IRIS AperTO

UNIVERSITÀ DEGLI STUDI DI TORINO

AperTO - Archivio Istituzionale Open Access dell'Università di Torino

A multivariate statistical approach for the estimation of the ethnic origin of unknown genetic profiles in forensic genetics

This is a pre print version of the following article:

Original Citation:

Availability:

This version is available http://hdl.handle.net/2318/1727847 since 2025-02-24T21:07:56Z

Published version:

DOI:10.1016/j.fsigen.2019.102209

Terms of use:

Open Access

Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)

13 September 2026

Journal Pre-proof

A multivariate statistical approach for the estimation of the ethnic origin of unknown genetic profiles in forensic genetics

Eugenio Alladio, Chiara Della Rocca, Filippo Barni, Jean-Michel Dugoujon, Paolo Garofano, Ornella Semino, Andrea Berti, Andrea Novelletto, Marco Vincenti, Fulvio Cruciani

PII: S1872-4973(19)30361-8

DOI: https://doi.org/10.1016/j.fsigen.2019.102209

Reference: FSIGEN 102209

To appear in: *Forensic Science International: Genetics*

Received Date: 2 August 2019

Revised Date: 11 October 2019

Accepted Date: 22 November 2019

Please cite this article as: { doi: https://doi.org/

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2019 Published by Elsevier.

Journal Pre-proof

A multivariate statistical approach for the estimation of the ethnic origin of unknown genetic profiles in forensic genetics

Authors:
Eugenio Alladio[a,b,c*], Chiara Della Rocca[d], Filippo Barni[a], Jean-Michel Dugoujon[e], Paolo Garofano[c], Ornella Semino[f], Andrea Berti[a], Andrea Novelletto[g], Marco Vincenti[b,c], Fulvio Cruciani[d,h]

Affiliations:
[a] Reparto CC Investigazioni Scientifiche di Roma, Sezione di Biologia, Viale Tor di Quinto 119, 00191 Roma, Italy.
[b] Dipartimento di Chimica, Università degli Studi di Torino, Via P. Giuria 7, 10125 Torino, Italy.
[c] Centro Regionale Antidoping e di Tossicologia "A. Bertinaria" di Orbassano (Torino), Regione Gonzole 10/1, 10030 Orbassano (Torino), Italy.
[d] Dipartimento di Biologia e Biotecnologie "Charles Darwin", Sapienza Università di Roma, Piazzale Aldo Moro 5, 00185 Roma, Italy.
[e] Centre National de la Recherche Scientifique (CNRS) and Université Toulouse III - Paul Sabatier, 118, route de Narbonne, 31062 Toulouse Cedex 9, France.
[f] Dipartimento di Biologia e Biotecnologie "L. Spallanzani", Università degli Studi di Pavia, Via Adolfo Ferrata 9, 27100 Pavia, Italy.
[g] Dipartimento di Biologia, Università degli Studi di Roma "Tor Vergata", Via della Ricerca Scientifica, 1, 00133 Roma, Italy.
[h] Istituto di Biologia e Patologia Molecolari, Consiglio Nazionale delle Ricerche, Rome, Italy

*corresponding author

Eugenio Alladio, PhD

Reparto CC Investigazioni Scientifiche di Roma, Sezione di Biologia

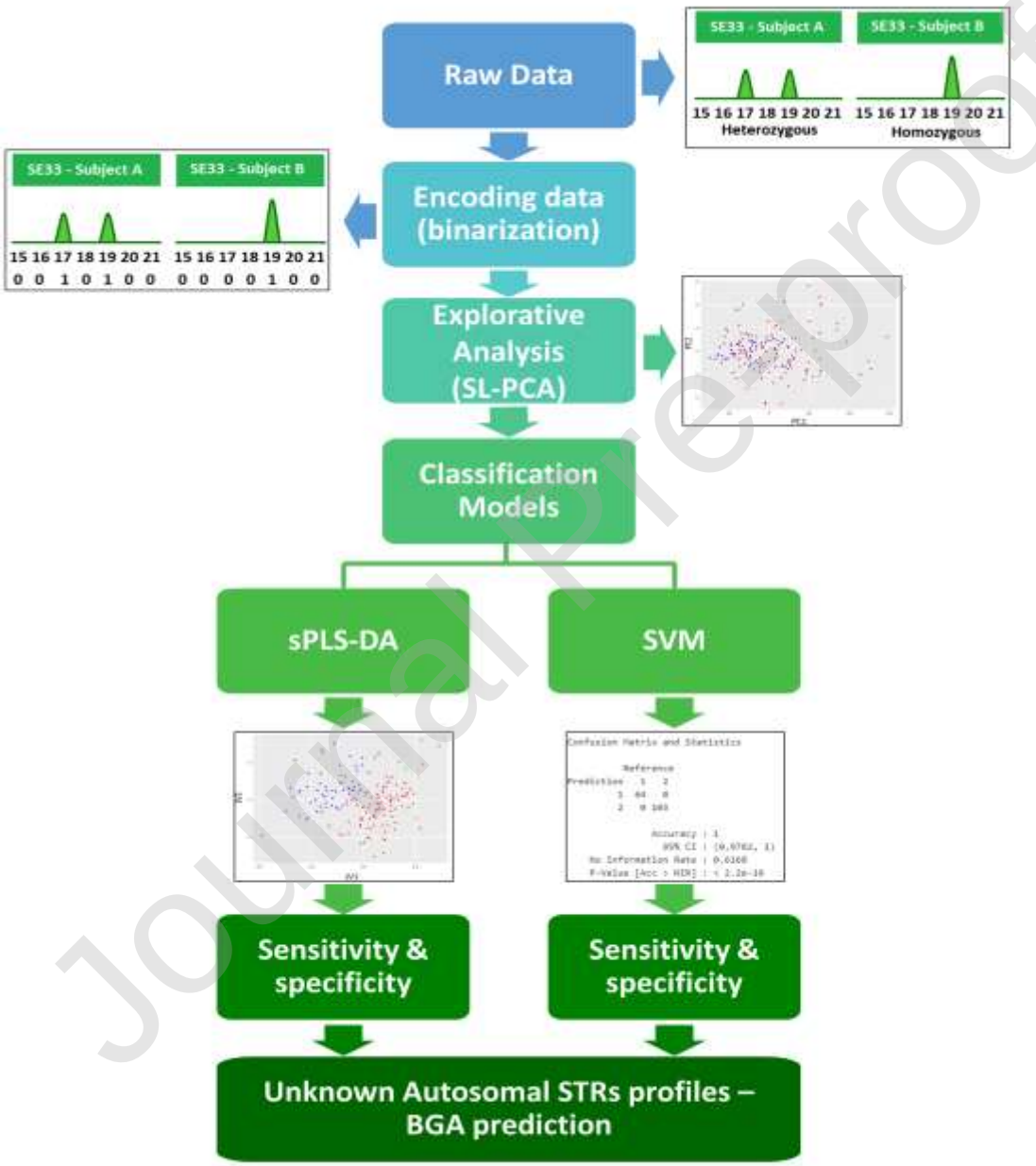
Viale Tor di Quinto 119, 00191 Roma, Italy.

eugenio.alladio@unito.it

eugenio.alladio@carabinieri.it

Journal Pre-proof

Graphical abstract



Journal Pre-proof

Journal Pre-proof

Highlights

- Multivariate protocols for estimating BGA using STRs DNA data are developed;
- Principal Components Analysis is adopted as exploratory approach;
- Partial Least Squares Discriminant Analysis and Support Vector Machines are used;
- Efficient discriminant models are obtained for multiple populations;
- This approach can easily test DNA STRs genotypes from unknown individuals.

1. Introduction

DNA profiling of biological evidence such as those recovered from crime scenes, mass-disaster areas or missing person investigations is one of the most challenging topics in forensic sciences[1–3]. Through the years, DNA typing has been more and more employed, exploiting large sets of genetic markers that can be simultaneously analyzed on a single biological sample or trace, even if containing only a few copies of DNA.

In the last decade, because of continuous technical developments in forensic genetics, DNA analysis moved towards the so-called Next Generation Sequencing (NGS) or Massive Parallel Sequencing (MPS). Currently, this technology enables genotyping at a large number of Short Tandem Repeats (STRs) loci in addition to an ever growing number of further markers such as, for example, autosomal and Y-chromosome Single Nucleotide Polymorphisms (SNPs) and mitochondrial DNA (mtDNA) variants[4]. Nowadays, STRs markers are widely utilized for personal identification in the interpretation process of single source samples and DNA mixtures collected e.g. during crime scene investigation activities[5–7]. On the other hand, the more evolutionarily stable SNPs, in the biparental and uniparental portions of the genome, are being used to infer the biogeographical ancestry and ethnic origins (generally named as BGA) of individuals and degraded samples [8–11]. Up to date, while autosomal STRs markers are the elective tool for

Journal Pre-proof

personal identification, they have been poorly employed as Ancestry Informative Markers (AIMs) as STR alleles equal in state occur in diverse populations, mostly because of recurrent mutation (homoplasy).

Bayesian statistics have been applied to estimate the ethnic affiliation of unknown genetic profiles[12,13] obtained with autosomal STRs in well-known software such as STRUCTURE[14], the Snipper App suite[15] and PopAffiliator 2[16]. These approaches perform Bayesian evaluations by inferring the relationships between the allele frequencies of specific populations and the alleles observed in the individuals, which are recognized as part of such populations. This is done by computing the likelihood values of membership to each of the tested population groups, according to their relative allele frequencies. An advantage of these methodologies is that prior information about the samples can be considered during the advancement the analysis[17]. In the case of multi-locus genotypes, the power to obtain large amounts of data from a single biological sample requires appropriate statistical strategies to extract as precise information as possible regarding its ancestry. In this context, multivariate data analysis techniques may provide useful advantages to infer ethnic affiliation or ancestry of unknown subjects' genetic profiles. These methods may simultaneously perform specific and sensitive discriminations among different groups. Software based on Likelihood Ratios (LR) traditionally involve the comparison of only two alternative hypotheses, while multivariate techniques may efficiently evaluate several population groups together. However, the likelihood-based methods for BGA estimation overcome this issue by computing the likelihood of membership to each of the populations under exam[17,18]. In the present study, we employ multivariate methodologies such as Sparse and Logistic Principal Component Analysis (SL-PCA)[19], Sparse Partial Least Squares-Discriminant Analysis (sPLS-DA)[20–22] and Support Vector Machines (SVM)[23–25] on autosomal STRs data sets. These multivariate techniques were selected as they turned capable of dealing with the nature of the genotypic data, which can be easily binarized. Our goal was to develop multivariate approaches for the interpretation of DNA profiles to better estimate the biogeographical ancestry information of personal genetic profiles, by

Journal Pre-proof

building dynamic and flexible models that could be easily modified according to the number of tested populations and the number of markers in the profile and the reference panel. Our multivariate statistics approach may represent a powerful tool for research purposes and the investigative authorities, too.

2. Materials & Methods

2.1. Datasets

Four different population datasets were selected for this study. All the datasets consisted of individual genotypes rather than allele frequencies. In order of decreasing heterogeneity, the first dataset was extracted from the NIST U.S. population database[26], and consisted of genotypic data for U.S. African-American (N = 342), Asian (N = 97) and Caucasian (N = 361). For this dataset, the following 24 markers were selected: D1S1656, D2S441, D2S1338, D3S1358, D5S818, D6S1043, D7S820, D8S1179, D10S1248, D12S391, D13S317, D16S539, D18S51, D19S433, D21S11, D22S1045, CSF1PO, FGA, Penta D, Penta E, SE33, TH01, TPOX, vWA[26]. Markers F13A01, F13B, FESFPS, LPL and Penta C, which are present in the NIST U.S. population database, were not considered in this study since they are usually not included in commercially available autosomal STR amplification kits commonly used in forensic laboratories.

The second dataset consisted of original unpublished genotypes from Northern and sub-Saharan African populations analyzed for 16 autosomal STRs loci using the AmpFℓSTR® NGM SElect™ PCR Amplification Kit from Thermo Fisher Scientific (D10S1248, vWA, D16S539, D2S1338, D8S1179, D21S11, D18S51, D22S1045, D19S433, TH01, FGA, D2S441, D3S1358, D1S1656, D12S391, SE33), i.e. 231 Northern Africans (67 Moroccan Berbers, 62 Algerian Berbers, 60 Libyan Arabs and 42 Northern Egyptians) and 197 sub-Saharan Africans (95 Cameroonians, 49 Chadians and 53 Senegalese). All the biological samples included in this dataset were randomly collected from informed people, whose genotypes were successfully tested for Hardy-Weinberg and linkage equilibrium. The obtained

Journal Pre-proof

results are currently undergoing publication process. This study ethically complies with the ISFG guidelines for the publication of genetic population data [27–29] and was formally approved by the Carabinieri Scientific Investigations Department of Rome (Italy).The third dataset comprised two central Asian populations genotyped for 15 autosomal STRs loci using the AmpFℓSTR® Identifiler™ PCR Amplification Kit panel from Applied Biosystems (D3S1358, vWA, FGA, D8S1179, D21S11, D18S51, D5S818, D13S317, D7S820, TH01, TPOX, CSF1PO, D19S433, D2S1338, D16S539), i.e. 65 unrelated Afghan [30] and 103 Iraqi[31] (mainly from central and southern Iraq provinces).

The fourth dataset comprised two populations genotyped for 16 autosomal STRs loci using the AmpFℓSTR® NGM SElect™ PCR Amplification Kit from Thermo Fisher Scientific (D10S1248, vWA, D16S539, D2S1338, D8S1179, D2S11, D18S51, D22S1045, D19S433, TH01, FGA, D2S441, D3S1358, D1S1656, D12S391, SE33), i.e. 209 unrelated Italian individuals[32], and 287 Eastern Europeans (223 Romanian and 64 Moldavian subjects)[33].

For each dataset, we evaluated the inter-population genetic differentiation using the F_{ST} statistics (measure of the co-ancestry for randomly chosen alleles within the same subpopulation relative to the entire population) [34,35] in order to have a convenient metrics to objectively measure genetic differentiation among populations when estimating BGA of individuals belonging to such populations. F_{ST} values were obtained using the software STRAF v. 1.0.5[36].

2.2. Multivariate Statistics

Multivariate models were built on the DNA profiles, where each STR profile was converted into a row of zeros and ones by means of an *in-house* code developed in the R software (version 1.1.463)[37] statistical environment. In details, for all the tested individuals, a value equal to 1 was reported for the alleles *x* and *y* (where *x* is equal to *y* in case of homozygosity) recorded for a specific marker *Z*, while a value equal to 0 was reported for the other *n* available alleles of the previously cited marker *Z*. Consequently,

Journal Pre-proof

the STRs DNA profile of each individual was converted into a series of zeros and ones (i.e. binary dataset). Since the matrices obtained by using such computational approach turned to show many zeros as compared to the number of ones, *sparse* algorithms had to be considered when calculating the multivariate models.

SL-PCA, sPLS-DA and SVM multivariate techniques were employed to obtain reliable models for the estimation of the BGA information of unknown genetic profiles. Multivariate modelling and calculations were carried out in R (version 3.6.0)[37]. The following functions and R packages were used in order to build in-house R code for computing the different models: sparse logistic Principal Component Analysis[19], *mixOmics*[38] and *e1071*[39]. In-house developed codes will be available to the readers upon requests to the authors. A workflow of the approach employed in this study is reported in Figure S1 in the Supplementary Material.

Initially, SL-PCA was utilized as an exploratory analysis tool to verify the capabilities of multivariate statistics in recognizing specific pattern regarding the biogeographical origins of the individuals based on their STR profiles, especially when dealing with binary data (as reported above). PCA, here employed in the *sparse* and *logistic* version reported in [19], is one of the most exploited technique in the field of multivariate statistics; it allows to graphically represent the information contained into large data matrices by providing useful visual representations of data distributions, similarity trends, classes and outliers[40]. In practice, PCA evaluates the original data collected for several “objects” (i.e. the encoded individuals), by re-modelling them within new Cartesian diagrams. The new axes of these diagrams represent the Principal Components (PCs), defined as a linear combination of the original variables to make them reciprocally orthogonal.

After the preliminary evaluation of SL-PCA modelling, sPLS-DA and SVM models were applied, to assess their predictive capabilities in blind inference of the ethnic affiliation of DNA profiles. sPLS-DA is the *sparse* version of the combination of Partial Least Squares (PLS) and Discriminant Analysis (DA) techniques[22,41,42]. In practice, sPLS-regression finds the factors that capture the greatest amount of variance in predictor variables by

Journal Pre-proof

simultaneously modelling those X predictors that optimally correlate the responses of the Y matrix. Briefly, the PLS algorithm indicates that the Y responses are proportional to the first principal component – named as Latent Variable (LV) – except for some residuals; then, residuals turn proportional to the second LV, except for new residuals, *etc*. Afterwards, the slopes of the regression line – named as PLS weights – are calculated as residual regression coefficients and indicates the direction of the first LV. The variables/predictors are not usually independent and PLS may provide a bilinear projection model, plus some residuals. Because of that, PLS admits that some X-data are not correlated to Y-responses; these data can represent noise or redundancy, thus indicating that PLS tolerates noisy or redundant data, unlike other regression methodologies. On the other hand, LDA is a supervised classification method whose goal is to discriminate different classes of objects by evaluating the optimal boundaries among them. Originally developed by Fisher[20], LDA allows discriminating objects of different classes by examining the probability distributions of the classes to which the objects may belong. Accordingly, each object is classified in the specific class which shows the highest score in terms of probability. Graphically, the probability distributions are expressed as ellipses at different probability levels for each class under examination. These ellipses are respectively tangential to a point that is located half-way among the class centers and a straight delimiter is adopted as a boundary to separate the ellipses and, consequently, the different classes. LDA provides a linear function of the variables and maximizes the ratio between the variances of each class; weights are adopted to provide the best classification of the objects so that LDA can select the direction achieving the maximum separation among the given classes.

Finally, SVM is a Multivariate Data Analysis (also known as Machine Learning) methodology usually adopted for pattern recognition tasks. Very concisely, this methodology was developed by Vapnik[24] with the aim to provide a decision rule in terms of a special type of hyperplanes, defined as “optimal separating hyperplanes” and also known as “delimiter” or “margin”[23], capable of recognizing and discriminating the objects of different sets or classes. The delimiter is optimized as the distance between the

Journal Pre-proof

separating decision boundaries (hyperplanes) and the closest objects to these hyperplanes, which are defined as support vectors. As reported by Vapnik[24], SVM techniques map the objects matrix X into a high-dimensional space called “feature space”; then linear or nonlinear functions (such as kernels) may be adopted in order to build an optimal separating hyperplane in this space.

All the multivariate models were assembled adopting the 70% of the available data as training set and the remaining 30% of data was employed as evaluation set. Repeated double cross-validation procedures were performed by applying a venetian blind design and a number of data splits equal to 5 (i.e. 80% of the available data of the training set was employed to build the models), in accordance with[43]. Finally, sensitivity and specificity parameters were calculated for all the sPLS-DA and SVM models, as follows: (i) sensitivity is equal to the proportion of individuals belonging to a specific bio-geographical origin that are correctly identified as such, while (ii) specificity is equal to the proportion of individuals belonging to another bio-geographical origin (with reference to the one that is considered by the model) and that are correctly identified as such.

3. Results & Discussion

3.1. Multivariate Statistics

3.1.1. SL-PCA Analysis

SL-PCA was first exploited to rapidly investigate the main features in the datasets. For the NIST dataset, (Figure 1a), three main clusters corresponding to the African-American, Caucasian and Asian individuals were observed in the space of the first two PCs (accounting for 88.03% of total variance). A good separation was also observed for the SL-PCA comparison involving the Northern African and the Sub-Saharan African individuals, where the first two PCs accounted for 65.19% of the total variance (Figure 2a). This result can be due to the fact that the Sahara Desert acted as a strong geographic barrier to gene flow between the cited populations in the last five thousand years[44]. On the contrary, a

Journal Pre-proof

robust separation could not be observed in the case of less genetically differentiated populations such as the Afghan and the Iraqi individuals (Figure 3(a), 77.62% of total variance), as well as the Italian and the Romanian subjects (Figure 4(a), 72.18% of total variance).

In summary, this traditional multivariate procedure allowed us to observe the pertinence of multivariate statistics in assessing and recognizing the biogeographical ancestry information by evaluating the autosomal STRs DNA profiles, only. However, they returned unsatisfactory results whenever the populations to be compared showed quite similar STR allele frequencies (F_{ST} = 0.006 and 0.002, for the Afghans vs. Iraqis and Italians vs. Romanians comparisons, respectively). We then sought to assay more sophisticated and classification-like multivariate models (such as sPLS-DA and SVM techniques) to possibly obtain satisfactory separations between the populations and hence better chances of individual assignment.

3.1.2. sPLS-DA and SVM models

Based on results provided by PCA modelling, sPLS-DA was applied to the same experimental sets to develop useful discrimination models. The predictive models were evaluated in terms of Root Mean Square Error in Cross-Validation (RMSECV)[45], i.e. the lower the RMSECV value, the higher the discrimination power of the model. Moreover, the number of LVs was determined through the evaluation of further quality parameters such as the Predictive Residual Error Sum of Square (PRESS), Q-residuals, Hotelling's T^2, Leverages and Y-Studentized residuals[45]. sPLS-DA results for the different datasets are reported in Figures 1-4(b). Sensitivity and specificity values were calculated for each sPLS-DA model, too (Table 1). Firstly, the African-American, Asian and Caucasian affiliations showed a satisfactory separation (i.e. over 84%) with the SL-PCA (Figure 1(b)). An average error rate equal to 15%, 6% and 16% was observed for the African-American, Asian and Caucasian populations, respectively.

Journal Pre-proof

The Northern African and the Sub-Saharan populations showed a good discrimination (Figure 2(b)), in agreement with the relatively high inter-population genetic diversity observed (F_{ST} = 0.011). An average error rate equal to 6% and 9% was observed for the Northern African and the Sub-Saharan individuals, respectively.

Thirdly, the Afghan and the Iraqi populations showed a better separation than in SL-PCA (Figure 3(b)). An average error rate equal to 7% and 13% was observed for the Afghan and Iraqi subjects, respectively. This result might be ascribed to the fact that such populations are geographically and genetically separated[46,47].

Finally, an unsatisfactory result was obtained for the discrimination of the Italian and the Romanian populations (Figure 4(b)), i.e. the populations showing the lowest F_{ST} value among our datasets. In fact, an average error rate equal to 48% and 37% was observed for the Italian and Romanian individuals, respectively, so that more consistent Machine Learning approaches should be considered.

In conclusion, sPLS-DA might represent a useful tool for improving the routine estimation of the BGA information of autosomal STRs DNA profiles.

SVM was applied to the experimental datasets and the corresponding sensitivity and specificity values are reported in Table 1, too. A 100% accuracy as observed for the following tested populations: African-American vs. Asian vs. Caucasian individuals (NIST U.S. data), Northern vs. Sub-Saharan individuals and Afghan vs. Iraqi individuals. No misclassifications were observed both with the cross-validated training set and the extracted test set. On the other hand, an accuracy equal to 89.1% was calculated for the SVM model as applied to the Italian vs. Romanian population. In the present case optimal sensitivity and specificity values were obtained equal to 87.1% and 90.6%, respectively), corresponding to an overall number of 27 misclassifications out of 209 Italian individuals and 27 misclassifications out of 287 Romanian subjects. Consequently, SVM turned out to be a very powerful model, with high specificity and sensitivity values, for all the ethnic groups, thus proving once again the reliability of multivariate statistics to extract BGA information from autosomal STRs DNA genetic profiles. Finally, the traditional Bayesian approach involving the calculation of likelihood values was used to calculate the

Journal Pre-proof

sensitivity and specificity values on the same data on which we tested the different sPLS-DA and SVM models. Similar results were obtained for the prediction of NIST populations and Northern vs Sub-Saharan African individuals. However, this approach provided slightly worse results than sPLS-DA for the comparison of Italian and Romanian individuals, while the specificity and the sensitivity values turned to be significantly lower when evaluating the Afghan and the Iraqi subjects. All the results are reported in Table 1.

Journal Pre-proof

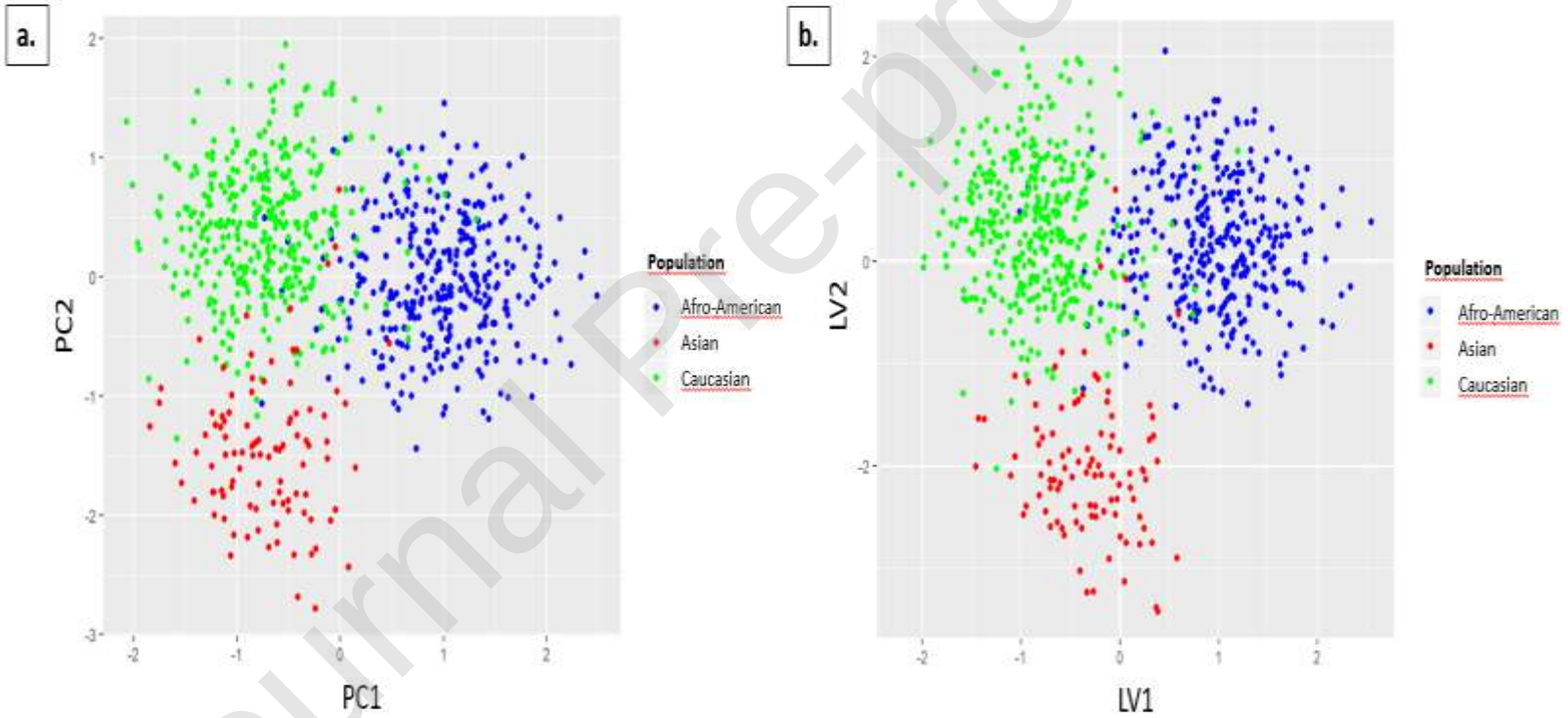


Figure 1. (a.) SL-PCA PC1 vs. PC2 PCA Scores Plot. (b.) sPLS-DA LV1 vs LV2 Scores Plot. Afro-American, Asian and Caucasian individuals are in blue, red and green points, respectively. Three main clusters can be observed in the figures.

Journal Pre-proof

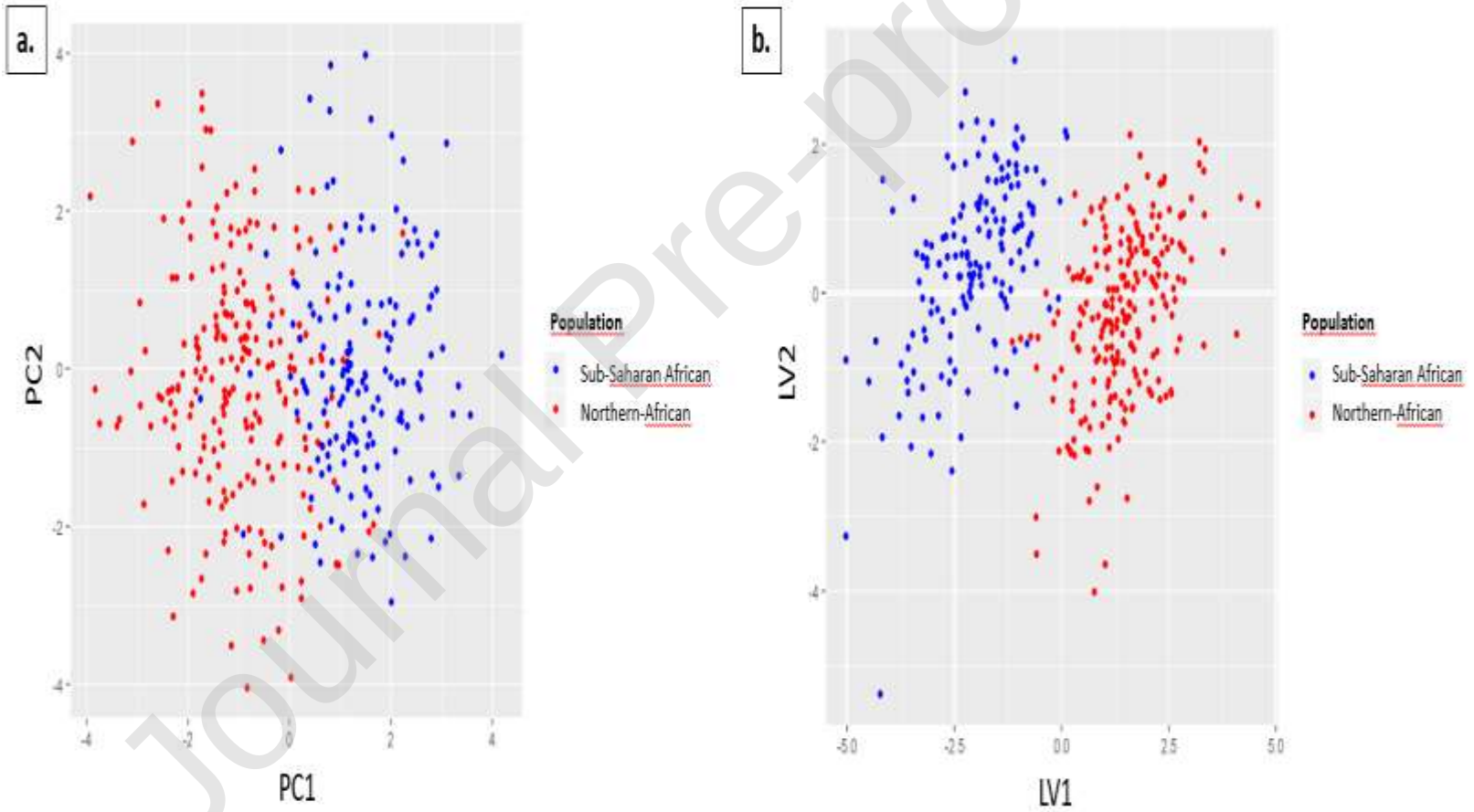


Figure 2 (a.) SL-PCA PC1 vs. PC2 PCA Scores Plot. and (b.) sPLS-DA LV1 vs LV2 Scores Plot for sub-Saharan (blue points) vs. Northern-African (red points) subjects.

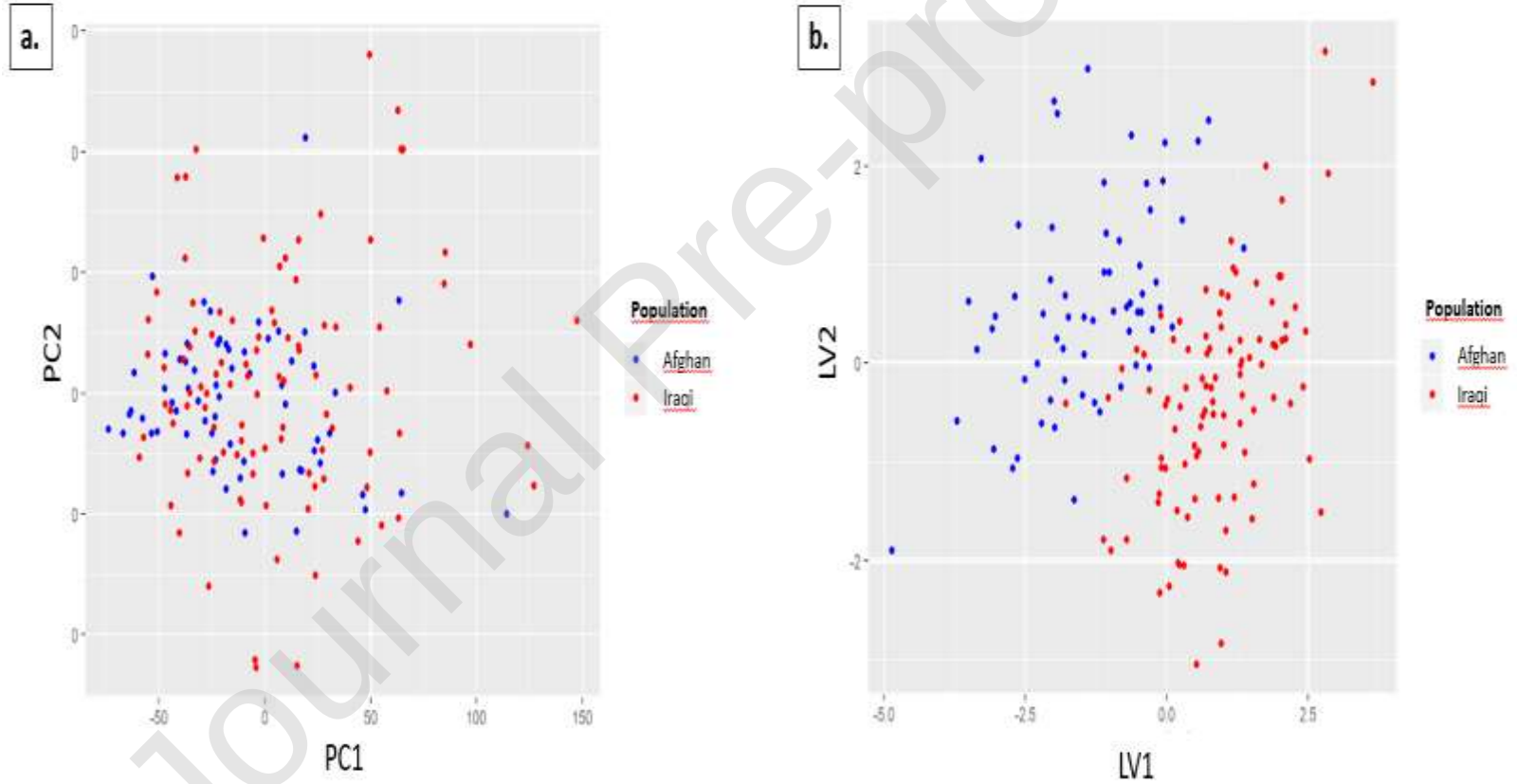


Figure 3. (a.) SL-PCA PC1 vs. PC2 PCA Scores Plot. and (b.) sPLS-DA LV1 vs. LV2 Scores Plot for Afghans (blue points) vs. Iraqis (red points).

Journal Pre-proof

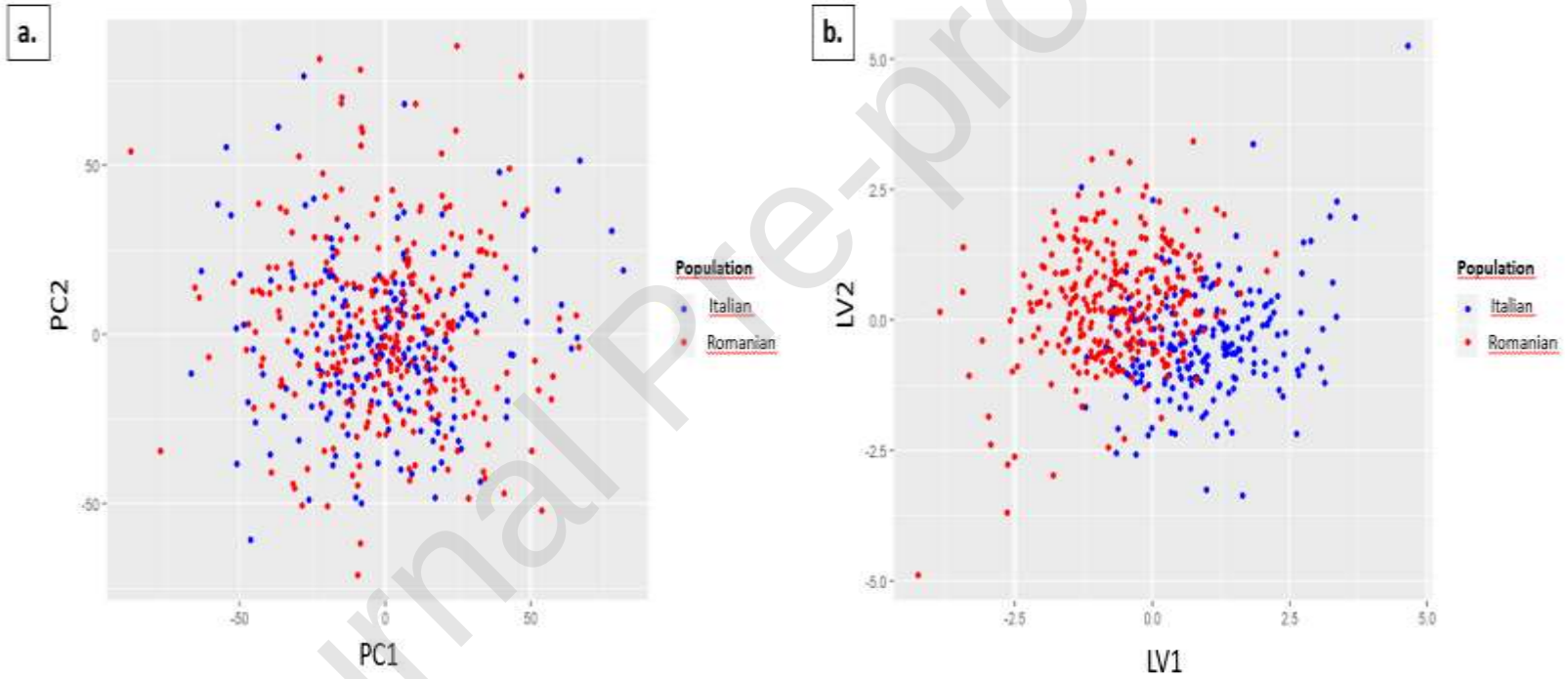


Figure 4. (a.) SL-PCA PC1 vs. PC2 PCA Scores Plot and (b.) sPLS-DA LV1 vs. LV2 Scores Plot for the Italian (blue points) vs. Romanian subjects (red points).

Journal Pre-proof

Journal Pre-proof

Table 1. Sensitivity and specificity values for sPLS-DA, SVM and Bayesian* models. The results are relative to the models validated by using a cross-validation strategy involving 5 cancellation groups and a venetian blind design. Similar results were obtained when testing the models on external validation datasets.

	Afro-American (NIST U.S.)			**Caucasian (NIST U.S.)**			**Asian (NIST U.S.)**			**Northern vs Sub-Saharan African**			**Afghan vs. Iraqi**			**Italian vs. Romanian**		
Model	*sPLS-DA*	*SVM*	*Bayesian*	*sPLS-DA*	*SVM*	*Bayesian*	*sPLS-DA*	*SVM*	*Bayesian*	*sPLS-DA*	*SVM*	*Bayesian*	*sPLS-DA*	*SVM*	*Bayesian*	*sPLS-DA*	*SVM*	*Bayesian*
Sensitivity	85.1%	100.0%	87.3%	94.2%	100.0%	95.8%	84.5%	100%	83.2%	94.4%	100.0%	91.9%	93.8%	100.0%	46.9%	52.2%	87.1%	46.4%
Specificity	92.1%	100.0%	93.6%	85.0%	100.0%	84.9%	89.8%	100%	88.1%	93.3%	100.0%	94.2%	95.7%	100.0%	50.5%	64.4%	90.6%	48.4%

* The calculation for the "Bayesian" model were performed using STRUCTURE software (version 2.3.4), whose parameters were, as follows: Length of Burnin Period: 100000; Number of MCMC reps after Burnin: 100000; No Admixture Ancestry model; Independent Allele Frequencies.

Journal Pre-proof

4. Conclusions

The present proof-of-concept study demonstrates the capability of multivariate statistics approaches to predict the population affiliation of autosomal genetic profiles that can be commonly recovered from any source, including crime scenes, mass-disaster and missing person investigations. sPLS-DA and SVM techniques drastically improved PCA, by providing optimal discrimination results (i.e. showing the lowest sensitivity value equal to 84%) and being capable of assessing the group affiliation of the examined DNA profiles according to their autosomal STRs loci. The predictive power of such multivariate techniques turned extremely high, indicating that the adoption of multivariate models may represent a powerful and useful tool for the investigative authorities to ease their decision processes when estimating the BGA of individuals. Future perspectives include the application of these multivariate strategies in discriminating even more locally-restricted populations. Further research studies with sPLS-DA and SVM techniques are already planned and will be performed in our laboratories using Next Generation Sequencing (NGS)/Massive Parallel Sequencing (MPS), by combining their data with the autosomal STRs results or developing the cited multivariate approaches on other forensic genetic markers such as Y-STR and SNPs. Moreover, an open-source and free-of-charge app is currently under development aiming to allow analysts to perform the described approaches for their routine BGA investigations.

Journal Pre-proof

Journal Pre-proof

Acknowledgements

This work was supported by: Sapienza University of Rome (grant n. RM11715C77B03CDC to FC); University of Pavia strategic theme “Towards a governance model for international migration: an interdisciplinary and diachronic perspective” (MIGRAT-IN-G) (OS); the Italian Ministry of Education, University and Research (MIUR): Dipartimenti di Eccellenza Program (2018–2022), Dept. of Biology and Biotechnology "L. Spallanzani", University of Pavia (OS).

Journal Pre-proof

Journal Pre-proof

References

[1] P. Gill, H. Haned, O. Bleka, O. Hansson, G. Dørum, T. Egeland, Genotyping and interpretation of STR-DNA: Low-template, mixtures and database matches-Twenty years of research and development., Forensic Sci. Int. Genet. 18 (2015) 100–17. doi:10.1016/j.fsigen.2015.03.014.

[2] A. Amorim, B. Budowle, Handbook of Forensic Genetics, WORLD SCIENTIFIC (EUROPE), 2016. doi:10.1142/q0023.

[3] M. Kayser, P. De Knijff, Improving human forensics through advances in genetics, genomics and molecular biology, Nat. Rev. Genet. 12 (2011) 179–192. doi:10.1038/nrg2952.

[4] V. Pereira, A. Freire-Aradas, D. Ballard, C. Børsting, V. Diez, P. Pruszkowska-Przybylska, J. Ribeiro, N.M. Achakzai, A. Aliferi, O. Bulbul, M.D.P. Carceles, S. Triki-Fendri, A. Rebai, D.S. Court, N. Morling, M.V. Lareu, Á. Carracedo, C. Phillips, Development and validation of the EUROFORGEN NAME (North African and Middle Eastern) ancestry panel, Forensic Sci. Int. Genet. (2019). doi:10.1016/j.fsigen.2019.06.010.

[5] P. Gill, C.H. Brenner, J.S. Buckleton, A. Carracedo, M. Krawczak, W.R. Mayr, N. Morling, M. Prinz, P.M. Schneider, B.S. Weir, DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures, Forensic Sci. Int. 160 (2006) 90–101. doi:10.1016/j.forsciint.2006.04.009.

[6] B. Budowle, A.J. Onorato, T.F. Callaghan, A. Della Manna, A.M. Gross, R.A. Guerrieri, J.C. Luttman, D.L. McClure, Mixture Interpretation: Defining the Relevant Features for Guidelines for the Assessment of Mixed DNA Profiles in Forensic Casework, J. Forensic Sci. 54 (2009) 810–821. doi:10.1111/j.1556-4029.2009.01046.x.

[7] D. Taylor, J.-A. Bright, J. Buckleton, J. Curran, An illustration of the effect of various sources of uncertainty on DNA likelihood ratio calculations, Forensic Sci. Int. Genet. 11 (2014) 56–63. doi:10.1016/j.fsigen.2014.02.003.

[8] P.M. Vallone, A.E. Decker, J.M. Butler, Allele frequencies for 70 autosomal SNP loci with U.S. Caucasian, African-American, and Hispanic samples, Forensic Sci. Int. 149 (2005) 279–286. doi:10.1016/j.forsciint.2004.07.014.

[9] H. Boonyarit, S. Mahasirimongkol, N. Chavalvechakul, M. Aoki, H. Amitani, N. Hosono, N. Kamatani, M. Kubo, P. Lertrit, Development of a SNP set for human identification: A set with high powers of discrimination which yields high genetic information from naturally degraded DNA samples in the Thai population, Forensic Sci. Int. Genet. 11 (2014) 166–173. doi:10.1016/j.fsigen.2014.03.010.

[10] J.M. Butler, Advanced Topics in Forensic DNA Typing: Methodology, Elsevier, 2012. doi:10.1016/C2011-0-04189-3.

[11] C. Phillips, C. Santos, M. Fondevila, Á. Carracedo, M.V. Lareu, Inference of Ancestry in Forensic Analysis I: Autosomal Ancestry-Informative Marker Sets, in: 2016: pp. 233–253. doi:10.1007/978-1-4939-3597-0_18.

[12] C.H. Brenner, Some mathematical problems in the DNA identification of victims in the 2004 tsunami and similar mass fatalities, Forensic Sci. Int. 157 (2006) 172–180. doi:10.1016/j.forsciint.2005.11.003.

[13] C.H. Brenner, B.S. Weir, Issues and strategies in the DNA identification of World Trade Center victims, Theor. Popul. Biol. 63 (2003) 173–178. doi:10.1016/S0040-5809(03)00008-

Journal Pre-proof

X.

[14] L. Porras-Hurtado, Y. Ruiz, C. Santos, C. Phillips, Á. Carracedo, M. V. Lareu, An overview of STRUCTURE: applications, parameter settings, and supporting software, Front. Genet. 4 (2013). doi:10.3389/fgene.2013.00098.

[15] C. Santos, C. Phillips, A. Gomez-Tato, J. Alvarez-Dios, Á. Carracedo, M.V. Lareu, Inference of Ancestry in Forensic Analysis II: Analysis of Genetic Data, in: 2016: pp. 255–285. doi:10.1007/978-1-4939-3597-0_19.

[16] L. Pereira, F. Alshamali, R. Andreassen, R. Ballard, W. Chantratita, N.S. Cho, C. Coudray, J.-M. Dugoujon, M. Espinoza, F. González-Andrade, S. Hadi, U.-D. Immel, C. Marian, A. Gonzalez-Martin, G. Mertens, W. Parson, C. Perone, L. Prieto, H. Takeshita, H. Rangel Villalobos, Z. Zeng, L. Zhivotovsky, R. Camacho, N.A. Fonseca, PopAffiliator: online calculator for individual affiliation to a major population group based on 17 autosomal short tandem repeat genotype profile, Int. J. Legal Med. 125 (2011) 629–636. doi:10.1007/s00414-010-0472-2.

[17] C. Santos, C. Phillips, A. Gomez-Tato, J. Alvarez-Dios, Á. Carracedo, M.V. Lareu, Inference of Ancestry in Forensic Analysis II: Analysis of Genetic Data., Methods Mol. Biol. 1420 (2016) 255–85. doi:10.1007/978-1-4939-3597-0_19.

[18] L. Porras-Hurtado, Y. Ruiz, C. Santos, C. Phillips, Á. Carracedo, M. V. Lareu, An overview of STRUCTURE: Applications, parameter settings, and supporting software, Front. Genet. 4 (2013). doi:10.3389/fgene.2013.00098.

[19] S. Lee, J.Z. Huang, J. Hu, Sparse Logistic Principal Components Analysis for binary data, Ann. Appl. Stat. 4 (2010) 1579–1601. doi:10.1214/10-AOAS327SUPP.

[20] M. Barker, W. Rayens, Partial least squares for discrimination, J. Chemom. 17 (2003) 166–173. doi:10.1002/cem.785.

[21] D. Ballabio, V. Consonni, Classification tools in chemistry. Part 1: linear models. PLS-DA, Anal. Methods. 5 (2013) 3790. doi:10.1039/c3ay40582f.

[22] K.-A. Lê Cao, D. Rossouw, C. Robert-Granié, P. Besse, A Sparse PLS for Variable Selection when Integrating Omics Data, Stat. Appl. Genet. Mol. Biol. 7 (2008). doi:10.2202/1544-6115.1390.

[23] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, B. Scholkopf, Support vector machines, IEEE Intell. Syst. Their Appl. 13 (1998) 18–28. doi:10.1109/5254.708428.

[24] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer New York, New York, NY, 2000. doi:10.1007/978-1-4757-3264-1.

[25] M. Forina, Fondamenta per la Chimica Analitica, 2012. http://www.sisnir.org/sisnir/download/fondamenta-per-la-chimica-analitica.

[26] C.R. Hill, D.L. Duewer, M.C. Kline, M.D. Coble, J.M. Butler, U.S. population data for 29 autosomal STR loci, Forensic Sci. Int. Genet. 7 (2013) e82–e83. doi:10.1016/j.fsigen.2012.12.004.

[27] Á. Carracedo, J.M. Butler, L. Gusmão, A. Linacre, W. Parson, L. Roewer, P.M. Schneider, New guidelines for the publication of genetic population data, Forensic Sci. Int. Genet. 7 (2013) 217–220. doi:10.1016/j.fsigen.2013.01.001.

[28] A. Carracedo, J.M. Butler, L. Gusmão, A. Linacre, W. Parson, L. Roewer, P.M. Schneider, Update of the guidelines for the publication of genetic population data, Forensic Sci. Int.

Journal Pre-proof

Genet. 10 (2014) A1–A2. doi:10.1016/j.fsigen.2014.01.004.

[29] L. Gusmão, J.M. Butler, A. Linacre, W. Parson, W. Parson, P.M. Schneider, A. Carracedo, Revised guidelines for the publication of genetic population data, Forensic Sci. Int. Genet. 30 (2017) 160–163. doi:10.1016/j.fsigen.2017.06.007.

[30] A. Berti, F. Barni, A. Virgili, G. Iacovacci, C. Franchi, C. Rapone, A. Di Carlo, C.M. Oddo, G. Lago, Autosomal STR Frequencies in Afghanistan Population, J. Forensic Sci. 50 (2005) 1–3. doi:10.1520/jfs2005237.

[31] F. Barni, A. Berti, A. Pianese, A. Boccellino, M.P. Miller, A. Caperna, G. Lago, Allele frequencies of 15 autosomal STR loci in the Iraq population with comparisons to other populations from the middle-eastern region, Forensic Sci. Int. 167 (2007) 87–92. doi:10.1016/j.forsciint.2006.03.005.

[32] A. Berti, F. Brisighelli, A. Bosetti, E. Pilli, Allele frequencies of the new European Standard Set (ESS) loci in the Italian population, Forensic Sci. Int. Genet. 5 (2011) 548–549. doi:10.1016/j.fsigen.2010.01.006.

[33] A. Benvisto, F. Messina, A. Finocchio, L. Popa, M. Stefan, G. Stefanescu, C. Mironeanu, A. Novelletto, C. Rapone, A. Berti, A genetic portrait of the South-Eastern Carpathians based on autosomal short tandem repeats loci used in forensics., Am. J. Hum. Biol. 30 (2018) e23139. doi:10.1002/ajhb.23139.

[34] B.S. Weir, C.C. Cockerham, Estimating F-Statistics for the Analysis of Population Structure, Evolution (N. Y). (2006). doi:10.2307/2408641.

[35] K.E. Holsinger, B.S. Weir, Genetics in geographically structured populations: Defining, estimating and interpreting FST, Nat. Rev. Genet. 10 (2009) 639–650. doi:10.1038/nrg2611.

[36] A. Gouy, M. Zieger, STRAF—A convenient online tool for STR data evaluation in forensic genetics, Forensic Sci. Int. Genet. 30 (2017) 148–151. doi:10.1016/j.fsigen.2017.07.007.

[37] R Core Team, R: A language and environment for statistical computing, (2015). https://www.r-project.org/.

[38] F. Rohart, B. Gautier, A. Singh, K.A. Lê Cao, mixOmics: An R package for 'omics feature selection and multiple data integration, PLoS Comput. Biol. 13 (2017). doi:10.1371/journal.pcbi.1005752.

[39] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, C. Chang, C. Lin, Package 'e1071' - Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien, (2019). https://cran.r-project.org/web/packages/e1071/e1071.pdf (accessed October 7, 2019).

[40] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods. 6 (2014) 2812–2831. doi:10.1039/C3AY41907J.

[41] K.-A. Lê Cao, S. Boitard, P. Besse, Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems., BMC Bioinformatics. 12 (2011) 253. doi:10.1186/1471-2105-12-253.

[42] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. Syst. 58 (2001) 109–130. doi:10.1016/S0169-7439(01)00155-1.

[43] P. Filzmoser, B. Liebmann, K. Varmuza, Repeated double cross validation, J. Chemom. 23 (2009) 160–171. doi:10.1002/cem.1225.

Journal Pre-proof

[44] E. D'Atanasio, B. Trombetta, M. Bonito, A. Finocchio, G. Di Vito, M. Seghizzi, R. Romano, G. Russo, G.M. Paganotti, E. Watson, A. Coppa, P. Anagnostou, J.-M. Dugoujon, P. Moral, D. Sellitto, A. Novelletto, F. Cruciani, The peopling of the last Green Sahara revealed by high-coverage resequencing of trans-Saharan patrilineages., Genome Biol. 19 (2018) 20. doi:10.1186/s13059-018-1393-5.

[45] M. Forina, S. Lanteri, M.C.C. Oliveros, C.P. Millan, Selection of useful predictors in multivariate calibration, Anal. Bioanal. Chem. 380 (2004) 397–418. doi:10.1007/s00216-004-2768-x.

[46] S. Dogan, C. Gurkan, M. Dogan, H.E. Balkaya, R. Tunc, D.K. Demirdov, N.A. Ameen, D. Marjanovic, A glimpse at the intricate mosaic of ethnicities from Mesopotamia: Paternal lineages of the Northern Iraqi Arabs, Kurds, Syriacs, Turkmens and Yazidis, PLoS One. (2017). doi:10.1371/journal.pone.0187408.

[47] J. Di Cristofaro, S. Buhler, S.A. Temori, J. Chiaroni, Genetic data of 15 STR loci in five populations from Afghanistan., Forensic Sci. Int. Genet. 6 (2012) e44-5. doi:10.1016/j.fsigen.2011.03.004.

Journal Pre-proof

Journal Pre-proof

Journal Pre-proof