and one Italian Bronze Age samples, in contrast with a series of findings on the same or similar individuals. All these samples are characterized by a missingness rate higher than 40% (I9006), suggesting that using PANE on projected PCs might be less biased than F4admix estimates.

We also used PANE to test the robustness of the support behind the hypothesis that the WHG contributions of British farmers came mostly from continental WHG rather than local British WHG [49]. Therefore, we ran PANE with the same settings as Brace et al. [49], modeling European Neolithic samples as a mixture of WHG and AN. We confirmed their results by finding WHG proportions in Iberian Early Neolithic samples similar to those in British Neolithic samples, suggesting a common WHG source (Additional file 2: Table S7). We then tested models with pairwise WHG individuals as possible sources and a single AN population (see the "Methods" section). We found that both Iberian and British samples consistently preferred Bichon or Villabruna-associated samples as WHG source, indicating their close relationship to the true source and confirming a shared origin in this cluster (Additional file 2: Table S8) [39]. This case illustrates how PANE can be a powerful tool to test and refine hypotheses in the ancient genomic field.

To assess the reliability of our methods in a different region, we leveraged the PCA of [20] to assess PANE performance on South-East Asian populations. In doing so, we harnessed the available 9 PCs modeling 15 groups as a combination of AN, Ganj Dareh Neolithic ancestry, West Siberian hunter-gatherer–related (WSHG), and Andamanese hunter-gatherer–related (AHG) and compared to the qpAdm results reported in Narasimhan et al. [20]. Our results (Additional file 2: Table S9) show that there is a high similarity in estimated ancestry proportions, with the only exception of AN and Ganj Dareh Neolithic which are under and overestimated, respectively.

Furthermore, we used PANE to model modern humans as a combination of Mbuti, Neanderthal, and Denisova using the first two PCs (Additional file 1: Fig. S7A). In doing so, we were able to detect the virtual absence of archaic contribution in Western Africa, and the exclusive contribution of Denisovan ancestry in Papuans [50] (Additional file 1: Fig. S7B).

Discussion

We present a global ancestry exploration approach, based on PCA and NNLS, that allows an accurate estimation of ancestry proportions in admixed groups or single samples. The approach leverages how the location of samples on the PC space can be related to the mean time of coalescence between pairs of samples [51], and to the recent observation that PC vectors are strongly related to *f4* metrics [25]. Specifically, in the case of an admixture event, samples will fall along a gradient and their putative admixture sources will be placed at the ends of the gradient [22, 51]. Our approach exploits the relative coordinates of the admixed samples and the ones of the putative sources in the PC space and summarizes the ancestry proportions of the targets through NNLS. Therefore, the groups selected to build the PC space need to describe the genetic differences between the populations of interest: adding an extreme outgroup, for example, will minimize the PC coordinates' differences between the target and proxy source set and limit the efficiency of PANE.

An advantage of the method is that it can leverage the entire PC space, allowing a large panel of donors to model a given target admixed group or multiple parallel analyses in case several different target groups and their relative proxy sources are analyzed.

We demonstrated that the approach is highly accurate in most of the scenarios tested. Regardless of the availability of true or proxy admixture sources, our approach correctly assigned ancestral proportions, with a low associated error. Moreover, in the rare cases of error, PANE appointed a group closely related ($F_{st} = 0.01$) to the source. Furthermore, our results demonstrate that ASAP reliably detects admixture contributions of 10% or more, while contributions as low as 5% are often missed when using proxy sources. In contrast, PANE can properly identify even a 5% contribution from the minor source when true sources are used, emphasizing that the splitting time between true sources and proxies (3 KYA) significantly impacts detection accuracy. The method tends to overestimate the major component in cases of imbalanced contributions. Although PANE performs well with balanced sources, it faces challenges with minor sources and imbalanced admixtures. These findings underscore the necessity for accurate reference panels and further exploration of the detection limits for small admixture proportions.

More importantly, PANE performs well even when pseudo-haploid data with missing variants are analyzed, with a maximum assignment error of 3%. Specifically, when the admixture sources have a split time > 24 KYA, PANE error estimates are lower than 0.014. On the other hand, for closely related admixture sources, the per-sample misassignment can reach 20% when multiple, closely related putative sources are available.

Compared to other global ancestry assignment tools, the approach is faster in terms of runtime while being as accurate (ChromoPainter) or more accurate than other tools (qpAdm) and, most importantly, provides ancestry estimates based on a straightforward formulation of user-defined ancestry sources with no need for in- or out-groups.

When tested on a real dataset of ancient and modern European and Asian genotypes, PANE confirmed the trends in ancestry composition observed in previous research, providing relevant information on the complex scenario of the continent. Notably, it estimated significant gene flows after the Neolithic period in Aegean Bronze Age Eurasian populations and confirmed previous findings about the shared origin of WHG ancestry in British and Iberian farmers. Furthermore, our results on a PCA built with South-East Asian individuals demonstrate that PANE can be used in different world regions. Furthermore, although this property should be thoroughly tested with extensive simulation and comparisons, PANE can infer the presence of archaic introgression in human populations.

Our approach relies on the assumption that the target group is indeed admixed. In the PCs, the target group might fall within a given cline as a result of a demographic scenario different from admixture (i.e., isolation by distance, or genetic drift) [51, 52]. Given that the PCA is the backbone of our tool, the accuracy of the assignment will be affected by any kind of bias that distorts the PC space (sample size differences between populations, ascertainment bias, linkage disequilibrium between SNPs, missing data). Furthermore, it is important to highlight that "PCA cannot distinguish between alternative models that have the same effect on mean coalescent time": thus, events, such as admixture, recent drift, or isolation by distance, might all produce similar projections [51]. It follows that PANE cannot be used as a formal test for admixture, but rather an exploratory tool, and integrated with different analyses, such as formal tests for admixture (such as F3 admixture statistics or LD decay methods [5, 26]), to test the admixture event further.

Conclusions

PANE is a powerful and flexible tool for analyzing ancestral proportions in individuals and mixed populations, offering a faster and often more precise alternative to other available methods. Thanks to the joint use of PCA and NNLS, the method can handle various situations, including scenarios with incomplete data. Although ancestral inferences can sometimes be biased by extreme genetic drift, PANE is still a useful tool for characterizing large-scale genomic datasets and can be widely used to complement more specific analyses, significantly contributing to the understanding of the genetic complexity of human populations.

Methods

Datasets

Simulated dataset

Genotype data with no missingness We used a simulated genotype dataset from Molinaro et al. composed of 13 simulated demes with different population sizes and split times ranging from 250 to 4000 generations, to represent a simplified scenario for current European-like (EUR 1–3), East Asian-like (ASN 1–3), and African-like (AFR 1–7) groups and 7 sister groups characterized by a split time from their closest population of 100 generations [28]. The data simulation was carried out with mspms and following a modified Van Dorp et al. model [53, 54]. The initial dataset consisted of eight admixed groups, obtained by combining pairs of the simulated Ghost populations (GST), all with ancestry proportions of 70–30%. The pairs of admixing GST were selected in order to cover a broad spectrum of split times. Specifically, we simulated admixture groups whose sources split time span from 75 to 9 KYA, six sources shared a bottleneck event and for three of these, we simulated an additional one. The initial set also comprised one admixed group characterized by a three-way admixture with the proportions of 60–20–20%, with African-like, European-like, and Asian-like ancestries, respectively.

We simulated an additional three-way admixture group, using the same highly divergent sources as above, but different ancestry proportions, namely 40% for the African-like ancestry and 20% for both the European and Asian-like ancestries. To test models with strongly imbalanced ancestry proportions, we also simulated three groups with 90–10% and three groups with 95–05% ancestry proportions. In this case, as well, we chose the admixture sources (GST) to cover a broad spectrum of split times. All admixture simulations were carried out with admix-simu (https://github.com/williamslab/admix-simu), creating 50 individuals per population, using a constant recombination rate (1.25×10^{-8}) and admixture time of 100 generations [28]. We obtained a simulated dataset of 4,745,025 SNPs, 20 non-admixed, and 16 admixed groups. After filtering for minor allele frequency with PLINK (maf 0.01), the final dataset comprised 284,249 SNPs [55]. PC analyses were performed on the final dataset, projecting the admixed target samples on the scaffold built from the non-admixed groups. The simulated GST populations acted as "true sources," while their sister groups acted as "proxy sources" for all our tests.

Ancient (pseudo-haploid) data To mimic the data quality of ancient DNA, we manipulated the simulated set by introducing both missing data as well as using pseudo-haploid genotypes. In each population, we introduced a variable missing rate (from 10 to 50%) in randomly selected positions, so every 10 individuals would be characterized by 10, 20, 30, 40, or 50% of missing data. Secondly, we created pseudo-haploid genotypes by randomly selecting at each locus one allele and assigning it as a homozygous genotype, eventually obtaining for each simulated group 100 pseudo-haploid genotypes from the original 50 diploid individuals. The missingness proportions were maintained after pruning.

PCA was performed after filtering for minor allele frequency (maf 0.01). For the pseudo-haploid datasets containing missing data, pruning was also performed (PLINK v1.9 indep-pairwise 50 10 0.1) [55].

After the filtering, the bim file of the modern simulated dataset contains 284,249 SNPs, the one in which only the true sources are pseudo-haploid has 135,211 SNPs, while the one in which the proxy sources are also pseudo-haploid has roughly 100,000 SNPs.

Real modern and ancient dataset

We downloaded the 1240 K + HO dataset (version V52.2, https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data) in EIGENSTRAT format [43, 44]. Such dataset includes present-day and ancient DNA data converted in PLINK format using convertf [22].

Starting from the.anno file and following the Aneli et al. method [48], we created a list of ancient and modern samples to keep from the 1240 K + HO dataset.

In particular, only the ancient ones (those that in the "Full date "column did not have the string "present") coming from Western Eurasian countries (latitudes higher than 22 and longitudes between − 15 and 60) and the present-day Mbuti individuals from Congo (string "Mbuti" in the "Genetic.ID" column) [5, 56] were selected. Then, we removed individuals that in the "Group.ID" column have the string "Ignore." Finally, we kept only those whose "Assessment" column contained the string "PASS." After removing any duplicates, we created a preliminary plink file with only 1240 K + HO ancient samples to whom we added other ancient samples taken from other published datasets [32, 46, 48, 57, 58].

For the modern samples, we selected individuals coming from Western Eurasian countries with latitudes higher than 22 and longitudes between − 15 and 60, but removing those coming from Uzbekistan, Kazakhstan, Algeria, Morocco, Tunisia, and Libya, as well as some populations from Russia and others showing "Ignore" within "Group.ID." Then, we selected the samples flagged as "PASS" in the "Assessment" column. In this way, we created a preliminary plink file with only the 1240 K + HO modern individuals to which we merged other modern-day samples taken from the Raveane et al. dataset [34].

From this modern dataset, we extracted only the autosome chromosome SNPs and excluded those monomorphic (–maf 0.00001) and with more than 5% missing data (–geno 0.05) using PLINK [55].

Then, we extracted the bulk of variants built on our modern dataset from the ancient dataset and finally merged all with PLINK1.9 excluding the ancient samples with less than 20,000 SNPs ($N = 1381$). To assess the relatedness between individuals and exclude close relatives, we calculated the kinship coefficient, pi-hat, which represents the probability that two randomly selected alleles at the same locus are identical by descent (IBD) between two samples. Using PLINK 1.9 with the options –genome –min 0.35, we selected pairs of individuals with a pi-hat value of 0.35 or higher, retaining a total of 7312 unrelated individuals.

We then filtered out samples with duplicated IDs retaining the one with a higher number of SNPs. We carried out a principal component analysis (PCA) using smartpca, through which we filtered out outliers obtaining a dataset of 6408 samples. Our final real dataset, containing 4740 ancient and 1668 modern samples with 206,363 SNPs, was converted to the EIGENSTRAT format using convertf.

Principal component analysis (PCA)

PCA was performed using smartpca (version 16000) available in the EIGENSOFT package [5].

The admixed (or target) populations were always projected, regardless of the dataset used. In the case of datasets with pseudo-haploid or ancient individuals, we projected all samples' genotypes onto the principal components inferred from the diploid/modern individuals using the lsqproject: YES option. For each analysis on the simulated and real genotype datasets, we run 10 PCs. Furthermore, we performed the same analysis using different numbers of PCs, in order to assess the performance of PANE (see Text S1).

Non-negative least squares (NNLS)

To perform the non-negative least squares, we used the NNLS function, as described in [1, 4, 7], which is an adaptation of the Lawson–Hanson NNLS implementation of non-negative least squares [59] available in the statistical software package R 3.5.1 [60].

We applied NNLS both population-wise and individual-wise. To estimate the NNLS population-wise, we estimated the average of each of the population PCs and then applied NNLS on the resulting vector. On the other hand, to estimate the NNLS individual-wise, the PCs of each individual were maintained as separate vectors.

Our approach is equivalent to that of Leslie et al. [7] and Montinaro et al. [2], which was used on chromosome painting profiles with the scope of minimizing potential biases due to incomplete lineage sorting.

In detail, PANE uses NNLS on a PCA matrix rather than a chromosome painting matrix, with the goal of minimizing the squared differences between the observed data (individual or population data points in the PC space) and the admixture model's predictions, constrained by non-negative coefficients:

$$min(x)||Px - y||_2^2; x \leq 0$$

where P is the PCA matrix of the sources and y is the vector of the tested individual/populations. In other words, we are minimizing the distance between predicted and observed coordinates in the PC space.

Error values reported in the text were calculated as the absolute average difference between the observed and the expected proportion assignment. We used a block jackknife approach to resample our set and estimate the standard errors. Given that the simulated set consisted of only chromosome 1, we could not use chromosomes as blocks, as usually it is done when the entire genome is available. We thus estimated the number of SNPs available after filtering and divided them into 20 blocks. For each resampling step, we removed one of such blocks and performed PCA on the remaining ones.

Standard errors were estimated on chromosome-based jackknife replications [61].

Testing post-admixture genetic drift effect on PANE estimates

We tested the effect of substantial drift, by simulating several two-way admixture events with variable admixture generation times: 50, 100, 150, 200, 250, 300, 350, 400, 450, and 500. Each simulated admixed group underwent a bottleneck, decreasing its effective population size from 10,000 to 1000, 50 generations after the admixture event had taken place. For the group that underwent the admixture event at 50 generations, we set the bottleneck event at 1 generation. In each analysis, we simulated 0.7–0.3 admixture contributions from the two sources, a 250-Mb sequence, a constant recombination rate of 1.25×10^{-8}, and a mutation rate of 1×10^{-8}. Eventually, we sampled 50 diploid individuals from the admixture groups and from the two groups contributing to the admixture event. We performed PC analyses by projecting the admixed group on the scaffold built by the sources. Simulations were carried out entirely with Msprime [53, 62].

qpAdm

To validate the results obtained from PANE, we performed the most widely used approach to assess the ancestry components and the relative proportions of the admixed population: qpAdm programs in the ADMIXTOOLS2 package [63], using precomputed f2 statistics. For each admixed individual, we tested as "left" populations all the possible true sources and used all the others as right populations. Subsequently, we selected the inference characterized by the largest p-value, irrespectively to their significance. Although there are many ways to harness qpAdm to obtain more reliable results, we decided to use a strategy comparable to the other tools harnessed here.

Rye

We applied Rye [27] converting the PCA output obtained by smartpca using a custom R script. We performed five different rounds using the first ten PCs.

ChromoPainter

ChromoPainter (CP) [64] was applied using the unlinked (-u) model, where, for each SNP in the target, we assign a score of $1/K$ to each reference haplotype that carries the same allele, where K is the total number of reference haplotypes that carry the same allele.

Analysis of Eurasian ancient and modern genotype data

We carried out a PCA computing 10 principal components (PCs) per each individual in our final dataset projecting ancient samples on the top of present-day genome variability ($N = 1668$). We used smartpca version 16,000 with autoshrink lsqproject options for this analysis. Subsequently, we selected ancient individuals previously analyzed in Lazaridis et al. [46] and filtered out samples with less than 180 K missing SNPs. This resulted in a dataset comprising 1380 ancient targets, of which 30 were chosen as sources for PANE. The selection of sources samples was based on the five main ancestral sources identified in Lazaridis et al. [46], namely Western Hunter-Gatherers (WHG), Eastern Hunter-Gatherers (EHG), Caucasus Hunter-Gatherers (CHG), Anatolian Neolithic, and Levant Neolithic. PANE was run using 10 PCs, and the estimates were correlated with F4admix results using all samples combined, as well as stratified by different ancestral sources.

Pearson correlation analysis was performed using the ggpubr library in R. To explore ancestry over time, we visualized single ancestry trends by either selecting individuals as indicated in the publication or by visually inspecting populations present in [46] figures.

To assess the reliability of our methods in a different region, we leveraged the PCA of [20] to assess PANE performance on South-East Asian populations. In doing so, we harnessed the available 9 PCs modeling 15 groups as a combination of Anatolia_N, Ganj_Dareh_N, WSHG (West Siberian hunter-gatherer–related), and AHG (Andamanese hunter-gatherer–related) and compared to the qpAdm results reported in [20].

Inference of archaic admixture in modern humans

To test the applicability of PANE in archaic admixture inference, we have projected 6441 modern and ancient genomes onto the PC space built using 182,314 SNPs from Chimp, Neanderthal (Vindija), and Denisova [40, 41]. PCA was carried out using smartpca [22] with lsqproject and autoshrink options set as NO. Next, we used PANE to model different individuals as a combination of Mbuti, Neanderthal, and Denisova, using the first 2 PCs.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03491-z.

Additional file 1: Supplementary Figure S1-S7. Fig. S1: Overview of the simulated scenario. We used a simulated genotype dataset composed by 13 simulated demes with different population sizes and split times ranging from 250 to 4,000 generations, to represent a simplified scenario for current European-like (EUR 1-3), East Asian-like (ASN 1-3) and African-like (AFR 1-7) groups and 7 sister groups characterized by a split time from their closest population of 100 generations [28]. Fig. S2: Effect of drift on PANE estimations, results showing the difference between the observed and expected major ancestry assignment by PANE (y-axis) across 10 admixture generations (x-axis), ranging from 50 to 500. For each admixture generation, the dots represent each simulated sample (50 per each group) and the boxplot shows the group distribution. The group average error increases with the number of admixture generations elapsed since the admixture event, indicating that substantial drift does have an impact on the ancestry assignment accuracy. Notably, in this scenario PANE carried a maximum 0.15 error in individual-based analyses, and a 0.12 error in the population average. Fig. S3: Ancestry proportion patterns in Anatolia. Following Lazaridis et al. 2022 [32, 46], we estimated ancestry composition in different Eurasian broad regions. The left and right panel shows the ancestry composition using PANE and F4admix (as estimated by [32, 46]). Fig. S4: Ancestry proportion patterns in the Steppe region. Following Lazaridis et al. 2022 [32, 46], we estimated ancestry composition in different Eurasian broad regions. The left and right panel shows the ancestry composition using PANE and F4admix (as estimated by [32, 46]). Fig. S5: Ancestry proportion patterns in Southern East Europe. Following Lazaridis et al. 2022 [32, 46], we estimated ancestry in different Eurasian broad regions. The left and right panel shows the ancestry composition using PANE and F4admix (as estimated by [32, 46]). Fig. S6: Ancestry proportion patterns in Italy. Following Lazaridis et al. 2022 [32, 46], we estimated ancestry in different Eurasian broad regions. The left and right panel shows the ancestry composition using PANE and F4admix (as estimated by [32, 46]). Fig. S7: Testing hybridization with other species. (A) Principal Component Analysis (PCA) used as input for the PANE framework. A total of 6,441 modern and ancient genomes are projected onto the genetic variability of Neanderthal, Denisovan, and chimpanzee genomes, using 182,314 SNPs for the projection. (B) PANE results showing average Denisovan and Neanderthal ancestry proportions (± 1 SD) in modern human populations, with populations sorted by Neanderthal ancestry levels.

Additional file 2: Supplementary Table S1-S9. Table S1: List of simulated populations. Column "% admix" indicates the admixed proportions for both the two-way (from Pop1 to Pop 14) and three-ways (Pop15-16). In Source 1, Source 2, and Source 3 columns are listed the sources used to create the admixed individuals, while Proxy 1, Proxy 2 and Proxy 3 columns indicate the populations closer to the real sources. The pairs of admixing Ghosts were selected to cover a broad spectrum of split times ("Minimum split time (KYA)" column), allowing us a deeper evaluation of the framework performance. Fixation index (Fst) for real sources and Proxies are shown in"Sources Fst" and "Proxies Fst" columns respectively. Table S2: PANE performances on simulated genotypes with no missingness. Results are listed in tables separated by methodology (Population-based and individual-based) and donors (real sources or closer populations) used. Each table shows PANE results, the errors (calculated as the difference from the PANE results and the expected proportion), and the standard errors (obtained from the jackknife approach - see M&M). Table S3: PANE performances with pseudo-haploid samples. Results are listed in tables separated by used methodology (Population-based and individual-based). Each table shows PANE results, the errors (calculated as the difference from the PANE results and the expected proportion), and the standard errors (obtained from jackknife approach- see M&M). Table S4: PANE performances with pseudo-haploid samples and limited reference genetic variation availability. Results are listed in tables separated by used methodology (Population-based and individual-based). Each table

shows PANE results, the errors (calculated as the difference from the PANE results and the expected proportion), and the standard errors (obtained from the jackknife approach - see M&M). Table S5: Comparison between PANE (columns 3:9), qpAdm (columns 10:16), Rye (columns 17:23), and ChromoPainter NNLS (columns 24:30) modeling the ancestral proportions of pseudo-haploid admixed populations given a set of pseudo-haploid sources. Individual-based approach was applied. Table S6: Ancestry compositions captured on real data by PANE compared with those obtained with F4admix in Lazaridis et al., 2022 [32, 46]. Table S7: PANE composition with the same settings as Brace et al. 2019 [49], modeling European Neolithic samples as a mixture of WHG and AN. Table S8: Models with pairwise WHG individuals as possible sources and a single Anatolian Neolithic population. Table S9: Assessing the performance of PANE on a South-East Asian dataset. We leveraged the PCA of Narasimhan et al. 2019 [20] to assess PANE performance on South-East Asian populations. In doing so, we harnessed the available 9 PCs modeling 15 groups as a combination of Anatolia_N, Ganj_Dareh_N, WSHG (West Siberian hunter-gatherer–related), and AHG (Andamanese hunter-gatherer–related) and compared to the qpAdm results reported in Narasimhan et al. 2019 [20].

Additional file 3: Supplementary Text S1. Text S1: Evaluation of PANE performance in relation to the number of PCs.

Acknowledgements
We would like to thank Nicole Soranzo for advice on the final stage of the manuscript preparation.

Peer review information
Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions
F.M. conceived the study and planned the design with L. D. G and L.M.. G.H., T.K., M.M., G.H., L.P. conceived specific analysis. G.H. and M.C.M. provided methodological support and shared software/source code L.D. G, L.M, A.R., F.S., F. M. performed the analyses. F.M, L.D.G, L.M, A.R., M.V. and F.S. wrote the manuscript with input from all authors. All authors read and approved the final manuscript.

Funding
LDG and MV were supported by #NEXTGENERATIONEU (NGEU) and funded by the Ministry of University and Research (MUR), National Recovery and Resilience Plan (NRRP), project MNESYS (PE0000006) – A Multiscale integrated approach to the study of the nervous system in health and disease (DN. 1553 11.10.2022). FM was supported by Fondazione con il Sud (2018-PDR-01136) and by the Italian Ministry of University and Research (2022P2ZESR). MV was supported by the Italian Ministry of University and Research (2022E8NN2N). FS is a PhD student within the European School of Molecular Medicine (SEMM). GH was supported by the Wellcome Trust (224575/Z/21/Z). LP is funded by the Italian Ministry of University and Research (PRIN 2022B27XYM). LM and TK were supported by KU Leuven BOF-C24 grant ZKD6488 C24M/19/075 and FWO grant G0A4521N (TK).

Data availability
Our approach to assessing ancestral composition using principal component analysis and NNLS (PANE) is available as an R package at github.com/lm-ut/PANE and in Zenodo https://doi.org/10.5281/zenodo.14016612 [30, 31]. This tool is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0). To view a copy of this license, visit https://creativecommons.org/licenses/by-nc-nd/4.0/deed.en. Requests for accessing previously published data used in this work should be directed to the corresponding authors of the publications where they were originally presented [19, 20, 32, 32–42]. The Allen Ancient Data Resource is available at [43, 44], Raveane et al. 2019 at https://capelligroup.wordpress.com/data/.

Declarations

Ethics approval and consent to participate
No ethical consent was required for this study.

Competing interests
All authors certify that they have no affiliations with or involvement in any organization or entity with any financial or non-financial interest in the subject matter or materials discussed in this manuscript. The authors declare that they have no conflicts of interest.

Received: 6 May 2024 Accepted: 30 January 2025
Published online: 11 February 2025

References

1. Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, et al. A genetic atlas of human admixture history. Science. 2014;343(6172):747–51.
2. Montinaro F, Busby GBJ, Pascali VL, Myers S, Hellenthal G, Capelli C. Unravelling the hidden ancestry of American admixed populations. Nat Commun. 2015;6(1):6596.
3. Busby GB, Band G, Le Si Q, Jallow M, Bougama E, Mangano VD, et al. Admixture into and within sub-Saharan Africa. eLife. 2016;5:e15266.

4. Ongaro L, Scliar MO, Flores R, Raveane A, Marnetto D, Sarno S, et al. The genomic impact of European colonization of the Americas. Curr Biol CB. 2019;29(23):3974-3986.e4.
5. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient admixture in human history. Genetics. 2012;192(3):1065–93.
6. Schlebusch CM, Skoglund P, Sjödin P, Gattepaille LM, Hernandez D, Jay F, et al. Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. Science. 2012;338(6105):374–9.
7. Leslie S, Winney B, Hellenthal G, Davison D, Boumertit A, Day T, et al. The fine-scale genetic structure of the British population. Nature. 2015;519(7543):309–14.
8. Pankratov V, Montinaro F, Kushniarevich A, Hudjashov G, Jay F, Saag L, et al. Differences in local population history at the finest level: the case of the Estonian population. Eur J Hum Genet EJHG. 2020;28(11):1580–91.
9. Skoglund P, Thompson JC, Prendergast ME, Mittnik A, Sirak K, Hajdinjak M, et al. Reconstructing prehistoric African population structure. Cell. 2017;171(1):59-71.e21.
10. Akbari A, Barton AR, Gazal S, Li Z, Kariminejad M, Perry A, et al. Pervasive findings of directional selection realize the promise of ancient DNA to elucidate human adaptation. bioRxiv. 2024; Available from: https://www.biorxiv.org/content/early/2024/09/15/2024.09.14.613021.
11. Allentoft ME, Sikora M, Refoyo-Martínez A, Irving-Pease EK, Fischer A, Barrie W, et al. Population genomics of post-glacial western Eurasia. Nature. 2024;625(7994):301–11.
12. Hui R, D'Atanasio E, Cassidy LM, Scheib CL, Kivisild T. Evaluating genotype imputation pipeline for ultra-low coverage ancient genomes. Sci Rep. 2020;10(1):18542.
13. Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, et al. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature. 2015;522(7555):207–11.
14. Harney É, Patterson N, Reich D, Wakeley J. Assessing the performance of qpAdm: a statistical tool for studying population admixture. Genetics. 2021;217(4):iyaa045.
15. Damgaard P de B, Marchi N, Rasmussen S, Peyrot M, Renaud G, Korneliussen T, et al. 137 ancient human genomes from across the Eurasian steppes. Nature. 2018;557(7705):369–74.
16. Haber M, Doumet-Serhal C, Scheib C, Xue Y, Danecek P, Mezzavilla M, et al. Continuity and admixture in the last five millennia of Levantine history from ancient Canaanite and present-day Lebanese genome sequences. Am J Hum Genet. 2017;101(2):274–82.
17. Hajdinjak M, Fu Q, Hübner A, Petr M, Mafessoni F, Grote S, et al. Reconstructing the genetic history of late Neanderthals. Nature. 2018;555(7698):652–6.
18. Harney É, May H, Shalem D, Rohland N, Mallick S, Lazaridis I, et al. Ancient DNA from Chalcolithic Israel reveals the role of population mixture in cultural transformation. Nat Commun. 2018;9(1):3336.
19. Lazaridis I, Mittnik A, Patterson N, Mallick S, Rohland N, Pfrengle S, et al. Genetic origins of the Minoans and Mycenaeans. Nature. 2017;548(7666):214–8.
20. Narasimhan VM, Patterson N, Moorjani P, Rohland N, Bernardos R, Mallick S, et al. The formation of human populations in South and Central Asia. Science. 2019;365:eaat7487.
21. Olalde I, Brace S, Allentoft ME, Armit I, Kristiansen K, Booth T, et al. The Beaker phenomenon and the genomic transformation of northwest Europe. Nature. 2018;555(7695):190–6.
22. Patterson N, Price AL, Reich D. Population structure and eigenanalysis PLOS Genet. 2006;2(12): e190.
23. Eren Yüncü, Ulaş Işıldak, Matthew P. Williams, Christian D. Huber, Olga Flegontova, Leonid A. Vyazov, et al. False discovery rates of *qpAdm*-based screens for genetic admixture. bioRxiv. 2023;2023.04.25.538339.
24. Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, et al. Genomic insights into the origin of farming in the ancient Near East. Nature. 2016;536(7617):419–24.
25. Peter BM. A geometric relationship of F2, F3 and F4-statistics with principal component analysis. Philos Trans R Soc B Biol Sci. 1852;2022(377):20200413.
26. Moorjani P, Patterson N, Hirschhorn JN, Keinan A, Hao L, Atzmon G, et al. The history of African gene flow into Southern Europeans, Levantines, and Jews. PLOS Genet. 2011;7(4): e1001373.
27. Conley AB, Rishishwar L, Ahmad M, Sharma S, Norris ET, Jordan IK, et al. Rye: genetic ancestry inference at biobank scale. Nucleic Acids Res. 2023;51(8): e44.
28. Molinaro L, Marnetto D, Mondal M, Ongaro L, Yelmen B, Lawson DJ, et al. A chromosome-painting-based pipeline to infer local ancestry under limited source availability. Genome Biol Evol. 2021;13(4):evab025.
29. Busing FMTA, Leeden RVD. Delete-m Jackknife for Unequal m. Stat Comput. 1999;9(1):3–8.
30. Molinaro L, de Gennaro L, Montinaro F. PANE - Principal component Ancestry proportions using NNLS Estimation. Zenodo; 2024. Available from: https://doi.org/10.5281/zenodo.14016612.
31. Molinaro L, de Gennaro L, Montinaro F. PANE - Principal component Ancestry proportions using NNLS Estimation. GitHub; 2024. Available from: https://github.com/lm-ut/PANE.
32. Lazaridis I, Alpaslan-Roodenberg S, Pinhasi R, Reich D. The genetic history of the Southern Arc: a bridge between West Asia and Europe. Harvard Dataverse; 2022. Available from: https://doi.org/10.7910/DVN/3AR0CD.
33. Fernandes DM, Mittnik A, Olalde I, Lazaridis I, Cheronet O, Rohland N, et al. The spread of steppe and Iranian-related ancestry in the islands of the western Mediterranean. Nat Ecol Evol. 2020;4(3):334–45.
34. Raveane A, Aneli S, Montinaro F, Athanasiadis G, Barlera S, Birolo G, et al. Population structure of modern-day Italians reveals patterns of ancient and archaic ancestries in Southern Europe. Sci Adv. 2019;5(9):eaaw3492.
35. Saupe T, Montinaro F, Scaggion C, Carrara N, Kivisild T, D'Atanasio E, et al. Ancient genomes reveal structural shifts after the arrival of Steppe-related ancestry in the Italian Peninsula. Curr Biol CB. 2021;31(12):2576-2591.e12.
36. Aneli S, Saupe T, Montinaro F, Solnik A, Molinaro L, Scaggion C, et al. The genetic origin of Daunians and the Pan-Mediterranean Southern Italian Iron Age context. Mol Biol Evol. 2022;39(2):msac014.
37. Haber M, Nassar J, Almarri MA, Saupe T, Saag L, Griffith SJ, et al. A genetic history of the Near East from an aDNA time course sampling eight points in the past 4,000 years. Am J Hum Genet. 2020;107(1):149–57.
38. Agranat-Tamir L, Waldman S, Martin MAS, Gokhman D, Mishol N, Eshel T, et al. The genomic history of the Bronze Age Southern Levant. Cell. 2020;181(5):1146-1157.e11.

39. Fu Q, Posth C, Hajdinjak M, Petr M, Mallick S, Fernandes D, et al. The genetic history of Ice Age Europe. Nature. 2016;534(7606):200–5.
40. Prüfer K, de Filippo C, Grote S, Mafessoni F, Korlević P, Hajdinjak M, et al. A high-coverage Neandertal genome from Vindija Cave in Croatia. Science. 2017;358(6363):655–8.
41. Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, et al. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature. 2010;468(7327):1053–60.
42. Mathieson I, Alpaslan-Roodenberg S, Posth C, Szécsényi-Nagy A, Rohland N, Mallick S, et al. The genomic history of southeastern Europe. Nature. 2018;555(7695):197–203.
43. Mallick S, Micco A, Mah M, Ringbauer H, Lazaridis I, Olalde I, et al. The Allen Ancient DNA Resource (AADR) a curated compendium of ancient human genomes. Sci Data. 2024;11(1):182.
44. Mallick S, Reich D. The Allen Ancient DNA Resource (AADR): a curated compendium of ancient human genomes. Harvard Dataverse; 2023. Available from: https://doi.org/10.7910/DVN/FFIDCW.
45. Li N, Stephens M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics. 2003;165(4):2213–33.
46. Lazaridis I, Alpaslan-Roodenberg S, Acar A, Açıkkol A, Agelarakis A, Aghikyan L, et al. The genetic history of the Southern Arc: a bridge between West Asia and Europe. Science. 2022;377(6609):eabm4247.
47. Raveane A, Molinaro L, Aneli S, Capodiferro MR, de Gennaro L, Ongaro L, et al. Assessing temporal and geographic contacts across the Adriatic Sea through the analysis of genome-wide data from Southern Italy. Genomics. 2022;114(4): 110405.
48. Aneli S, Mezzavilla M, Bortolini E, Pirastu N, Girotto G, Spedicati B, et al. Impact of cultural and genetic structure on food choices along the Silk Road. Proc Natl Acad Sci U S A. 2022;119(47): e2209311119.
49. Brace S, Diekmann Y, Booth TJ, van Dorp L, Faltyskova Z, Rohland N, et al. Ancient genomes indicate population replacement in Early Neolithic Britain. Nat Ecol Evol. 2019;3(5):765–71.
50. Jacobs GS, Hudjashov G, Saag L, Kusuma P, Darusallam CC, Lawson DJ, et al. Multiple deeply divergent Denisovan ancestries in Papuans. Cell. 2019;177(4):1010-1021.e32.
51. McVean G. A genealogical interpretation of principal components analysis. PLOS Genet. 2009;5(10):1–10.
52. Novembre J, Stephens M. Interpreting principal component analyses of spatial population genetic variation. Nat Genet. 2008;40(5):646–9.
53. Kelleher J, Etheridge AM, McVean G. Efficient coalescent simulation and genealogical analysis for large sample sizes. PLOS Comput Biol. 2016;12(5): e1004842.
54. van Dorp L, Balding D, Myers S, Pagani L, Tyler-Smith C, Bekele E, et al. Evidence for a common origin of blacksmiths and cultivators in the Ethiopian Ari within the last 4500 years: lessons for clustering-based inference. PLOS Genet. 2015;11(8):1–49.
55. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015;4(1):s13742–015–0047–8.
56. Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, et al. Insights into human genetic variation and population history from 929 diverse genomes. Science. 2020;367(6484):eaay5012.
57. Reitsema LJ, Mittnik A, Kyle B, Catalano G, Fabbri PF, Kazmi ACS, et al. The diverse genetic origins of a Classical period Greek army. Proc Natl Acad Sci. 2022;119(41): e2205272119.
58. Posth C, Zaro V, Spyrou MA, Vai S, Gnecchi-Ruscone GA, Modi A, et al. The origin and legacy of the Etruscans through a 2000-year archeogenomic time transect. Sci Adv. 2021;7(39):eabi7673.
59. Lawson CL, Hanson RJ. Solving least squares problems. Society for Industrial and Applied Mathematics; 1995. Available from: https://epubs.siam.org/doi/abs/10.1137/1.9781611971217.
60. R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2020. Available from: https://www.R-project.org/
61. Büsing C, Koster AMCA, Kutschka M. Recoverable robust knapsacks: the discrete scenario case. 2011;5:379–92.
62. Baumdicker F, Bisschop G, Goldstein D, Gower G, Ragsdale AP, Tsambos G, et al. Efficient ancestry and mutation simulation with msprime 1.0. Genetics. 2022;220(3):iyab229.
63. Maier R, Flegontov P, Flegontova O, Işıldak U, Changmai P, Reich D. On the limits of fitting complex models of population history to *f*-statistics. Nordborg M, Przeworski M, Balding D, Wiuf C, editors. eLife. 2023;12:e85492.
64. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. PLOS Genet. 2012;8(1): e1002453.

Publisher's Note