Science Bulletin 64 (2019) 1381–1384

Contents lists available at ScienceDirect

Science Bulletin

journal homepage: www.elsevier.com/locate/scib

Science Bulletin
www.scibull.com

News & Views

Biogeographical informativeness of Y-STR haplotypes

Jacobo Pardo-Seco [a,b], Alberto Gómez-Carballa [a,b], Xabier Bello [a,b], Federico Martinón-Torres [b], Antonio Salas [a,b,*]

[a] Unidade de Xenética, Instituto de Ciencias Forenses (INCIFOR), Facultade de Medicina, Universidade de Santiago de Compostela, and GenPoB Research Group, Instituto de Investigaciones Sanitarias (IDIS), Hospital Clínico Universitario de Santiago (SERGAS), Galicia, Spain
[b] Grupo de Investigación en Genética, Vacunas, Infecciones y Pediatría (GENVIP), Universidade de Santiago de Compostela, and Hospital Clínico Universitario de Santiago (SERGAS), Galicia, Spain

Research on biogeographical ancestry (BGA) is becoming of growing interest in forensic genetics and in the biomedical literature [1]. Thus, for instance, the need to predict ethnicity of an unknown suspect based on DNA profiles found at the crime scene is of maximum interest in criminalistics [2], and several autosomal SNP panels have been designed and tested for BGA investigations [3,4]. Most of these panels aim at discriminating three main continental groups (sub-Saharan Africans, Europeans, and Asians) by way of testing a number of ancestry informative markers (AIMs) that run from a few dozens to a few hundred [5] (see more background in Supplementary data online).

BGA can be also explored from uniparental markers using e.g., phylogeographic [6,7] or statistical approaches [8,9]. The main aim of the present study is to explore the BGA content of Y-STR haplotypes that represent worldwide populations. The results are relevant to different fields of research, including molecular anthropology and forensic genetics. In addition, all the classification procedures developed here are implemented in the frontend web development Y-Biogeographical Ancestry Tool (or Y-BAT; www.y-bat.eu).

We investigate BGA of Y-chromosome STRs (Y-STRs) in populations worldwide (n = 19630 Y-STR haplotypes) and grouping the samples in different continental schemes (Table S1 online): 3-way continental ancestry (sub-Saharan Africa, Eastern Asia, and Europe), 7-way continental ancestry (sub-Saharan Africa, Eastern Asia, Europe, Southeast Asia, South Asia, and Native America), and within European ancestry (Eastern Europe, Southeastern Europe, Western Europe, and Uralic Yukaghir). Most of the analyses were carried out using a probabilistic-based procedure based on quadratic discrimination analysis (QDA) [9] (see Methods in Supplementary Information for all technical details on databases and statistical procedures).

Almost all the haplotypes can be correctly classified when considering the 3-way continental classification context (Table S2; Fig. S1 online). The classification success gets worse when considering the 7-way continental ancestry and the within European ancestry context, but there is still a strong and informative geographical stratification of Y-haplotypes (Table S2; Figs. S2, S3 online).

Simulation experiments indicate that heterogeneous sample sizes of the training classification sets have little impact when classifying haplotypes into the three main ancestral groups (Africa, Eastern Asia, and Europe), but can have a deep impact when classifying haplotypes into a larger set of continental metapopulations or within continents (Figs. S4–S6 online). The apparent lower performance of some population set in the classification do not necessarily relates to the mathematical model employed or the intrinsic BGA content of the Y-STR haplotypes, but also to the ancestry/demographic characteristics of the testing/training population set employed. For instance, the American metapopulation (that considers only Native American groups) receives the main inference of ancestry in America (Fig. 1b), but also an important proportion in Europe and Middle East. This can be due to the fact that all Native American populations have variable European components, especially on the Y-chromosome (due to the gender-bias existing in America on uniparental inherited markers [10]). On the other hand, the BGA inference within Europe performs worst for Southeastern and Eastern, but it performs reasonably well for Western and Uralic-Yukaguir.

While waiting for further analyses aimed at measuring other potential parameters affecting classification (e.g., population genetic diversity), it seems most recommendable to use almost equal classification population sample sizes. Population training sample sizes of ~250 haplotypes seem to perform well when comparing the results with very large sample sizes in the classification algorithm (Supplementary data online).

Haplotype ancestries were interpolated into geographic maps to allow a better visualization of classification probabilities. The ancestries represented in the maps are those derived from the previous simulation experiments (briefly summarized in Tables S3 and S5 online) that uses homogeneous training sets instead the original unbalanced sample sizes (because the performance of the classification is better for the latter). Fig. 1a displays the maps for the 3-way continental classification probabilities. In good agreement with the probabilities shown in Fig. S1 and Table S2 (online), the maps indicate a good concordance of the Y-STR variation with geography. This stratification is also clear when

* Corresponding author.
E-mail address: antonio.salas@usc.es (A. Salas).

https://doi.org/10.1016/j.scib.2019.07.025

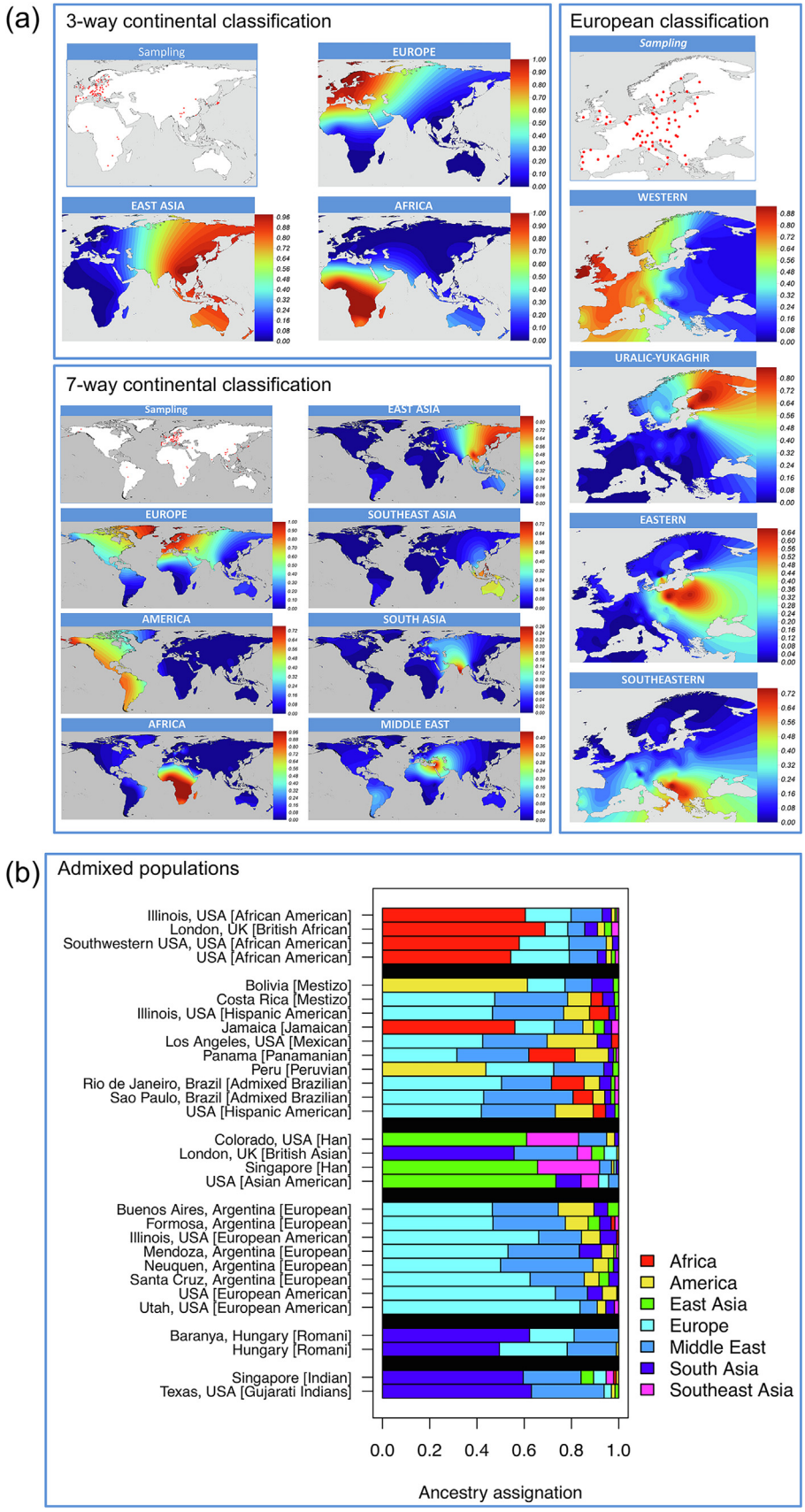

Fig. 1. (a) Interpolated geographic ancestry ascriptions under a 3-way continental model, a 7-way continental model, and in a within European population context. (b) Ancestral classification of admixed populations. The histograms were grouped by main expected ancestry in order to better visualize the different patterns of admixture.

examining the 7-way continental classification (Fig. 1a). It is important to interpret the interpolated maps to the light of the sampling points. For instance, in the Native American ancestry map (Fig. 1a), the Native American component is concentrated in the South America and most northern latitude of the continent; coinciding with the main sampling points for Native American variation; there are not sampling point in most of the North American continent. Even thought, the interpolated maps are very clear at signaling a strong geographic stratification of Y-STR variation at a continental level. We also examined interpolated classification probabilities within European regions (Fig. 1a). Again, the overall picture is that Y-STR variation is strongly stratified within Europe. The maps indirectly reflect the fact that incorrectly ascribed haplotypes are generally classified into neighboring regions giving rise to well stratified geographic maps of Y-STR haplotype variation.

Next, we examined BGA in admixed population groups from Purps et al. [11], and we followed the same "ethnic/population" labels used in the original source. Although, this exercise was done recently in Ecuadorian samples [8], we have now the opportunity to evaluate a wider population context that includes e.g., "African-American" populations, Asian admixed groups, Romani, etc. As shown in Table S6 (online) and Fig. 1b, the ancestries inferred by the algorithm of classification agree well with expectations [12,13], and validates this procedure for future anthropological or forensic genetic studies of admixed populations. For instance, the "African-American" groups have a predominant African ancestry (~58% on average) with a lower European/Middle East (~32% on average) component. The European admixed groups were sampled in America; and although these samples have a main European ancestry, all have a significant Native American component.

In addition, the analysis reveals that pseudo-ethnic labelling of populations can be misled, because the same pseudo-ethnic category might be applied to different admixed patterns. Thus, for instance, while the Asians from USA are most likely originally from Eastern Asia, the British Asians have a main ancestry in South Asia. The "mestizo" from Bolivia and Peru have a predominant Native American ancestry, those from Jamaica have a main African ancestry, whereas other "Mestizo" groups have a main European/Middle East ancestry. The African component is also present at low levels in most "Mestizo" groups with the exception of Bolivians (fitting with previous results indicating that the African component is found almost exclusively in the Yungas Bolivian region [14]) and Peruvians. Overall, these results suggest that pseudo-ethnic labelling might have consequences in different fields of biomedical research (e.g., substructure in case-control association studies, forensic genetics, etc.) [15].

Another interesting finding is that the Romani have a main South Asian male ancestry (~53%) [13], admixed to a lower level with Europe (~25%) and Middle East (~20%); their pattern of admixture is very similar to the one found in Indians from Singapore and Gujarati Indians from USA (Fig. 1b).

Different mathematical procedures for haplotype classification were tested; the results indicate that, overall, QDA performs better than the K-nearest neighbors (K-NN) and the Support Vector Machine (SVM) algorithm and has the advantage of providing a probability of classification instead the categoric classification offered by the other two procedures (Fig. S7 online).

The present study paves the ground for new investigations in the area of BGA inference from Y-STR haplotypes. The methods developed here were implemented in Y-BAT, a web-tool that allows exploring the most likely geographical origin of a single Y-STR haplotypes, and more generally, classifying batches of haplotypes from forensic and/or molecular anthropological studies.

In agreement with previous studies, the results indicate that Y-STRs haplotypes contain an important amount of BGA information that allows classifying haplotypes into main continental or even sub-continental regions. Further efforts could be focused on improving classification algorithms and validation exercises. Although the origin of a Y-STR haplotype in non-admixed populations does not necessarily correlates with the main genome ancestry of carriers, there is an expected correlation (more diffused in admixed populations) that makes inference of BGA an useful tool to assist police investigation [9], especially when investigating Y-haplotypes from DNA databases or when there is not leftover DNA for further analysis from e.g., the crime scene, but the Y-STR haplotype is already available. Therefore, Y-STR contain important BGA information that is inexpensive and can be easily treated in routine forensic casework in order to illuminate police investigations.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgments

A.S. received support from the project GePEM ISCIII/PI16/01478/Cofinanciado FEDER of the Instituto de Salud Carlos III. F.M.T. received support from project ReSVinext ISCIII/PI16/01569/Cofinanciado FEDER.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scib.2019.07.025.

References

[1] Tian C, Gregersen PK, Seldin MF. Accounting for ancestry: population substructure and genome-wide association studies. Hum Mol Genet 2008;17:R143–50.
[2] Gannett Lisa. Biogeographical ancestry and race. Stud Hist Philos Sci Part C: Stud History Philos Biol Biomed Sci 2014;47:173–84.
[3] Salas A, Amigo J. A reduced number of mtsnps saturates mitochondrial DNA haplotype diversity of worldwide population groups. PLoS One 2010;5:e10218.
[4] Phillips C, Prieto L, Fondevila M, et al. Ancestry analysis in the 11-m madrid bomb attack investigation. PLoS One 2009;4:e6583.
[5] Pardo-Seco J, Martinón-Torres F, Salas A. Evaluating the accuracy of aim panels at quantifying genome ancestry. BMC Genom 2014;30:543.
[6] Salas A, Catelli L, Pardo-Seco J, et al. Y-chromosome peruvian origin of the 500-year-old inca child mummy sacrificed in cerro aconcagua (argentina). Sci Bull 2018;63:1457–9.
[7] Pereira L, Černý V, Cerezo M, et al. Linking the sub-saharan and west eurasian gene pools: maternal and paternal heritage of the tuareg nomads from the african sahel. Eur J Hum Genet 2010;18:915–23.
[8] Toscanini U, Gaviria A, Pardo-Seco J, et al. The geographic mosaic of ecuadorian Y-chromosome ancestry. Forensic Sci Int Genet 2018;33:59–65.
[9] Egeland T, Bøvelstad HM, Storvik GO, et al. Inferring the most likely geographical origin of mtDNA sequence profiles. Ann Hum Genet 2004;68:461–71.
[10] Toscanini U, Gusmão L, Berardi G, et al. Y chromosome microsatellite genetic variation in two native american populations from argentina: population stratification and mutation data. Forensic Sci Int Genet 2008;2:274–80.
[11] Purps J, Siegert S, Willuweit S, et al. A global analysis of y-chromosomal haplotype diversity for 23 str loci. Forensic Sci Int Genet 2014;12C:12–23.
[12] Salas A, Jaime JC, Álvarez-Iglesias V, et al. Gender bias in the multiethnic genetic composition of central argentina. J Hum Genet 2008;53:662–74.
[13] Gómez-Carballa A, Pardo-Seco J, Fachal L, et al. Indian signatures in the westernmost edge of the european romani diaspora: new insight from mitogenomes. PLoS One 2013;8:e75397.
[14] Taboada-Echalar P, Álvarez-Iglesias V, Heinz T, et al. The genetic legacy of the pre-colonial period in contemporary bolivians. PLoS One 2013;8:e58980.
[15] Salas A, Elson JL. Mitochondrial DNA as a risk factor for false positives in case-control association studies. J Genet Genom 2015;42:169–72.

Antonio Salas is a full professor at the University of Santiago de Compostela (USC; Galicia, Spain). Prof. Salas' research focuses on genomics in the field of molecular anthropology, medical and forensic genetics/genomics. In the last few years, he is showing special interest to the field of host infectomics with special focus on genomics and transcriptomics. He is the head of GenPob (Population Genetics in Biomedicine) research group at the Instituto de Investigaciones Sanitarias - IDIS) and member of the Instituto de Ciencias Forenses (INCIFOR) of the USC; both located in the city of Santiago de Compostela, Spain.