scientific reports

Check for updates

OPEN

Multivariate statistical approach and machine learning for the evaluation of biogeographical ancestry inference in the forensic field

Eugenio Alladio[1,3], Brando Poggiali[2], Giulia Cosenza[2] & Elena Pilli[2]✉

The biogeographical ancestry (BGA) of a trace or a person/skeleton refers to the component of ethnicity, constituted of biological and cultural elements, that is biologically determined. Nowadays, many individuals are interested in exploring their genealogy, and the capability to distinguish biogeographic information about population groups and subgroups via DNA analysis plays an essential role in several fields such as in forensics. In fact, for investigative and intelligence purposes, it is beneficial to inference the biogeographical origins of perpetrators of crimes or victims of unsolved cold cases when no reference profile from perpetrators or database hits for comparative purposes are available. Current approaches for biogeographical ancestry estimation using SNPs data are usually based on PCA and Structure software. The present study provides an alternative method that involves multivariate data analysis and machine learning strategies to evaluate BGA discriminating power of unknown samples using different commercial panels. Starting from 1000 Genomes project, Simons Genome Diversity Project and Human Genome Diversity Project datasets involving African, American, Asian, European and Oceania individuals, and moving towards further and more geographically restricted populations, powerful multivariate techniques such as Partial Least Squares-Discriminant Analysis (PLS-DA) and machine learning techniques such as XGBoost were employed, and their discriminating power was compared. PLS-DA method provided more robust classifications than XGBoost method, showing that the adopted approach might be an interesting tool for forensic experts to infer BGA information from the DNA profile of unknown individuals, but also highlighting that the commercial forensic panels could be inadequate to discriminate populations at intra-continental level.

Inference of individual biogeographic ancestry plays an essential role in several genetics fields, from population/anthropological studies with the interpretation of genetic admixture in populations or human population expansion, movement, and interaction (e.g.[1]) to medical applications with the evaluation of disease susceptibility (e.g.[2]). Moreover, other disciplines including epidemiology, pharmacogenomics, and forensics, can benefit from biogeographic ancestry testing. In addition, ancestry analysis is of increasing relevance to crime investigations. Identifying perpetrators of crimes or victims of unsolved cold cases by DNA analysis may be hindered by few or no investigative leads and the consequent absence of reference profiles from perpetrators or database hits. In such cases, there is a need for additional genetic information, such as biogeographical ancestry (BGA), that left the trace sample at the crime scene. The best way to assign an individual into a particular population via genetic testing is to use ancestry informative markers (AIMs) –markers characterized by essential differences in allele frequencies between populations[3–5]. As proposed by several studies (for example[6–25]), short tandem repeats (STRs), single nucleotide polymorphisms (SNPs), insertion/deletion polymorphisms (InDels), and microhaplotypes can be used as AIMs for ancestry inference. However, autosomal single nucleotide polymorphisms are the best choices due to their inherent stability, high density of genome-wide distribution, and pronounced frequency

[1]Department of Chemistry, University of Turin, Turin, Italy. [2]Department of Biology, Forensic Molecular Anthropology Laboratory, University of Florence, Florence, Italy. [3]Present address: Centro Regionale Antidoping e di Tossicologia "A. Bertinaria", Orbassano, Torino, Italy. ✉email: elena.pilli@unifi.it

variation among populations. Recently, the application of massively parallel sequencing technologies for BGA tool developments allowed the simultaneous analysis of a significant number of SNPs than the SNaPshot-based minisequencing technology as attested by the development for forensics of commercial and non-commercial panels[26–29]. To date, the statistical clustering methods most used for the inference of the biogeographical ancestry of a person or trace relies on are PCA, STRUCTURE[8,30] and GenoGeographer[31,32]. However, although they provide easy ways to visualize data clustering, these methods are empirical and not adequate for ancestry inference in the forensic field.

In the last decades, analysts have gradually started to take into account all the variables/predictors simultaneously (i.e., in a multivariate way)[33–42], since this approach allows to extract from the datasets more information than just looking at them in a univariate way, mainly when large amounts of noisy or redundant data occur. One of the attempts to use a multivariate statistical approach to identify clusters of genetically structured populations was described by Jombart et al.[43] in 2010. In their paper, Discriminant Analysis of Principal Components (DAPC) was applied to simulated data and the performance of their approach was compared to that obtained using STRUCTURE. Multivariate data analysis techniques can be roughly divided into two main categories: (i) pattern recognition techniques; (ii) regression/calibration models. Very concisely, pattern recognition models can be again divided into two categories: (i) unsupervised models (where the information about the a priori classification of each of the individuals/instances/samples under exam is missing); (ii) supervised/classification models (where the a priori classification of each of the instances under exam is known). One of the most known unsupervised methodologies (also known as exploratory analyses) is Principal Components Analysis (PCA)[44]. On the other hand, supervised/classification modeling techniques can present an important family of strategies known as discrimination models, such as Partial Least Squares-Discriminant Analysis (PLS-DA)[45], that aim to calculate specific boundaries in the multidimensional space that allow separating the different objects within their corresponding classes[46]. Therefore, for the first time, we decided to adopt a multivariate methodology in the present study, such as Partial Least Squares-Discriminant Analysis (PLS-DA) on several SNPs datasets involving instances of different populations showing different BGA. Our main aim was to build robust multivariate models to interpret the results of BGA inference by using different BGA panels that have been developed for this purpose. The PLS-DA approach has been already adopted by Alladio et al.[30] on DNA STRs data to infer the biogeographical ancestry of unknown individuals, and the developed models provided interesting performances in terms of sensitivity, specificity, and accuracy. However, there are no examples of using such machine learning tool on the more informative SNPs data, especially in terms of ancestry, so that the authors decided to extend this approach on a large amount of data and individuals, too.

Simultaneously, a second and very popular supervised learning algorithm named XGBoost (eXtreme Gradient Boosting) was tested on the collected data to evaluate the performance of another machine learning approach and compare its results with PLS-DA. As well as PLS-DA, no example of this approach for BGA inference have been reported in literature dealing with SNPs data.

The BGA panels evaluated in this study are EUROFORGEN Global AIMs SNP (128 AISNPs here, EUROFORGEN)[28], Verogen® ForenSeq™ DNA Signature Prep Kit (55 AISNPs here, ForenSeq)[27], MAPlex—Multiplex for the Asia–Pacific (144 AISNPs here, MAPlex)[29], and Thermo Fisher HID Ion AmpliSeq™ Ancestry Panel (165 AISNPs here, Thermo Fisher)[26].

Methods

Datasets. The SNPs dataset evaluated in this study was composed of 3,557 individuals from 1,000 Genomes project (2,504 individuals from 26 populations)[47], Simons Genome Diversity Project (SDGP) (279 individuals from 130 populations) (https://www.simonsfoundation.org/simons-genome-diversity-project/) and Human Genome Diversity Project (HGDP) (929 individuals from 54 populations)[48].

The individuals shared from the three projects were removed.

All the tested multivariate models were calculated in two steps:

- the first models involved the evaluation of the whole data available by evaluating the different BGA categories in the form of "continental" ancestry, such as African, American, Asian (combining Central, East, North, and South Asia populations), European (involving Middle East populations, too), and Oceanian individuals;
- the following models were built on each continent separately (i.e., Asia, Africa, America, Europe, and Oceania by considering the most represented populations (i.e., 80 individuals, at least).

Multivariate modeling. PCA, PLS-DA, and XGBoost models were applied on the collected data derived from the different BGA panels. The AIMs profile of each individual (instance) was transformed into a row of zeros and ones by using a one-hot encoding strategy developed in the R environment (version 4.0.2)[49]. In detail, for all the tested subjects, a value equal to 1 was reported for the *n* SNP recorded for each specific AIM, while a value equal to 0 was reported for the other available SNPs of the previously cited AIM. Consequently, each instance's AIMs profile consisted of a string of 0 and 1 (i.e., one-hot encoding). As similarly reported in[30], PCA and PLS-DA approaches were used to obtain trustworthy and cross-validated models to infer the BGA information of the available instances. The following R packages were exploited for this purpose: *correlationfunnel*[50], *dplyr*[51], *ggplot2*[52], *mdatools*[53], *mixOmics*[54], *mlr*[55], *plotly*[56], *pls*, *plsVarSel*[57] and *xgboost*[58].

Principal components analysis (PCA, also known as *eigenvector analysis*) was preliminarily employed to perform exploratory analyses and dimension reduction studies on the available data[59–61]. It is commonly used to graphically represent the acquired data by evaluating any subgroup or cluster within the instances and assessing the correlation among the collected features[44,62]. Starting from the original dataset (consisting of a matrix X), PCA aims to eliminate redundant and noisy information by selecting a small number of variables, leading to a

better understanding of the data structure. PCA calculates new orthogonal (i.e., uncorrelated) variables, named Principal Components (PCs), that represent a linear combination of the original variables aimed to reproduce the structure of the original data (X), but in an optimal and interpretable way. In practice, PCA approach decomposes the original matrix X into the product of two new matrixes, named T and P, plus a matrix of residuals E, as follows:

$$X_{(n,p)} = T_{(n,f)} \times P_{(f,p)} + E_{(n,p)}$$

where n is the number of instances (here, the genotyped subjects), p is the number of predictors (features, here the measured AIMs), and f represents the number of selected principal components. The first PC is oriented in the direction of the maximum variance. Afterward, the second PC is oriented towards the maximum residual variance, chosen among the infinite orthogonal directions regarding the first component, etcetera. T matrix contains the objects' coordinates (called *scores*) in the new multivariate space delimited by a f-number of PCs (i.e., scores plot). P matrix gives the variable vector coordinates (called *loadings*) in the new multivariate system (i.e., loadings plot). These linear functions are calculated according to specific weighting coefficients representing the linkage between the original variables and the new components. The loadings are the elements of the eigenvector of the variance–covariance matrix of the original X matrix. Each eigenvector has a corresponding eigenvalue that indicates the amount of variance explained (EV) by each PC. In the present study, only the first f PCs were selected to account for a specific percentage (around 80% of cumulative explained variance, CEV) of the system's overall variance.

Since the PCA approach is primarily an exploratory (unsupervised) data analysis approach, further PLS-DA and XGBoost models were calculated to build properly supervised classification algorithms. These models were evaluated to develop tools capable of predicting and inferring the BGA of new (i.e., unknown) samples and individuals, together with classification probabilities and scores. The adoption of classification models in BGA inference should overcome PCA only since supervised models maximize the covariance between the independent variables (i.e., the SNPs data) and the dependent response (i.e., the BGA of the collected individuals). Moreover, the supervised approaches are particularly suitable for forensic tasks like the one examined in this study since they are appropriately made to predict new unknown samples.

Subsequently, PLS-DA was adopted to investigate the covariance between a matrix X of predictors (i.e., the measured AIMs) and the BGA responses included in a matrix Y. In particular, the main point of multivariate classification models like PLS-DA is to investigate the relationships between X and Y and build a model capable of predicting the BGA responses of new samples whose genotypes will be measured in future caseworks. PLS again calculates new components, called *latent variables* (LV), computed by evaluating X and Y matrixes simultaneously. In particular, from a geometric point-of-view, the latent variables represent a slightly rotated version of the Principal Components[63–66]. While PCA maximizes the X matrix variance, the PLS approach iteratively maximizes the covariance between X and Y. For this purpose, the components calculated on Y are rotated to maximize the covariance concerning the components calculated on X. The iterative process ends when no more helpful information can be extracted from X and Y matrices. Briefly, PLS algorithms can be summarized by the following steps:

1. Calculating two matrices E (= X) and F (= Y) whose columns are centered and normalized;
2. Initializing a vector u is with random values before starting the iterative process;
3. Calculating $w \sim E^T u$, where w represents the weights (coefficients) relative to X and the symbol ~ means "*to normalize the result of the operation*", as suggested by[64];
4. Calculating $t \sim Ew$, where t represents the new scores of X;
5. Calculating $q \sim F^T t$, where q represents the weights (coefficients) relative to Y;
6. Calculating $u \sim Fq$, where u represents the new scores of Y;
7. If t has not converged, then the iterative algorithm moves back to step 3. Otherwise, if t has converged, a b value is computed. This value allows predicting Y from t by following the equation $b = t^T u$. Simultaneously, the loadings of X are computed by following the equation $p = E^T t$.
8. Finally, the effect of t is subtracted from both E and F matrixes, as follows: $E_{final} = E\text{-}tp^T$ and $F_{final} = F\text{-}btc^T$. In particular, the scalar b values are represented by a diagonal matrix B.

The sum of squares of X (Y) explained by the latent vector is computed as $p^T p$ (b^2 for Y), and the percentage of the variance explained (EV) by the PLSR model is obtained by dividing the explained sum of squares by the corresponding total sum of squares[64]. The discriminant version of PLS (PLS-DA) is computed by classifying the objects through X's regression (PLS) and a matrix Y that contains binary responses. In particular, Y consists of G columns equal to the number of categories (i.e., BGA classes). Each column contains the class membership information of the corresponding n individuals (instances). Since the response is binary (or one-hot encoded), if an instance belongs to a specific g-th category, it shows a response equal to 1 within the g-th column. Otherwise, its response is coded as 0.

Finally, the XGBoost algorithm, reported in[58], was tested since it has become a viral classification algorithm, mainly when data are expressed in a one-hot encoded format (i.e., binary data involving only 0 and 1 values). It frequently shows improved performances compared to the well-known supervised classification approaches. It shows several strengths typical of the tree-based algorithms, such as the capability of handling categorical features (in terms of one-hot encoding) and the possibility of making no assumptions about the distributions of the collected data. The main computational weakness of XGBoost algorithm is the necessity of setting many parameters (*hyperparameters*) before obtaining the best (robust and cross-validated) models. In particular, a grid search approach was used in this study to optimize several hyperparameters such as *eta* (i.e., the learning rate, a

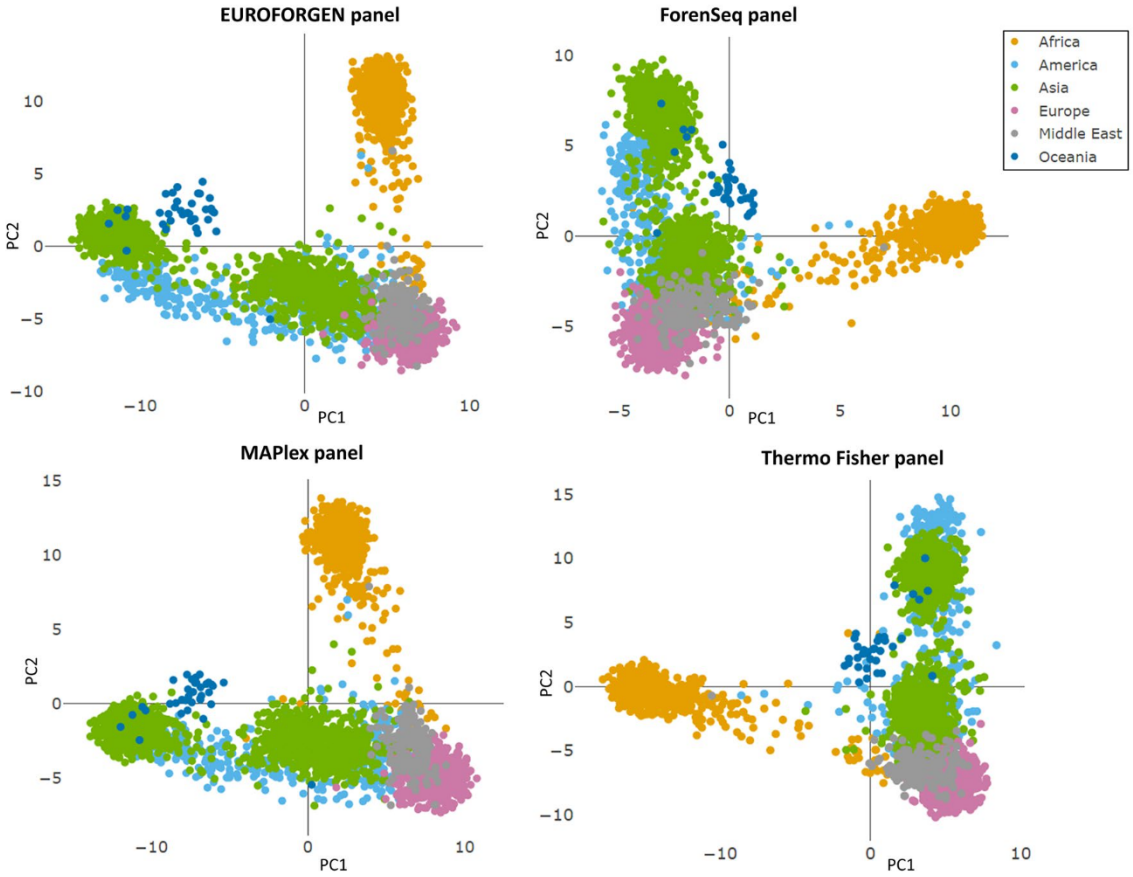

Figure 1. PCA Scores plots showing the PCA models obtained for the different evaluated AIMs panels.

number between 0 and 1), *gamma* (i.e., the minimum number of splitting for a node, from 0 up to 14), *max_depth* (i.e., a number that indicates how deeply each evaluated tree can grow, from 1 up to 5), *min_child_weight* (i.e., a value defining the level of impurity sustainable for a node, between 1 and 9), *subsample* (i.e., a value describing the proportion of samples to be randomly sampled during the evaluation of each tree, between 0 and 1), *colsample_bytree* (i.e., a number describing the proportion of features selected by each tree, between 0.5 and 1), and *nrounds* (i.e., the number of trees that can be sequentially built within the model[55,67]. XGBoost models were computed on all the available SNPs collected in this study at inter-continental and continental levels. All XGBoost models (as well as the PLS-DA ones) were expressed in terms of sensitivity, specificity, and accuracy. Confusion matrixes and AUC values of the calculated Receiver operating characteristic (ROC) curves were evaluated, too. AUC values equal to 0.5 suggest no discrimination. In contrast, AUC values between 0.7 and 0.8 indicate that the model has acceptable discrimination and AUC values between 0.8 and 0.9 suggest an excellent capacity of discrimination and values equal to or greater than 0.9 indicate outstanding discrimination[68].

The hyperparameters' tuning of all the developed models (PLS-DA and, mainly, XGBoost) was performed using a grid search approach and employing a fivefold cross-validation approach with venetian blinds sampling design. The models reported in this study are, therefore, the best we obtained from tuning our models on the data obtained for the different panels (the values of the hyperparameters are not reported). Root Mean Square Error in Cross Validation (RMSECV) was evaluated when building the PCA and PLS-DA models to define the optimal number of components for the developed models.

All experiments were achieved in accordance with relevant guidelines and regulations.

Results and discussion

PCA, PLS-DA and XGBoost models at inter-continental level. As proposed in different papers[28,69–71], PCA was first performed to preliminary investigate the available datasets involving the four selected AIMs panels for BGA inference. As expected, for the first level of BGA (i.e., inter-continental BGA) inference, several separate clusters corresponding to African, American, Asian, European, and Oceanian individuals were observed in the space of the first two PCs (Fig. 1). This result turned straightforward for all the evaluated AIMs panels.

After an initial PCA analysis with the Asian continent in its entirety, the Asian was subdivided into its regions due to its breadth -within our dataset, individuals were belonging to different regions of Asia- and the fact that the prediction of the biogeographical origin within the Asian continent has been and is a subject extensively studied in the forensic field[25,72–76].

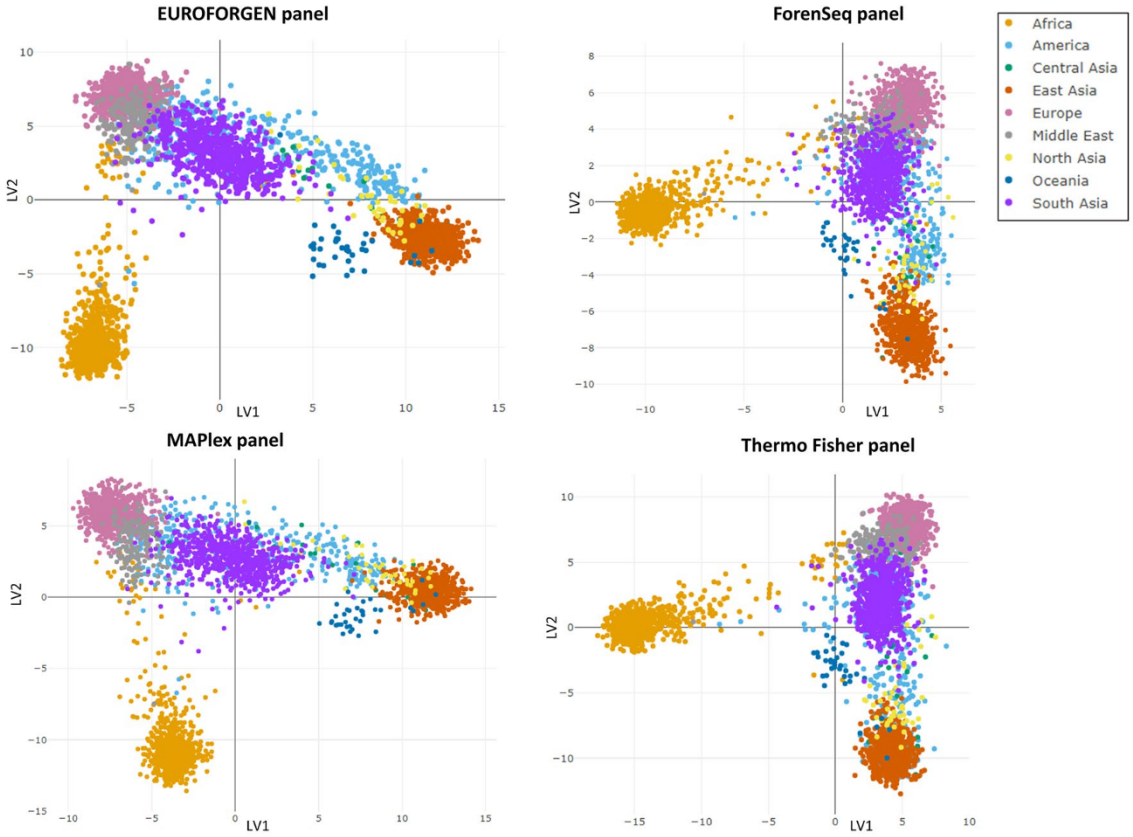

Figure 2. PLS-DA Scores plots showing the models obtained for the different evaluated AIMs panels.

As in our dataset, if considering Asia composed by Central, East, North, and South Asia populations, PCA plot highlights that African, East Asian, Oceanian, and (partially) North Asian and European individuals showed a better differentiation from the other tested individuals, while American, South Asian, Central Asian, and Middle East subjects provided an overlap in the PCA space. In addition, better separation of the evaluated populations can be observed in Additional file 1: Fig. S1 also involving three principal components (for a total amount of CEV% equal to 83%).

All the evaluated AIMs panels show a similar degree of separation among individuals belonging to different continental areas. However, only the African individuals reveal a separate cluster in all the panels, presumably due to the history of humans in Africa that is complex and includes demographic events that influenced patterns of genetic variation across the continent, and the fact that modern humans first appeared in Africa roughly 250,000–350,000 years before present and subsequently migrated to other parts of the world[77].

As shown in Additional file 1: Fig. S1a,b, the African individuals generate an elongated cluster (dark yellow) that extends towards the gray one corresponding to the Middle East region. By evaluating the African individuals closest to the Middle East cluster, we observed that they belong to the populations of northern Africa. The Middle East cluster is in the middle of the European and South Asian clusters and partly overlaps. The light blue cluster that corresponds to the admixed and non-admixed American population is projected toward the European cluster and partly overlaps with it, suggesting that admixed American individuals have an important proportion of European ancestry[78].

As it can be observed in Fig. 1 and in Additional file 1: Fig. S1, the distribution of the populations in the space of the PCs perfectly reflects the distribution of the populations in the globe: indeed, geographically distant populations are located distantly in the PCA plot, while geographically close populations, regardless of whether they belong to a continent or another, are close in the PCA plot.

Similar PCA plots were obtained by Glusman et al.[79] and Haber et al.[80] using a significantly greater number of SNPs, 300,000 and 240,000 respectively than those tested in all the forensic panels. Therefore, as previously highlighted[28,69,70], despite the limited number of SNPs, the performance of each panel across populations was generally consistent even if some genetic markers performed more than others.

However, although PCA analysis allows us to assign an individual to his/her population of origin through a visual, intuitive, and easy to interpret approach, it does not provide significant divergence between populations, and obviously, it cannot be used alone in forensic context because it does not provide an accurate statistical estimate of the weight of the evidence[69].

PLS-DA was then applied to the same experimental sets based on PCA modeling results to develop more reliable discrimination models to classify the variables. As a result, for the first level of BGA (i.e., inter-continental BGA) inference, African, American, East Asian, South Asian, Central Asian, North Asian, European, and

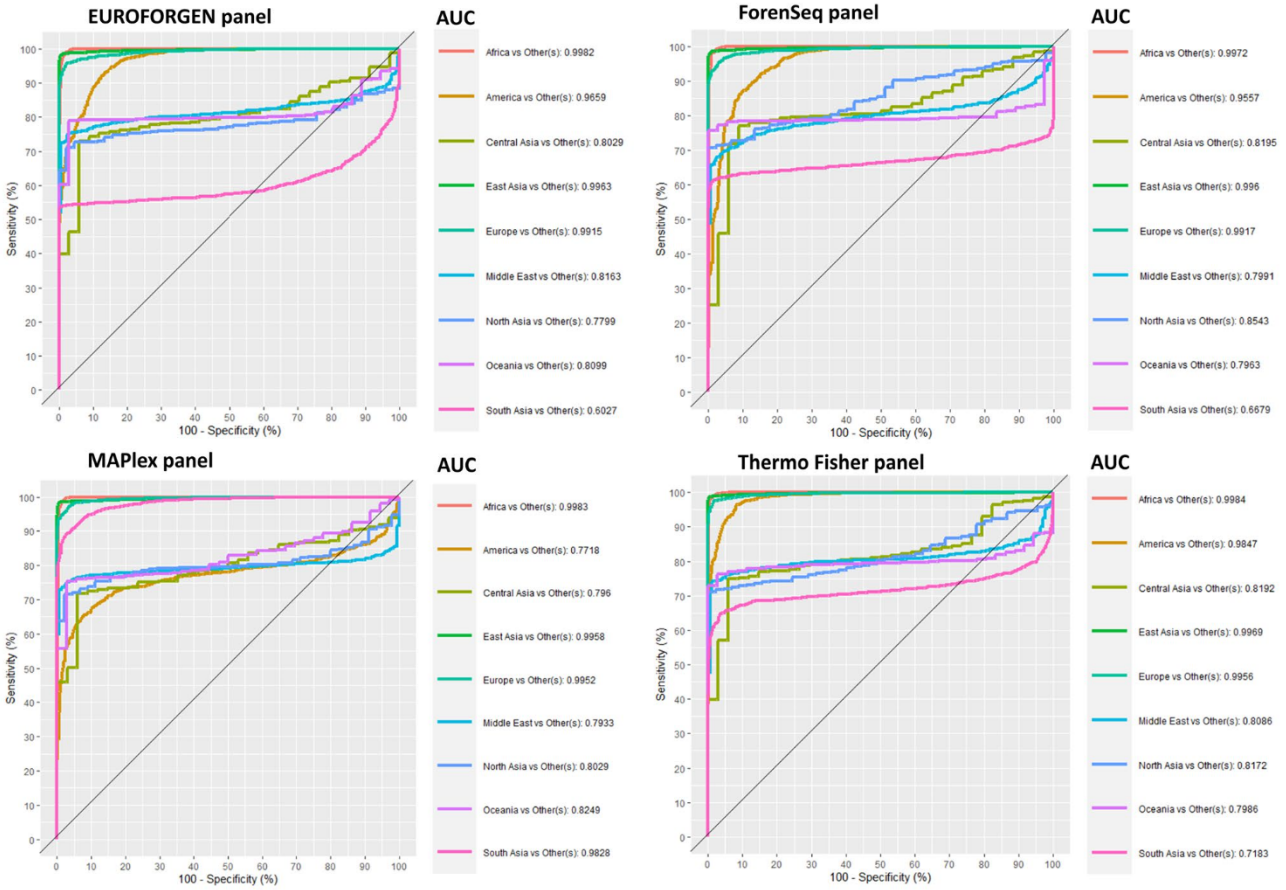

Figure 3. ROC curves, sensitivity, specificity, and AUC values for the tested continental populations.

Oceanian individuals were effectively separated using models involving two latent variables (LVs) (Fig. 2). This result turned noteworthy for all the evaluated panels.

Even if the PCA and PLS-DA plots may seem similar, the obtained Receiver Operating Characteristic (ROC) curves, together with the values of sensitivity, specificity, and AUC highlight the importance of a statistical tool to infer BGA. PLS-DA models for African, American, Asian, European, and Oceanian individuals provided optimal predictions with the CEV% values higher than 98% for all populations in all panels investigated except Oceania—Euroforgen (CEV% 88%), ForenSeq (CEV% 79%), MAPlex (CEV% 86%) and Thermo Fisher (CEV% 79%), and America in ForenSeq (CEV% 95%) panel-. The Oceania population results might be affected by the small number of individuals in the dataset showing this ancestry. All the developed models provided a CEV% higher than 80%, and all the tested AIMs panels proved reliable results that remarked the necessity to use a proper classification model, rather than PCA modeling, to infer BGA robustly.

In addition, through the PLS-DA model, the MaPlex panel ability to differentiate the set of individuals from South Asian to others was estimated with a high degree of accuracy (AUC = 0.9828). As expected from the preliminary assessment of MaPlex[29], no other panel considered in this study was found to be comparable with it in enhancing South Asian differentiation (Fig. 3). Outstanding discrimination was obtained for East Asian populations in all panels considered associated with less discrimination for Central and North Asian probably due to the limited number of Asian population samples in our dataset, the use of unsuitable markers to discriminate these areas, and the fact that Asia has been a critical hub of human migration and population admixture[81–83].

As shown in Fig. 3, there are some populations showing poor sensitivity and specificity values. As an example, South Asian individuals have low values for EUROFORGEN, ForenSeq and Thermo Fisher panels, while they are classified with promising results using the MAPlex panel. Similar behaviours are also observed for Middle East and Oceania individuals. These results reflect the fact that some panels, like MAPlex, have been developed to deeply investigate specific populations (i.e., Asia–Pacific populations) and their classification might be prone to better identify such individuals[29]. On the other hand, some populations (like Oceanian and Middle East subjects) showed a lower number of available individuals, compared to the other tested populations, so that the classification performance are not optimal and might be improved by raising the number of investigated subjects.

In accordance with Phillips et al.[29], our results indicated enhanced South Asian differentiation (AUC = 0.98) using MaPlex panel compared to other forensic panels (Fig. 4), but no increased differentiation between West Eurasian and East Asian populations was detected.

Afterward, the best XGBoost model obtained after the grid search approach provided the following performances (Table 1) in terms of sensitivity, specificity, and AUC. XGBoost algorithm was tested to compare its

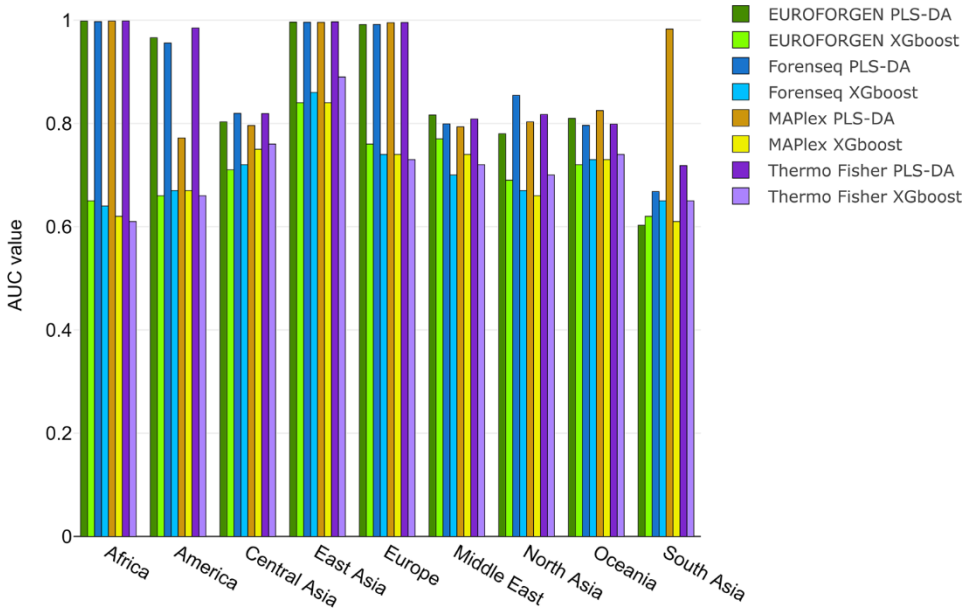

Figure 4. Comparison between AUC values of different populations obtained from PLS-DA and XGBoost model at inter-continental level considering Asian divided into regions.

	EUROFORGEN			ForenSeq			MAPlex			Thermo Fisher		
Populations	Sensitivity	Specificity	AUC	Sensitivity	Specificity	AUC	Sensitivity	Specificity	AUC	Sensitivity	Specificity	AUC
Africa	0.53	0.77	0.65	0.51	0.78	0.64	0.48	0.76	0.62	0.47	0.76	0.61
America	0.46	0.86	0.66	0.48	0.86	0.67	0.48	0.87	0.67	0.44	0.87	0.66
Central Asia	0.70	0.73	0.71	0.72	0.72	0.72	0.71	0.78	0.75	0.73	0.80	0.76
East Asia	0.78	0.91	0.84	0.83	0.88	0.86	0.79	0.88	0.84	0.84	0.93	0.89
Europe	0.63	0.89	0.76	0.58	0.89	0.74	0.63	0.85	0.74	0.58	0.88	0.73
Middle East	0.72	0.83	0.77	0.62	0.79	0.70	0.68	0.79	0.74	0.71	0.73	0.72
North Asia	0.56	0.82	0.69	0.61	0.73	0.67	0.59	0.74	0.66	0.57	0.82	0.70
Oceania	0.70	0.73	0.72	0.70	0.77	0.73	0.70	0.77	0.73	0.71	0.77	0.74
South Asia	0.50	0.73	0.62	0.55	0.75	0.65	0.49	0.73	0.61	0.55	0.75	0.65

Table 1. Sensitivity, specificity, and AUC values of the optimal XGBoost model built at inter-continental level for all panels investigated.

performances with those from PLS-DA to evaluate another ML model aimed to obtain optimal and feasible inference models for BGA prediction.

As it can be seen by the values reported in Table 1, XGBoost model provides interesting results, but slightly lower than those of PLS-DA models, especially when comparing the AUC values (Fig. 4).

As shown in Fig. 4, optimal AUC values (close to 1) were observed for African, American, East Asian, and European populations using PLS-DA method, while lower results (around 0.8) were obtained for Central Asia, Middle East, North Asia, Oceania, and South Asia (with the exception of MAPlex panel involving a PLS-DA model) areas. The best results were achieved when using PLS-DA modeling, showing AUC values substantially higher than those obtained by XGBoost. The worst predictions were those involving the South Asian populations overall with AUC values around 0.6. In parallel, STRUCTURE software was tested as a benchmark comparison. The AUC of STRUCTURE was calculated by comparing the ancestry predictions from STRUCTURE software with the real ancestry origins of the tested populations and individuals. Firstly, the number of K clusters (i.e., populations) we selected for our comparison with STRUCTURE was equal to the number of ancestry populations we tested for the different PLS-DA and XGBoost models at inter-continental and inter-continental levels. Then, using CLUMPP together with STRUCTURE, we were able to obtain the Q-matrices containing the membership coefficients for each individual in each cluster. Therefore, each individual was assigned to the ancestry (k-th cluster) showing the highest membership coefficient: this approach allowed us to obtain ROC curves and AUC values for comparing STRUCTURE approach to the predictions and the performance provided by PLS-DA and XGBoost models.

Comparison between AUC values of different populations obtained from PLS-DA, XGBoost and STRUCTURE model at inter-continental level is reported in the Fig. 5. As it can be observed in Fig. 5, better performance

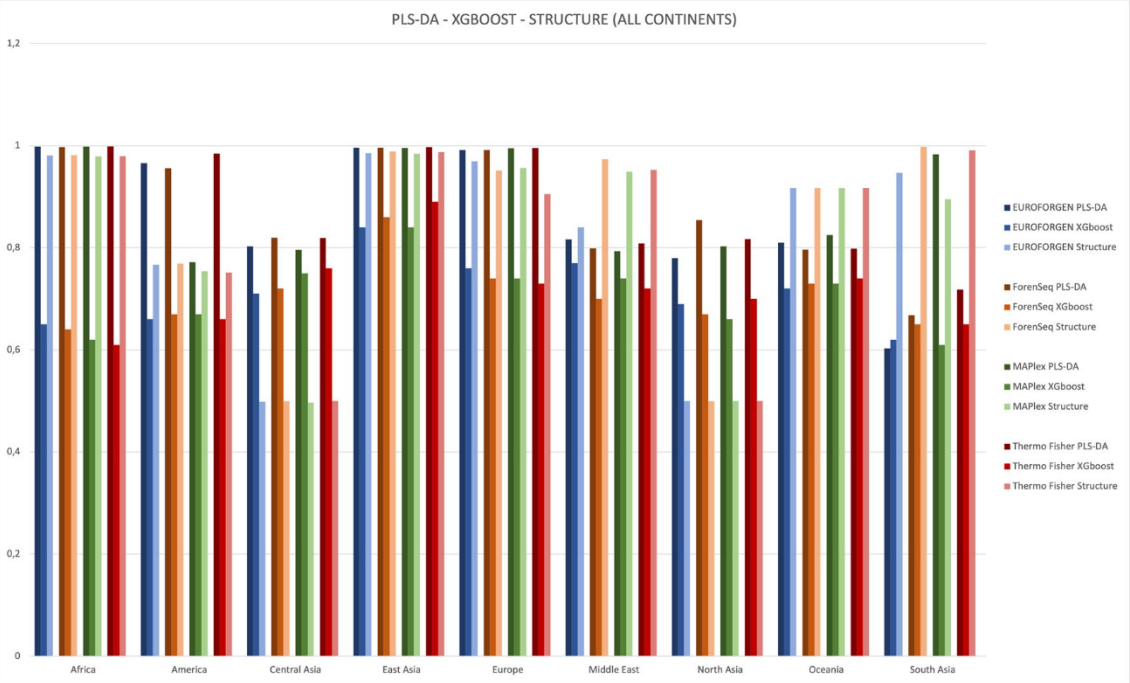

Figure 5. Comparison of AUC values of different populations obtained from PLS-DA, XGBoost, and STRU CTURE at inter-continental level considering Asian divided into regions.

was achieved when using PLS-DA modeling rather than STRUCTURE for diverse continents such as Africa, America, Europe and most of Asia (central, east and north Asian) for all panels investigated. Different results were observed in south Asia, Middle East and Oceania where STRUCTURE model seems to work best in almost all panels investigated with the exception of MaPlex panel in South Asia. The worst predictions were those involving XGBoost with AUC values on average lower than STRUCTURE except for Central Asian and North Asia.

PCA, PLS-DA and XGBoost models at intra-continental level. PCA model was assessed to infer BGA at continental level and, as expected[28,69], unsatisfactory separations were observed (an example is shown in Fig. 6 for MAPlex panel). In particular, the following countries and populations were evaluated for the different geographical areas:

- Africa: African Caribbeans, Gambia, Kenya, Nigeria, Sierra Leone;
- America: Colombia, Mexican Ancestry from Los Angeles, Mexico, Peru, Puerto Rico;
- Asia: Bangladesh, China, India, Japan, Pakistan, Sri Lanka;
- Europe: Finland, France, Great Britain, Italy, Spain, Israel.

These countries and populations were selected since they showed more than 80 genotyped individuals in the analyzed dataset; therefore, Oceanian individuals were not considered since the number of genotyped subjects was too limited. As observed in Fig. 6, no significant differences or clusters were detected when using PCA exploratory strategy. Considering Asian population plot, Japan and China provided a different cluster when compared to the other Asian countries but despite the MAPlex panel was specifically developed to provide differentiation of Asian population, can discriminate South from East Asian populations but the sub-populations in these geographical areas cannot be separated from each other. Similar results were observed for all the other BGA AIMs panels (Additional file 1: Figs. S2, S3, S4, S5).

In summary, if this traditional multivariate approach allows us to suggest the BGA of known individuals at the inter-continental level, it fails at intra-continental level, presumably due to the statistical method that is incapable to classify the variables.

Therefore, the application of the PCA model can be considered inadequate for forensic BGA inference goals. For this reason, we adopted proper classification models, such as PLS-DA and XGBoost, to improve our models' performance and obtain adequate separations among the populations.

Therefore, PLS-DA and XGBoost models were evaluated at intra-continental level. Figure 7 reports the models and the performance results of the PLS-DA model built to discriminate among the African population.

In the African scenario, the best results were achieved by EUROFORGEN and Thermo Fisher panels, but also MAPlex panel provided interesting results.

The AUC values of the EUROFORGEN panel (Fig. 7) between 0.8 and 0.9 for two out of five populations analyzed and greater than 0.9 for the remaining three, suggest an excellent capacity of discrimination and

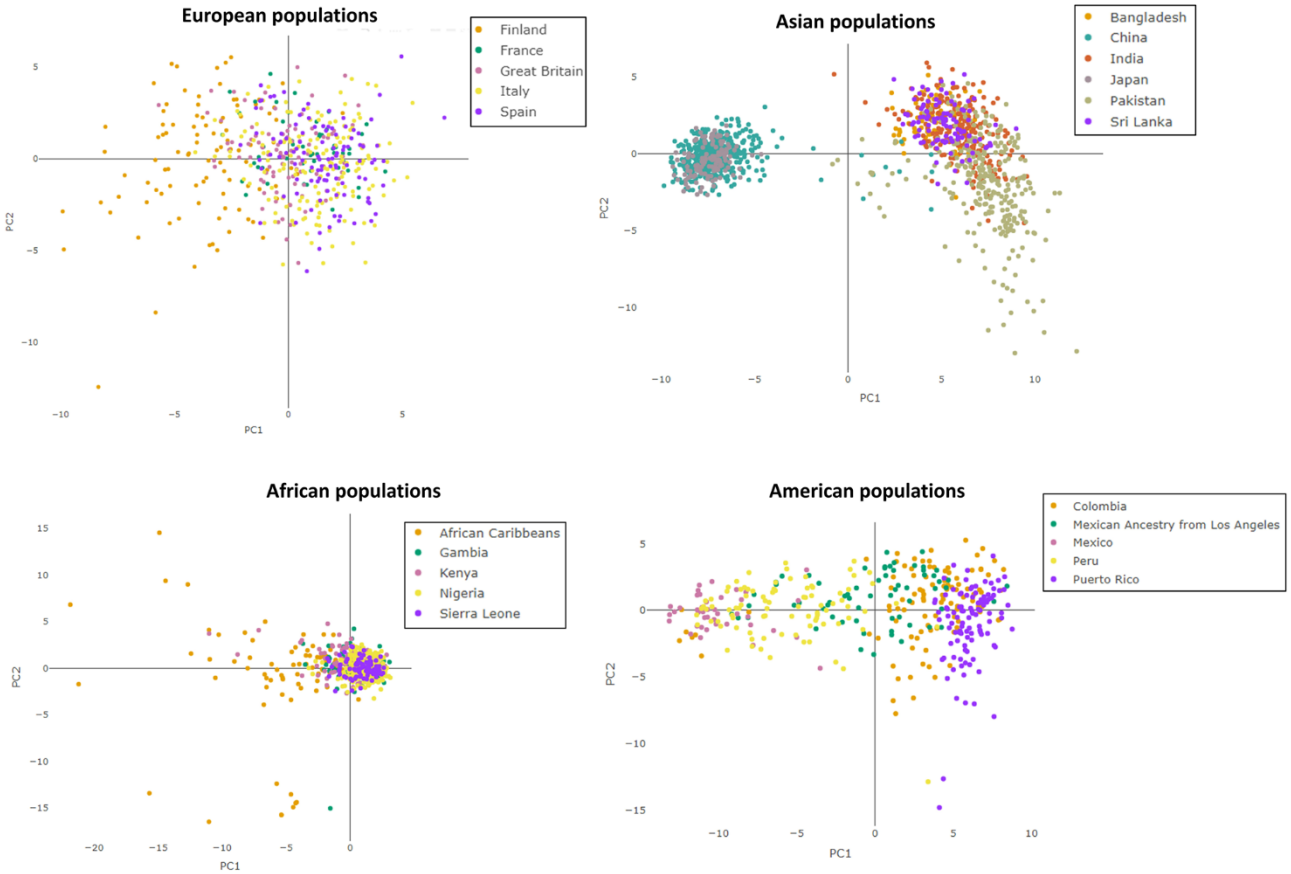

Figure 6. PCA Scores plots showing the PCA models obtained for the different countries and populations tested using the Maplex panel.

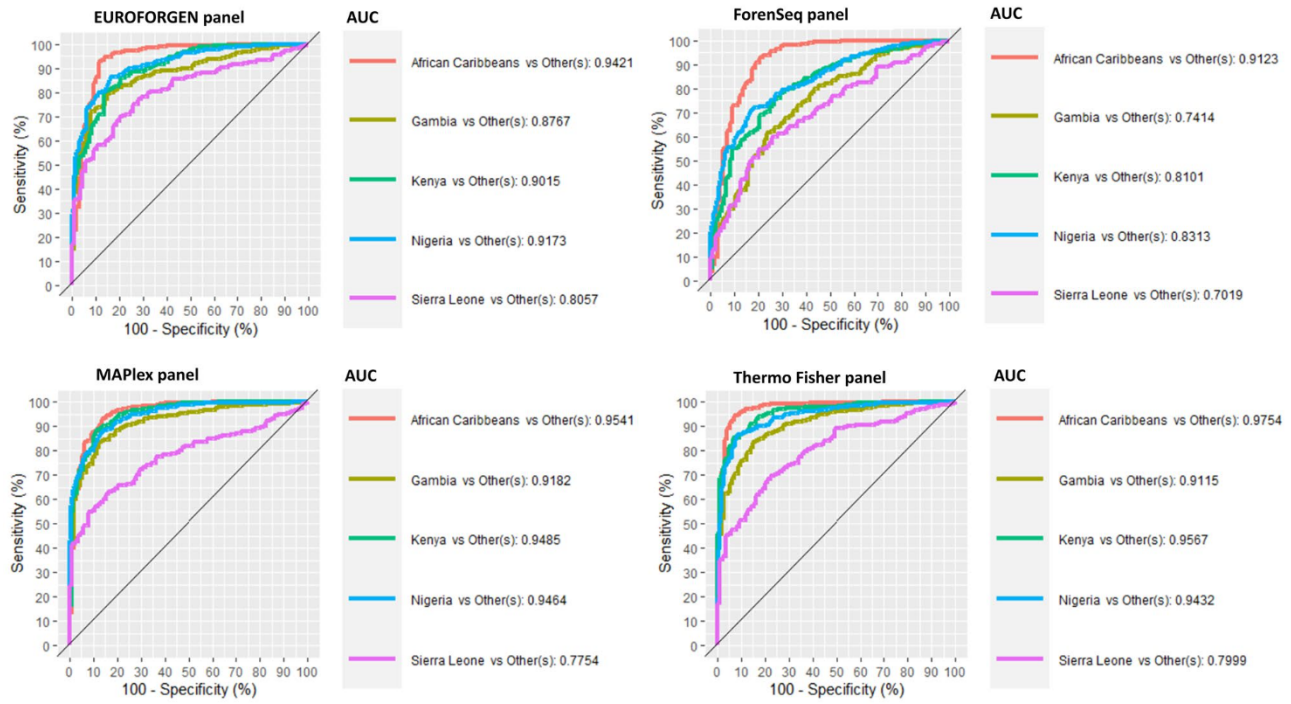

Figure 7. ROC curves, sensitivity, specificity, and AUC values for African countries and populations.

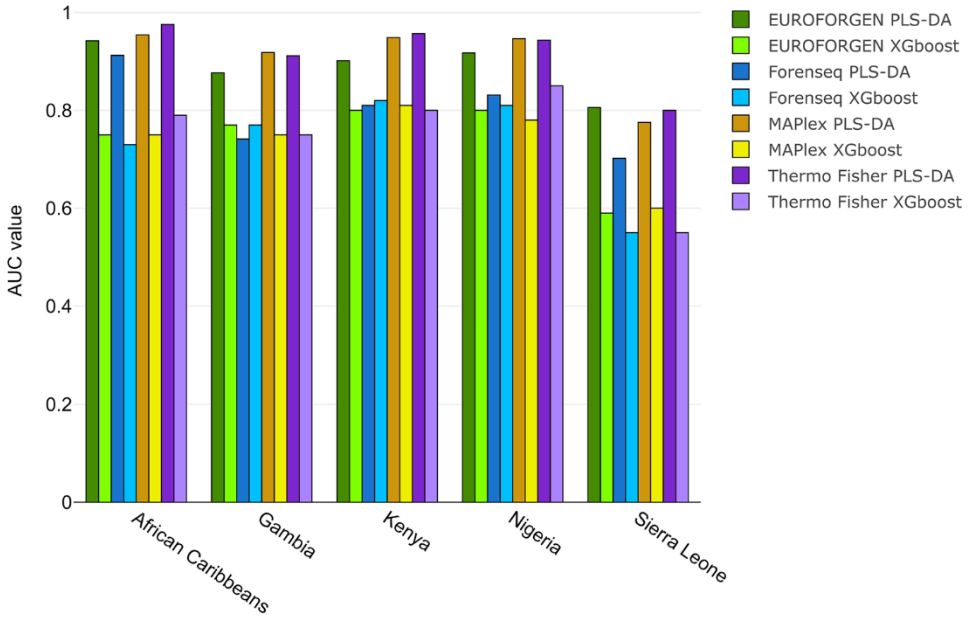

Figure 8. Comparison of AUC values obtained fromPLS-DA and XGBoost model for African population.

outstanding discrimination, respectively, of the SNPs in the panel. Thermo Fisher and MaPlex panel obtained similar results.

Presumably, due to the limited numbers of markers in the panel, the worst classification performances were provided by the ForenSeq panel with an average AUC value of 0.798, the lowest value compared to the other panels. These results can also be assessed from the scores plots reported in Additional file 1: Fig. S6 where several clusters are visible from the PLS-DA models built using the different AIMs panels.

The AUC value very close to 100% observed for the African population in all panels tested (Fig. 3) highlights their outstanding discrimination at the inter-continental level and a slightly less capability, albeit excellent in most of the panels, at intra-continental level (Fig. 7). Indeed, the average AUC values for all panels in African population range from an acceptable discrimination for Forenseq panel (average AUC value = 0.798) to an outstanding discrimination for MaPlex and Thermo Fisher panel with the average AUC values equal to 0.92 and 0.91 respectively.

The XGBoost model was also performed, and Tables S1 in Additional file 1 shows the sensitivity, specificity, and AUC values for African populations.

AUC values of PLS-DA and XGBoost model were compared (Fig. 8).

Interesting AUC values (around 0.9) were observed for African Carribean, Gambian, Kenyan, and Nigerian individuals, while the worst results (0.8 for PLS-DA, 0.6 for XGBoost) were obtained for the subjects from Sierra Leone presumably influenced by the lower number of individuals in the population. Again, the best performances were achieved using PLS-DA modeling.

In the American framework (Fig. 9), no specific panel or model outperformed the others. Good discrimination results were observed using EUROFORGEN and MAPlex panels for the individuals from Mexico and Peru, and Puerto Rico (in all cases, AUC value is higher than 0.97), and Colombia (for MAPlex only with an AUC value of 0.85). On the other hand, the Thermo Fisher panel showed the best results in discriminating the individual of Mexican ancestry living in Los Angeles (US) (AUC value of 0.88), but also ForenSeq panel provided remarkable results (AUC value of 0.84). Thermo Fisher panel also provided reliable classification results (AUC value of 0.98) when dealing with subjects from Puerto Rico (as well as EUROFORGEN (0.97) and MAPlex (0.99) panels). These results can also be observed from the scores plots reported in Additional file 1: Fig. S7, showing several clusters among the tested countries and populations.

In addition, in the American scenario, all panels investigated except MaPlex show AUC values higher than 0.95 at inter-continental level (Fig. 3), and a very slightly less capability of discrimination was observed at inter-continental level with the average AUC values higher than 0.90 for all panels (Fig. 9). Therefore, particular attention should be paid with the MaPlex panel. In this case, the AUC value at inter-continental level is much lower (AUC = 0.77) than the average value obtained at intra-continental level (AUC mean = 0.93), showing a better discrimination at intra-continental level rather than at inter-continental one. This might be because there is a lower variability in the analyzed data (as well as in the number of tested populations) and, in this scenario, the algorithms are capable of predicting and inferring BGA with improved performances.

Tables S2 in additional file 1 shows the sensitivity, specificity, and AUC values of XGBoost model for American population. AUC values of PLS-DA and XGBoost models were compared (Fig. 10).

As shown in Fig. 10, optimal AUC values (around 1 for PLS-DA) were observed when inferring the BGA for individuals from Mexico, Peru, and Puerto Rico, while lower performances (around 0.8 for PLS-DA) were

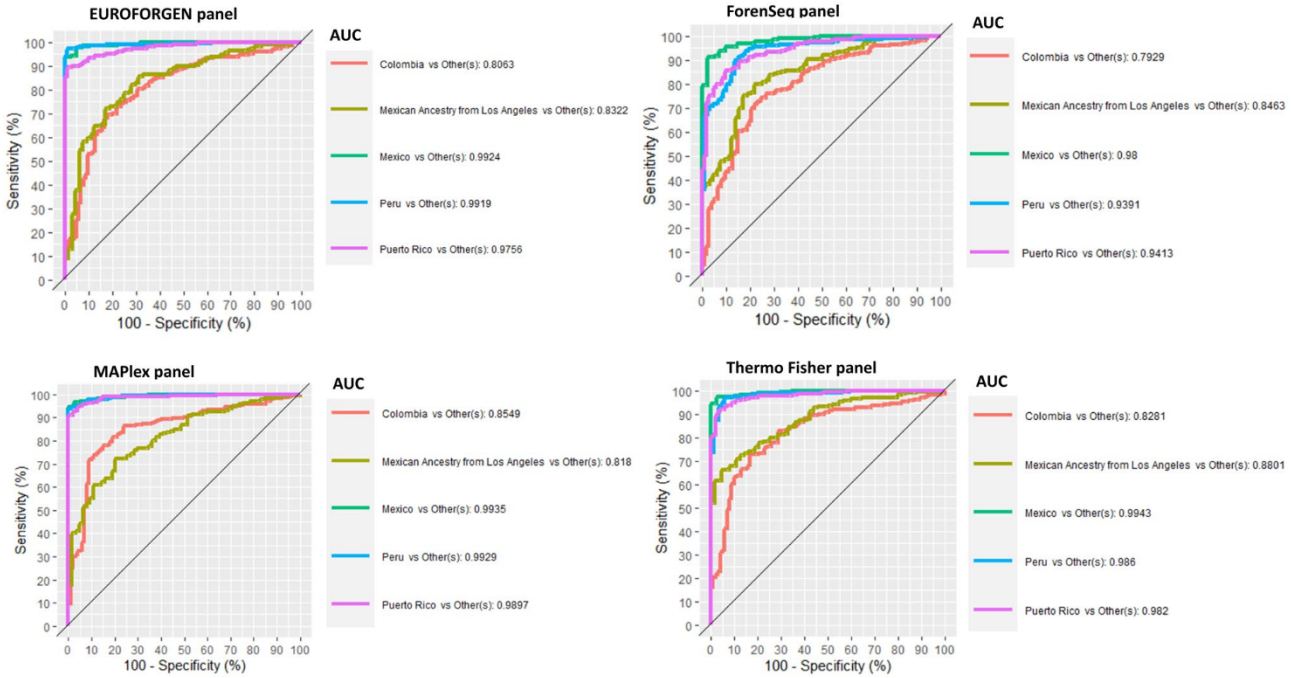

Figure 9. ROC curves, sensitivity, specificity, and AUC values for American countries and populations.

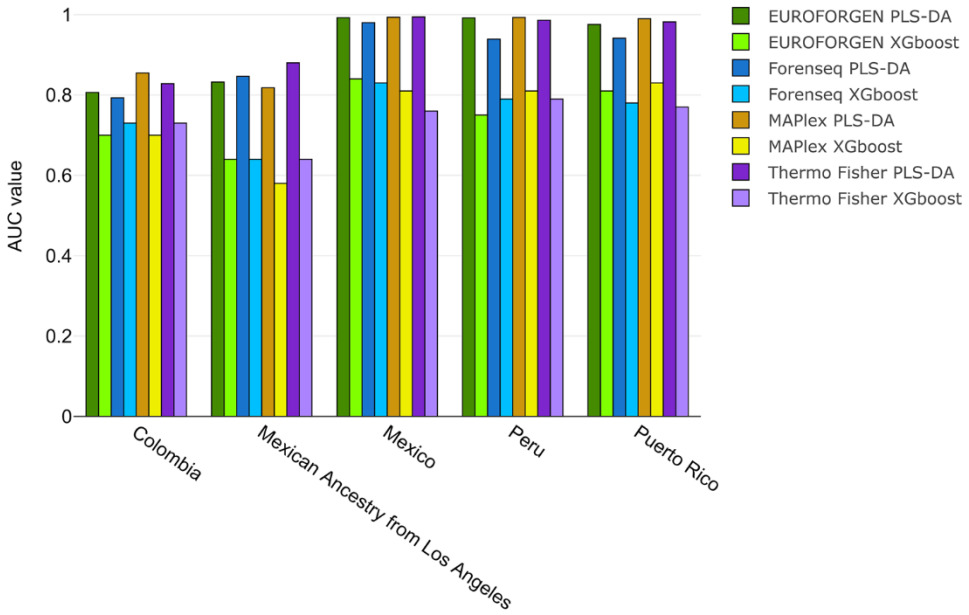

Figure 10. Comparison of AUC values obtained from PLS-DA and XGBoost model for American population.

obtained when evaluating Colombian and Mexican Ancestry from Los Angeles individuals. Again, the best performances were achieved using PLS-DA modeling.

In the Asian framework (Fig. 11), similar results were obtained. On average, the best results were obtained when evaluating the Thermo Fisher and MAPlex panels, especially for the individuals from China, Japan, and Pakistan with AUC values equal to 0.99, 0.98 and 0.95, respectively, for Thermo panel and 0.98, 0.98 and 0.86 for MaPlex panel. Excellent discrimination was achieved also for India, Bangladesh and Sri Lanka with AUC greater than 0.80, showing the ability of these two panels to differentiate sub-populations.

The scores plot provided two separated clusters; the first one consists of China and Japan, while the second cluster reported the individuals from Bangladesh, India, Pakistan, and Sri Lanka (Additional file 1: Fig. S8).

Tables S3 in additional file 1 shows the sensitivity, specificity, and AUC values of XGBoost model for Asian population. AUC values of PLS-DA and XGBoost models were compared (Fig. 12).

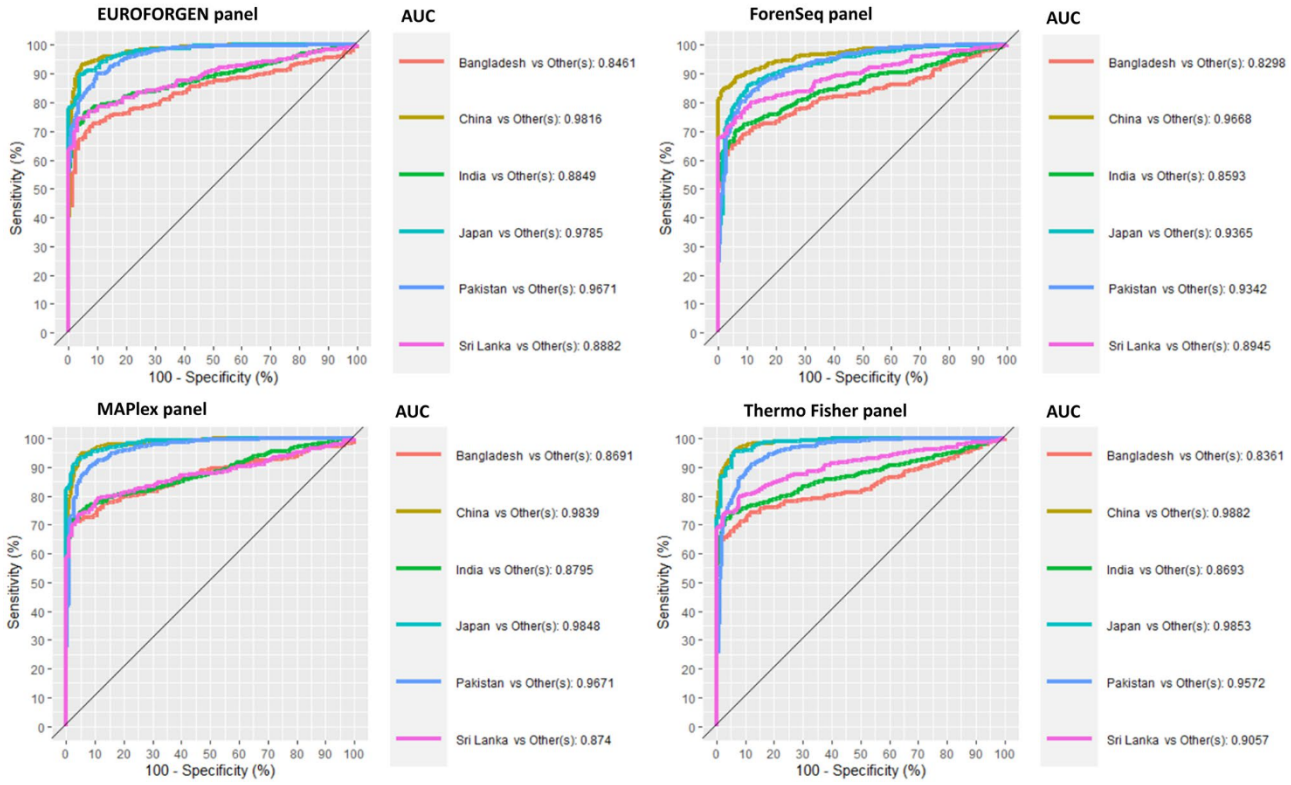

Figure 11. ROC curves, sensitivity, specificity, and AUC values for Asian countries and populations.

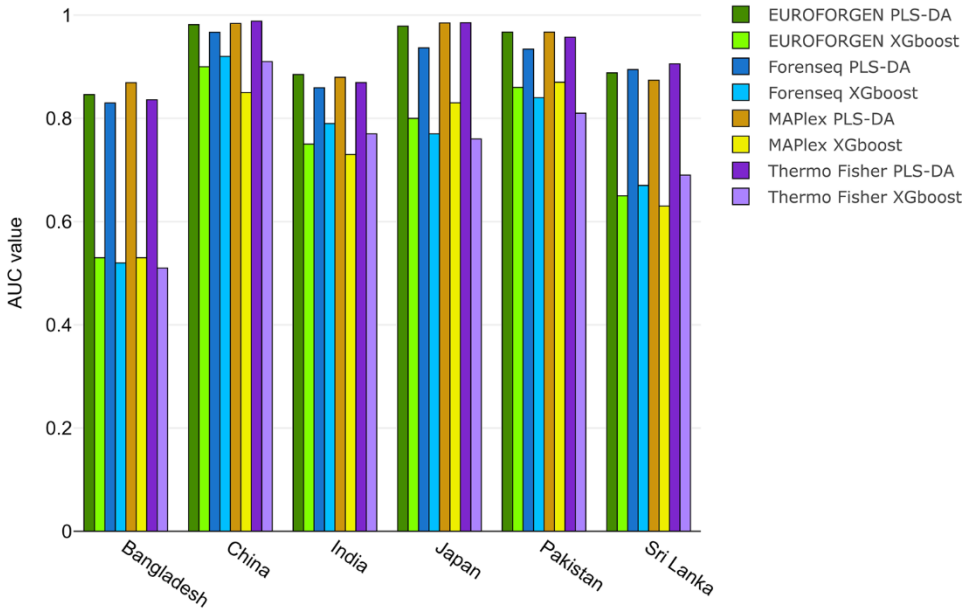

Figure 12. Comparison of AUC values obtained from PLS-DA and XGBoost model for Asian population.

The best AUC values (around 1 for PLS-DA) were obtained when inferring the BGA for individuals from China, Japan, and Puerto Rico, while lower results (around 0.8 for PLS-DA) were obtained when evaluating individuals from Bangladesh, India, and Sri Lanka. The lowest results were showed by the XGBoost model on Bangladesh subjects and, once again, the best performances were achieved with PLS-DA modeling.

Finally, no specific AIMs panel or model outperformed the others when evaluating the European countries and populations except for the ForenSeq panel that presents the worst results, presumably due to the low numbers

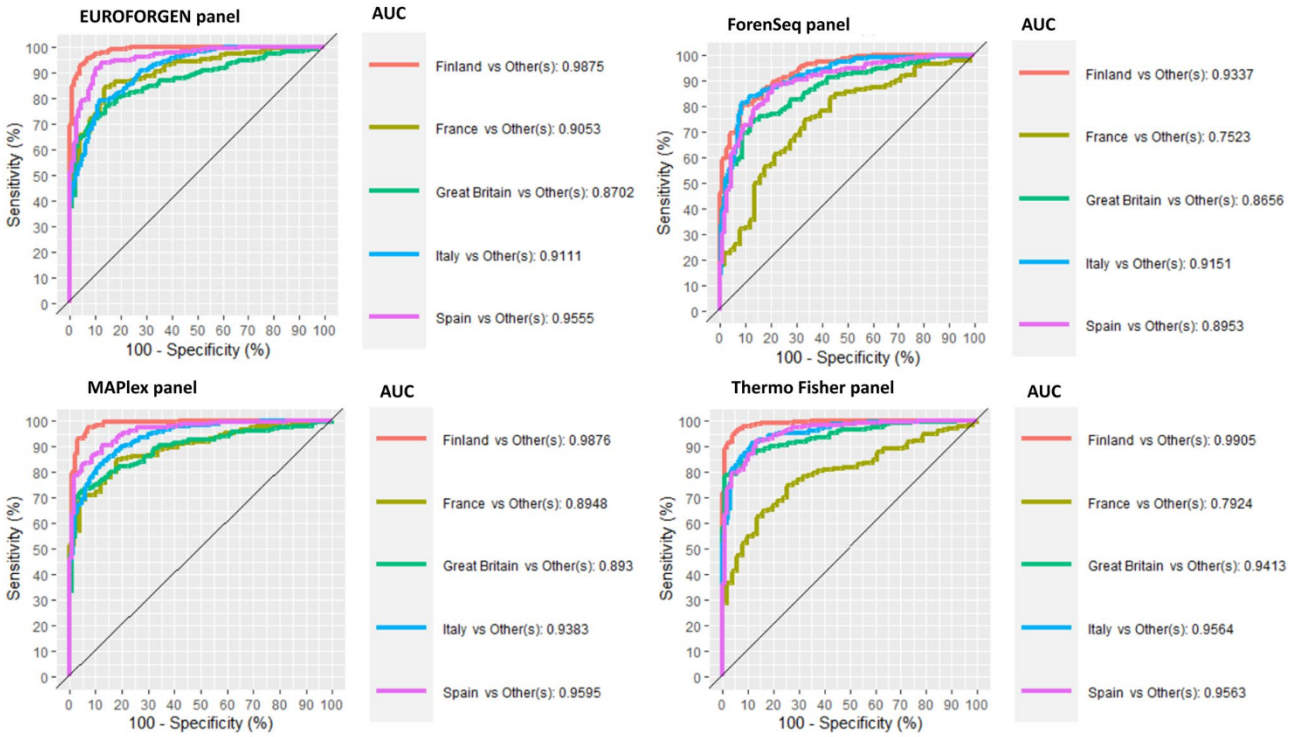

Figure 13. ROC curves, sensitivity, specificity, and AUC values for European countries and populations.

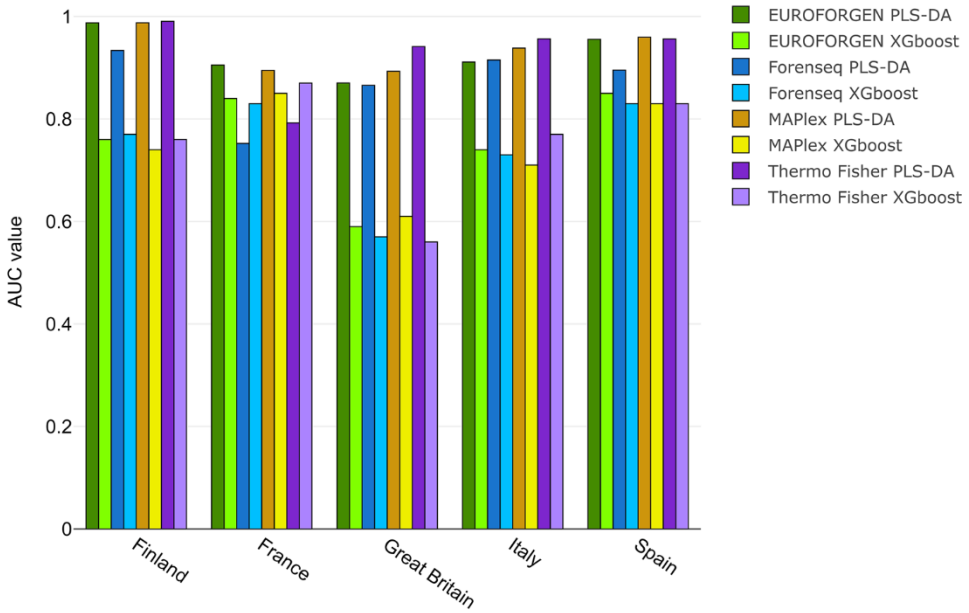

Figure 14. Comparison of AUC values obtained from PLS-DA and XGBoost model for European population.

of markers analyzed. The scores plot provided several separate clusters for all the evaluated populations, and these results can also be observed from the scores plots reported in Additional file 1: Fig. S9.

As shown in Fig. 13, the best discrimination result was achieved for Finland populations (AUC ≥ 0.93) for all panels investigated. It has to be noted that the best results for the French individuals were obtained with EUROFORGEN and MAPlex AIMs panels, while for the other groups (Italians, English, Spanish, and Finns) the results are comparable.

Tables S4 in additional file 1 shows the sensitivity, specificity, and AUC values of XGBoost model for European population. AUC values of PLS-DA and XGBoost models were compared (Fig. 14).

Optimal AUC values (around 0.9–1) were observed for all the PLS-DA models in this scenario, instead of the XGBoost models showing significantly lower results.

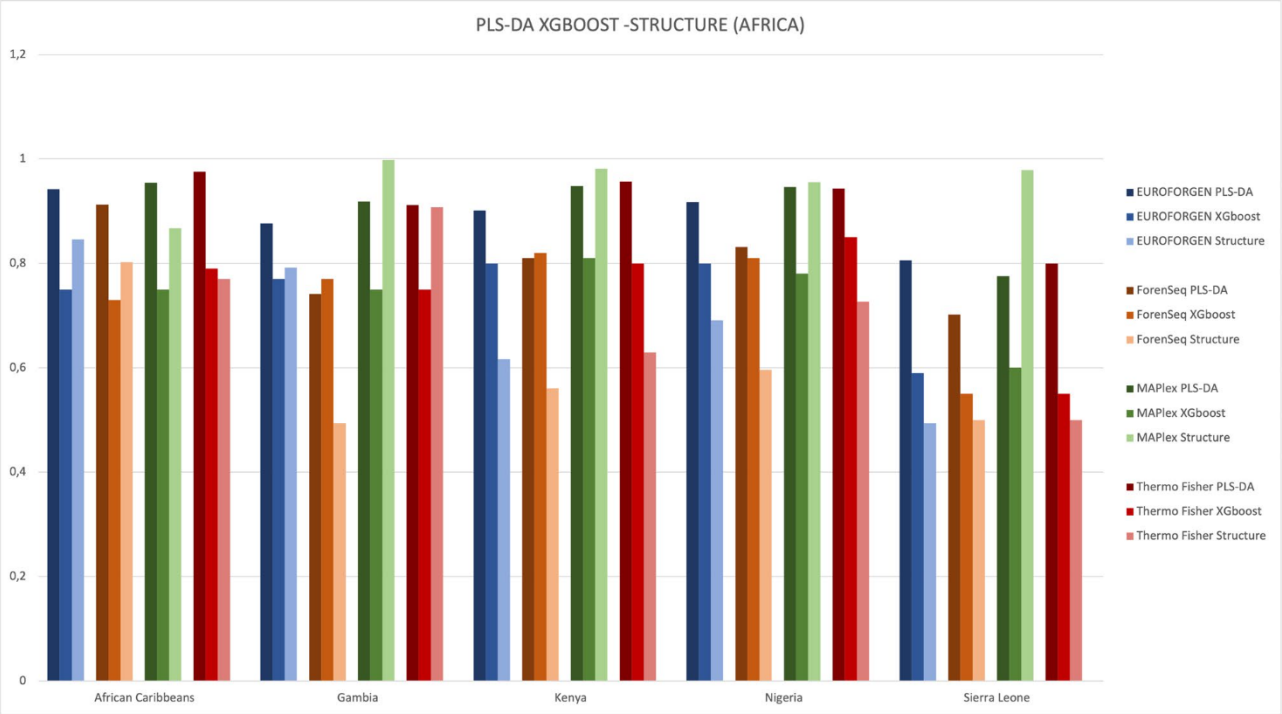

Figure 15. Comparison of AUC values obtained from PLS-DA, XGBoost and STRUCTURE model for African population.

STRUCTURE approach was also compared with PLS-DA and XGBoost model at intra-continental level by evaluating the populations selected for the Africans, as an example. The results in terms of comparison of the AUC values are reported in the Fig. 15. As already observed at inter-continental level, PLS-DA, and in most cases XGBoost, provided, on average, better performance in terms of accuracy when compared to STRUCTURE approach also at intra-continental level.

Comparing the ROC curves of all forensic panels both at inter-continental level and at intra-continental level, a decrease in the accuracy in inferring BGA at intra-continental level was observed. This decrease may be explained by the natural geographical distribution of some populations: "populations that share geographical borders and cultural practices are closely related genetically and these populations show similar genetic patterns"[72], by the SNPs in forensic panels, selected with the aim of discriminating populations at continental level[28,69,73], and by their number which is relatively low compared to that used in other genetic fields through NGS technology.

PLS-DA and XGBoost at intra-continental level provided, on average, better performance in terms of accuracy when compared to STRUCTURE approach. In particular, the obtained results showed that PLS-DA performed better than STRUCTURE at both inter- and intra-continental level. Similar results were achieved by Jombart et al.[43] when using a supervised classification approach like DAPC in comparison with STRUCTURE. Furthermore, both PLS-DA and STRUCTURE methods provide graphical outputs for interpreting the results of the obtained classification models. STRUCTURE provides the results in form of bar plot (being extremely helpful, for instance, when interpreting admixtures) while PLS-DA modelling shows a scatter plot for the tested populations, aimed to evaluate the goodness of the developed classification and allowing to project new individuals into the calculated Scores plots. On the other hand, GenoGeographer approach shows a brilliant use of Likelihood Ratio modelling, since it allows to compare the tested populations and the predictions in terms of $Log_{10}LR$. Similarly, our XGBoost and PLS-DA approaches provide numerical results for the performance of the models (in terms of ROC curves) and the classifications of new individuals (in terms of probability of classification for the new tested individuals).

Conclusions

Ancestry analysis is of increasing relevance to crime investigations in all situations in which few or no investigative leads are available and genetic information about the donor of the trace or skeleton found at the crime-scene could be of help to police investigations to find unknown perpetrators of crime or identify missing persons. Therefore, the present study investigated the application of multivariate data analysis modelling to discriminate and predict the BGA of several populations by evaluating the AIMs markers and panels available in the market for forensic purposes. PLS-DA and XGBoost supervised models drastically improved the traditionally used PCA approach, by supplying satisfactory classification results and showing a capability of BGA discrimination of the diverse forensic panels. Moreover, the comparison between these models with STRUCTURE approach highlighted a better BGA prediction of PLS-DA than STRUCTURE at both inter- and intra-continental level. Different results were observed at inter-continental level only in south Asia, Middle East, and Oceania where

STRUCTURE model seems to work best in almost all panels investigated with the exception of MaPlex panel in South Asia. Despite the high classification performances of PLS-DA, a decrease in the accuracy in inferring BGA was observed for all panels investigated when moving to more geographically restricted populations presumably due to the type of selected SNPs and their limited number in all forensic panels. In addition, particular attention should be paid to the database. Since the analysis of ancestry inference is performed by comparing the sample genotype with one or more known reference population groups, well-characterized databases with high-quality genotyping results of well-defined reference populations are critical. This work represents a proof-of-concept study suggesting the possibility to use supervised algorithms such as PLS-DA or XGBoost (and, eventually, other multivariate models) as a tool for the investigative police forces to estimate the BGA of suspects and persons of interests. Despite the findings, our work does not aim to suggest the use of PLS-DA and XGBoost as improved alternative methods to those involving likelihood-ratio computations and further investigations will be conducted to fully investigate the performance of these approaches and their use for forensic purposes. Future perspective will involve the evaluation of class-modelling ML approaches like SIMCA (Soft-Independent Model of Class Analogy) and the computation of likelihood ratios for each classification; these approaches will allow the forensic expert to obtain interesting information to interpret the results of critical unknown individuals such as admixtures and the cases where AIMs profiles do not belong to any of the included reference populations.

Data availability

All data generated or analysed during this study are included in this article and its supplementary information files.

Received: 14 October 2021; Accepted: 13 April 2022
Published online: 28 May 2022

References

1. Elhaik, E. *et al.* Geographic population structure analysis of worldwide human populations infers their biogeographical origins. *Nat Commun.* https://doi.org/10.1038/ncomms4513 (2014).
2. Halder, I. *et al.* Biogeographic ancestry, self-identified race, and admixture-phenotype associations in the Heart SCORE Study. *Am. J. Epidemiol.* **176**, 146–155. https://doi.org/10.1093/aje/kwr518 (2012).
3. Shriver, M. D. *et al.* Skin pigmentation, biogeographical ancestry and admixture mapping. *Hum. Genet.* **112**, 387–399. https://doi.org/10.1007/s00439-002-0896-y (2003).
4. Rosenberg, N. A. *et al.* Genetic structure of human populations. *Science* **298**, 2381–2385. https://doi.org/10.1126/science.1078311 (2002).
5. Rosenberg, N. A., Li, L. M., Ward, R. & Pritchard, J. K. Informativeness of genetic markers for inference of ancestry. *Am. J. Hum. Genet.* **73**, 1402–1422 (2003).
6. Qu, S. *et al.* Establishing a second-tier panel of 18 ancestry informative markers to improve ancestry distinctions among Asian populations. *Forensic Sci. Int.* **41**, 159–167. https://doi.org/10.1016/j.fsigen.2019.05.001 (2019).
7. Phillips, C. *et al.* Development of a novel forensic STR multiplex for ancestry analysis and extended identity testing. *Electrophoresis* **34**, 1151–1162. https://doi.org/10.1002/elps.201200621 (2013).
8. Phillips, C. Forensic genetic analysis of bio-geographical ancestry. *Forensic Sci. Int.* **18**, 49–65. https://doi.org/10.1016/j.fsigen.2015.05.012 (2015).
9. Santos, C. *et al.* Completion of a worldwide reference panel of samples for an ancestry informative Indel assay. *Forensic Sci. Int.* **17**, 75–80. https://doi.org/10.1016/j.fsigen.2015.03.011 (2015).
10. Gettings, K. B. *et al.* A 50-SNP assay for biogeographic ancestry and phenotype prediction in the U.S. population. *Forensic Sci. Int.* **8**, 101–108. https://doi.org/10.1016/j.fsigen.2013.07.010 (2014).
11. Pakstis, A. J. *et al.* 52 additional reference population samples for the 55 AISNP panel. *Forensic Sci Int Genet.* **19**, 269–271. https://doi.org/10.1016/j.fsigen.2015.08.003 (2015).
12. Kidd, J. R. *et al.* Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples. *Investig Genet.* https://doi.org/10.1186/2041-2223-2-1 (2011).
13. Oldoni, F. *et al.* Population genetic data of 74 microhaplotypes in four major U.S. population groups. *Forensic Sci. Int.* **49**, 102398. https://doi.org/10.1016/j.fsigen.2020.102398 (2020).
14. Suárez, D. *et al.* Ancestry analysis using autosomal SNPs in northern South America, reveals interpretation differences between an AIM panel and an identification panel. *Forensic Sci. Int.* **326**, 110934 (2021).
15. Guanglin, H. *et al.* Massively parallel sequencing of 165 ancestry-informative SNPs and forensic biogeographical ancestry inference in three southern Chinese Sinitic/Tai-Kadai populations. *Forensic Sci. Int.* **52**, 102475. https://doi.org/10.1016/J.FSIGEN.2021.102475 (2021).
16. Kuo, Y. H., Vanderzwan, S. L., Kasprowicz, A. E. & Sacks, B. N. Using ancestry-informative SNPs to quantify introgression of European alleles into North American red foxes. *J. Hered.* **110**, 782–792. https://doi.org/10.1093/JHERED/ESZ053 (2019).
17. Pereira, V. *et al.* Evaluation of the precision of ancestry inferences in South American Admixed populations. *Front Genet.* https://doi.org/10.3389/FGENE.2020.00966 (2020).
18. Truelsen, D., Pereira, V., Phillips, C., Morling, N. & Børsting, C. Evaluation of a custom GeneRead^{TM} massively parallel sequencing assay with 210 ancestry informative SNPs using the Ion S5^{TM} and MiSeq platforms. *Forensic Sci. Int.* **50**, 102411. https://doi.org/10.1016/J.FSIGEN.2020.102411 (2021).
19. Simayijiang, H., Børsting, C., Tvedebrink, T. & Morling, N. Analysis of Uyghur and Kazakh populations using the Precision ID Ancestry Panel. *Forensic Sci. Int.* **43**, 102144. https://doi.org/10.1016/J.FSIGEN.2019.102144 (2019).
20. Pakstis, A. J. *et al.* The population genetics characteristics of a 90 locus panel of microhaplotypes. *Hum. Genet.* **140**, 1753–1773. https://doi.org/10.1007/S00439-021-02382-0 (2021).
21. Cheung, E. Y. Y., Phillips, C., Eduardoff, M., Lareu, M. V. & McNevin, D. Performance of ancestry-informative SNP and microhaplotype markers. *Forensic Sci. Int.* **43**, 102141. https://doi.org/10.1016/J.FSIGEN.2019.102141 (2019).
22. Bulbul, O. *et al.* Ancestry inference of 96 population samples using microhaplotypes. *Int. J. Leg. Med.* **132**, 703–711. https://doi.org/10.1007/S00414-017-1748-6 (2018).
23. de la Puente, M. *et al.* Development and evaluation of the ancestry informative marker panel of the VISAGE basic tool. *Genes* **12**, 1284. https://doi.org/10.3390/GENES12081284 (2021).
24. Xiao-Ye, J. *et al.* Development a multiplex panel of AISNPs, multi-allelic InDels, microhaplotypes and Y-SNP/InDel loci for multiple forensic purposes via the NGS. *Electrophoresis* https://doi.org/10.1002/ELPS.202100253 (2021).

25. Zhu, Q. *et al.* A targeted ancestry informative InDels panel on capillary electrophoresis for ancestry inference in Asian populations. *Electrophoresis* **42**, 1605–1613. https://doi.org/10.1002/ELPS.202100016 (2021).
26. Al-Asfi, M. *et al.* Assessment of the precision ID ancestry panel. *Int. J. Leg. Med.* **132**, 1581–1594. https://doi.org/10.1007/s00414-018-1785-9 (2018).
27. Jäger, A. C. *et al.* Developmental validation of the MiSeq FGx forensic genomics system for targeted next generation sequencing in forensic DNA casework and database laboratories. *Forensic Sci. Int.* **28**, 52–70. https://doi.org/10.1016/j.fsigen.2017.01.011 (2017).
28. Eduardoff, M. *et al.* Inter-laboratory evaluation of the EUROFORGEN Global ancestry-informative SNP panel by massively parallel sequencing using the Ion PGM™. *Forensic Sci. Int.* **23**, 178–189. https://doi.org/10.1016/j.fsigen.2016.04.008 (2016).
29. Phillips, C. *et al.* MAPlex: A massively parallel sequencing ancestry analysis multiplex for Asia-Pacific populations. *Forensic Sci. Int.* **42**, 213–226. https://doi.org/10.1016/j.fsigen.2019.06.022 (2019).
30. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959. https://doi.org/10.1093/GENETICS/155.2.945 (2000).
31. Tvedebrink, T., Eriksen, P. S., Mogensen, H. S. & Morling, N. GenoGeographer: A tool for genogeographic inference. *Forensic Sci. Int.* **6**, e463–e465 (2017).
32. Mogensen, H. S., Tvedebrink, T., Børsting, C., Pereira, V. & Morling, N. Ancestry prediction efficiency of the software GenoGeographer using a z-score method and the ancestry informative markers in the Precision ID Ancestry Panel. *Forensic Sci. Int.* **44**, 102154. https://doi.org/10.1016/J.FSIGEN.2019.102154 (2020).
33. Leardi, R., Seasholtz, M. B. & Pell, R. J. Variable selection for multivariate calibration using a genetic algorithm: Prediction of additive concentrations in polymer films from Fourier transform-infrared spectral data. *Anal. Chim. Acta.* **461**, 189–200 (2002).
34. Jouan-Rimbaud, D., Massart, D. L., Leardi, R. & De Noord, O. E. Genetic algorithms as a tool for wavelength selection in multivariate calibration. *Anal. Chem.* **67**, 4295–4301. https://doi.org/10.1021/ac00119a015 (1995).
35. Kowalski, B. R. & Seasholtz, M. B. Recent developments in multivariate calibration. *J. Chemom.* **5**, 129–145. https://doi.org/10.1002/cem.1180050303 (1991).
36. Zadora, G., Neocleous, T. & Aitken, C. G. G. Recent developments in likelihood ratio models for multivariate compositional data. *Sci. Justice.* **50**, 30. https://doi.org/10.1016/j.scijus.2009.11.023 (2010).
37. Bozza, S., Broséus, J., Esseiva, P. & Taroni, F. Bayesian classification criterion for forensic multivariate data. *Forensic Sci. Int.* **244**, 295–301 (2014).
38. Aitken, C. G. G. & Lucy, D. Evaluation of trace evidence in the form of multivariate data. *J. R. Stat. Soc. Ser. C* **53**, 109–122. https://doi.org/10.1046/j.0035-9254.2003.05271.x (2004).
39. Kumar, N., Bansal, A., Sarma, G. S. & Rawal, R. K. Chemometrics tools used in analytical chemistry: An overview. *Talanta* **123**, 186–199 (2014).
40. Geladi, P. Analysis of multi-way (multi-mode) data. *Chemom. Intell. Lab. Syst.* **7**, 11–30 (1989).
41. Rijk, J. C. W. *et al.* Metabolomics approach to anabolic steroid urine profiling of bovines treated with prohormones. *Anal. Chem.* **81**, 6879–6888. https://doi.org/10.1021/ac900874m (2009).
42. Bro, R. *Multi-way Analysis in the Food Industry Models, Algorithms, and Applications* (1998).
43. Jombart, T., Devillard, S. & Balloux, F. Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC Genet.* **11**, 1–15. https://doi.org/10.1186/1471-2156-11-94/FIGURES/9 (2010).
44. Bro, R. & Smilde, A. K. Principal component analysis. *Anal. Methods* **6**, 2812–2831 (2014).
45. Ballabio, D. & Consonni, V. Classification tools in chemistry Part 1: linear models PLS-DA. *Anal Methods* **5**, 3790–3798 (2013).
46. Alladio, E. *et al.* A multivariate statistical approach for the estimation of the ethnic origin of unknown genetic profiles in forensic genetics. *Forensic Sci. Int.* **45**, 102299 (2020).
47. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature.* **526**, 68–74. https://doi.org/10.1038/nature15393 (2015).
48. Bergström, A. *et al.* Insights into human genetic variation and population history from 929 diverse genomes. *Science* https://doi.org/10.1126/science.aay5012 (2020).
49. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2015).
50. Dancho, M. *correlationfunnel: Speed Up Exploratory Data Analysis (EDA) with the Correlation Funnel* (2020).
51. Wickham, H., François, R., Henry, L., Müller, K. *dplyr: A Grammar of Data Manipulation* (2020).
52. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag, 2016).
53. Kucheryavskiy, S. mdatools: R package for chemometrics. *Chemom. Intell. Lab. Syst.* **198**, 103937 (2020).
54. Rohart, F., Gautier, B., Singh, A. & Lê Cao, K. A. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput Biol.* **13**, e1005752 (2017).
55. Bischl, B. *et al.* {mlr}: machine learning in R. *J. Mach. Learn. Res.* **17**, 1–5 (2016).
56. Sievert, C. *plotly for R* (2018).
57. Mehmood, T., Liland, K. H., Snipen, L. & Sæbø, S. A review of variable selection methods in Partial Least Squares Regression. *Chemom. Intell. Lab. Syst.* **118**, 62–69 (2012).
58. Chen, T. *et al. xgboost: Extreme Gradient Boosting* (2021).
59. Statheropoulos, M., Smaragdis, E., Tzamtzisa, N. & Georgakopoulosb, C. Principal component analysis for resolving coeluting substances in gas chromatography-mass spectrometry doping control analysis. *Anal. Chim. Acta.* **2670**, 53–61 (1996).
60. Smoliński, A., Walczak, B. & Einax, J. Exploratory analysis of data sets with missing elements and outliers. *Chemosphere* **49**, 233–245 (2002).
61. Stanimirova, I., Walczak, B., Massart, D. L. & Simeonov, V. A comparison between two robust PCA algorithms. *Chemom. Intell. Lab. Syst.* **71**, 83–95 (2004).
62. Ralston, P., Depuy, G. & Graham, J. H. Graphical enhancement to support PCA-based process monitoring and fault diagnosis. *ISA Trans.* **43**, 639–653 (2004).
63. Godoy, J. L., Vega, J. R. & Marchetti, J. L. Relationships between PCA and PLS-regression. *Chemom. Intell. Lab. Syst.* **130**, 182–191 (2014).
64. Abdi, H. Partial least square regression (PLS regression). *Encycl. Res. Methods Soc. Sci.* **6**(4), 792–795 (2003).
65. Geladi, P. & Kowalski, B. R. Partial least-squares regression: A tutorial. *Anal. Chim. Acta* **185**, 1–17 (1986).
66. Wold, S., Sjöström, M. & Eriksson, L. PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **58**, 109–130 (2001).
67. Guang, P. *et al.* Blood-based FTIR-ATR spectroscopy coupled with extreme gradient boosting for the diagnosis of type 2 diabetes: A STARD compliant diagnosis research. *Medicine* **99**, e19657 (2020).
68. Hosmer, D. & Lemeshow, S. *Applied Logistic Regression, 3rd Edition.* (Wiley, 2013). https://www.wiley.com/en-us/Applied+Logistic+Regression%2C+3rd+Edition-p-9780470582473. Accessed 17 Jun 2021.
69. Pereira, V., Mogensen, H. S., Børsting, C. & Morling, N. Evaluation of the Precision ID Ancestry Panel for crime case work: A SNP typing assay developed for typing of 165 ancestral informative markers. *Forensic Sci. Int.* **28**, 138–145 (2017).
70. Churchill, J. D., Novroski, N. M. M., King, J. L., Seah, L. H. & Budowle, B. Population and performance analyses of four major populations with Illumina's FGx Forensic Genomics System. *Forensic Sci. Int.* **30**, 81–92. https://doi.org/10.1016/j.fsigen.2017.06.004 (2017).
71. Santos, C. *et al.* Inference of ancestry in forensic analysis II: Analysis of genetic data. *Methods Mol. Biol.* **1420**, 255–285. https://doi.org/10.1007/978-1-4939-3597-0_19 (2016).

72. Bulbul, O. *et al.* Improving ancestry distinctions among Southwest Asian populations. *Forensic Sci. Int.* **35**, 14–20 (2018).
73. Ramani, A. *et al.* Differentiation of Asian population samples using the Illumina ForenSeq kit. *Forensic Sci. Int. Genet.* https://doi.org/10.1016/j.fsigen.2020.102318 (2020).
74. Xavier, C. *et al.* Forensic evaluation of the Asia Pacific ancestry-informative MAPlex assay. *Forensic Sci. Int.* **48**, 102344 (2020).
75. Mizuno, F., Naka, I., Ueda, S., Ohashi, J. & Kurosaki, K. The number of SNPs required for distinguishing Japanese from other East Asians. *Leg. Med. (Tokyo).* https://doi.org/10.1016/J.LEGALMED.2021.101849 (2021).
76. Sun, K. *et al.* Evaluation of 12 Multi-InDel markers for forensic ancestry prediction in Asian populations. *Forensic Sci. Int.* https://doi.org/10.1016/J.FSIGEN.2019.102155 (2019).
77. Schlebusch, C. M. & Jakobsson, M. Tales of human migration, admixture, and selection in Africa. *Annu. Rev. Genom. Hum. Genet.* **19**, 405–428. https://doi.org/10.1146/annurev-genom-083117-021759 (2018).
78. Secolin, R. *et al.* Distribution of local ancestry and evidence of adaptation in admixed populations. *Sci. Rep.* **9**, 1–12. https://doi.org/10.1038/s41598-019-50362-2 (2019).
79. Glusman, G., Mauldin, D. E., Hood, L. E. & Robinson, M. Ultrafast comparison of personal genomes via precomputed genome fingerprints. *Front. Genet.* https://doi.org/10.3389/fgene.2017.00136 (2017).
80. Haber, M. *et al.* Genetic evidence for an origin of the Armenians from Bronze Age mixing of multiple populations. *Eur. J. Hum. Genet.* **24**, 931–936. https://doi.org/10.1038/ejhg.2015.206 (2016).
81. Scott, E. M. *et al.* Characterization of greater middle eastern genetic variation for enhanced disease gene discovery. *Nat. Genet.* **48**, 1071–1079. https://doi.org/10.1038/ng.3592 (2016).
82. Tay, G. K., Henschel, A., Daw Elbait, G. & Al Safar, H. S. Genetic diversity and low stratification of the population of the United Arab Emirates. *Front Genet.* https://doi.org/10.3389/fgene.2020.00608 (2020).
83. Palstra, F. P., Heyer, E. & Austerlitz, F. Statistical inference on genetic data reveals the complex demographic history of human populations in Central Asia. *Mol. Biol. Evol.* **32**, 1411–1424. https://doi.org/10.1093/molbev/msv030 (2015).

Author contributions

Conceptualization, E.P.; Methodology, E.A; Formal Analysis, B.P., G.C., E.A.; Investigation, E.P.; Data Curation: E.P., E.A.; Visualization, E.P., E.A.; Writing—Original Draft Preparation, E.P., E.A.; Writing—Review & Editing, E.P., E.A. All authors have read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-12903-0.

Correspondence and requests for materials should be addressed to E.P.

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

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