West Asian sources of the Eurasian component in Ethiopians: a reassessment

Ludovica Molinaro[1*], Francesco Montinaro[1], Toomas Kivisild[1,2], Luca Pagani[1,3*]

[1]Estonian Biocentre, Institute of Genomics, University of Tartu, Estonia
[2]Department of Human Genetics, KU Leuven, Leuven, Belgium
[3]Department of Biology, University of Padova, Italy

*To whom correspondence may be addressed: lu.molinaro8@gmail.com; lp.lucapagani@gmail.com

Summary

Previous genome-scale studies of populations living today in Ethiopia have found evidence of recent gene flow from an Eurasian source, dating to the last 3,000 years[1,2,3,4]. Haplotype[1] and genotype data based analyses of modern[2,4] and ancient data (aDNA)[3,5] have considered Sardinia-like proxy[2], broadly Levantine[1,4] or Neolithic Levantine[3] populations as a range of possible sources for this gene flow. Given the ancient nature of this gene flow and the extent of population movements and replacements that affected West Asia in the last 3000 years, aDNA evidence would seem as the best proxy for determining the putative population source. We demonstrate, however, that the deeply divergent, autochthonous African component which accounts for ~50% of most contemporary Ethiopian genomes, affects the overall allele frequency spectrum to an extent that makes it hard to control for it and, at once, to discern between subtly different, yet important, Eurasian sources (such as Anatolian or Levant Neolithic ones). Here we re-assess pattern of allele sharing between the Eurasian component of Ethiopians (here called "NAF" for Non African) and ancient and modern proxies area after having extracted NAF from Ethiopians through ancestry deconvolution, and unveil a genomic signature compatible with population movements that affected the Mediterranean area and the Levant after the fall of the Minoan civilization.

Results and Discussion

To determine the most likely source of the Eurasian gene flow into the ancestral gene pool of present-day Ethiopians we have used a combination of ancestry deconvolution (AD) and allele sharing methods[6]. AD refers to analyses that determine the likeliest ancestry composition of genomes of individuals with mixed ancestry at fine haplotype resolution. These methods have

allowed us to i) exploit high quality modern data and ii) harness the power of allele sharing tools on genetic fractions with no or reduced African contributions. Such a strategy, while potentially beneficial, introduce a novel source of bias which we aimed to explore here. Particularly, after AD of 120 Ethiopian genomes[7], we assigned each genomic SNP into one of the following four categories based on the method likelihoods (see Methods for further details): 1) confidently non African (NAF); 2) low confidence non African (X); 3) low confidence African (Y) and 4) confidently African (AF, consistently filtered out from our analyses). While basing our inference on the NAF component alone, we here demonstrate that the component X does account for a minority of the genome and, when analysed together with NAF does not qualitatively change the results. Furthermore, when joining together the NAF and AF confidently assigned components (to create "Joint" components) we recapitulate the signals of the global population (prior to ancestry deconvolution), showing that the X and Y components are not holding a considerable or peculiar genetic signature and hence ruling out, in this study, the role of ancestry deconvolution as a potential source of artifacts. For the sake of clarity, out of the four admixed Ethiopian populations available from Pagani et al. 2015 (Amhara, Oromo, Somali, Wolayta), we report results only on the NAF component of Amhara. Comparable results for the other three populations, which we chose not to lump into a heterogeneous Ethiopian super-population to emphasize potential population-specific peculiarities, are provided in Supplementary Information.

A preliminary exploration of the NAF genomes through ADMIXTURE (Figure S5) and projected PCA showed them to fall within the range of Eurasian populations, close to ancient populations with a high Anatolian Neolithic component (e.g. Anatolia_N and Minoans) (Figure 1 and S1-S4). Notably, several Jewish populations from North Africa cluster with NAF as well. The affinity between Anatolian Neolithic and NAF was further highlighted by *f3* outgroup statistic, in contrast to results obtained with the genomes before ancestry deconvolution (Supplementary Figure S6). Overall, whole-genome sequences of all the Ethiopian populations appear closer to ancient Near Eastern populations such as: Minoans, Natufian, Levant Neolithic and Anatolian Neolithic. On the other hand, their NAF components appear closer to populations with a high Anatolian rather than Levantine (such as Minoans, Sardinians and Anatolia Neolithic) component. The highest genetic affinity to the NAF components was observed among North African (Tunisian, Libyan and Moroccan) Jews (See Figure S6), as already seen in the PCA clustering (See Figures 1, S1-S4).

We further dissected the observed affinity between NAF and Anatolian Neolithic-like populations through a set of *f4* tests aimed at refining through more and more stringent comparisons the best proxy population for the Eurasian layer (Figure 2). The whole-genomes, with both African and Non-African component, are significantly closer to a Levantine ancestry rather than Anatolian (Z-Score 2.98), with them being closer to Levant_ChL individuals than Levant_N. On the other hand, NAF is shown to be closer to a Neolithic ancestry from Anatolia rather than any Levantine one (Z-score -2.847) and, among Levantine populations, notably closer to Levantine Chalcholitic than to Bronze Age groups or contemporary Lebanese. We further compare the best proxies for the Non African component using the top scoring populations from Outgroup *f3* analyses. Minoans appear to be as close to NAF as Anatolian Neolithic individuals (Z-Scores < 1). When we delved into the North African Jews signals, they broadly show affinity with NAF with particular reference to Jews from Tunisian. Similar trends were observed for

all other Ethiopian populations (Figure S7 and Table S1) and did not change when considering alternative combinations of deconvoluted components (Figure 2). Given that our ability to pinpoint the actual source of the NAF component is inherently limited by the availability of ancient and modern populations, we used qpGraph (Supplementary Figures S8,S9 and S10) and qpAdm to model NAF as a mixture of the major axes of genetic diversity that best described the Mediterranean area at the time of the studied event, following Lazaridis et al. 2016. When looking at the global genomes, our qpAdm results replicate a Levant_N origin for the Eurasian component of Ethiopians[3] (Figure 3, left column). For further results on the other Ethiopian populations see Table S2 and Supplementary Figure S11. In sum, similarly to Minoan and Tunisian Jewish populations, the non African component of Ethiopian populations can be best modelled as a mixture of ~85% Anatolian_N and ~15% CHG composition of ancestries (Figure 3, columns 2,3,4).

While this mixed ancestry component likely reached Ethiopia only within the last 3,000 years, these results should not be interpreted as involving a direct connection or descent line between Neolithic Anatolia and Ethiopia. Instead, these results can potentially be seen as informative for the identification of candidates among the available ancient and modern populations which, following geographic and chronological considerations, may be suitable proxies for one or more populations that mediated the Eurasian gene flow to East Africa. Of the ones analyzed here, Minoans and Tunisian Jews seem to provide the two closest matches to NAF, adding on top of the genetic evidence a criteria of space/time compatibility. A tentative links between these three groups may be provided by the maritime trade routes connecting Crete (home to the Minoan culture) to the Levant[8,9,10] and by the shuffling role played by a horde of nomads who navigated throughout the Mediterranean Sea 3 kya: the Sea People. These tribes left traces of their passage both in Crete, in Anatolia, when they fought the Hittite Empire and in Egypt and the Levant, and are told to have settled in the land of Canaan, known also as Palestine[11]. Interestingly, among those tribes that settled in Palestine there were: Denyen, Tjeker and Peleset. Although there are different theories around the origin of each of the tribes, there are suggestions that link the Denyen with the tribe of Dan, from which Jews from Ethiopia have been said to descend and Peleset to their neighboring Philistines[12]. The role of Sea People may therefore be crucial in explaining a temporary presence of a Minoan-like ancestry in the Levant, bringing Anatolian-like components to levels as high as 85%. A pulse of populations with Anatolian-rich ancestry has just been recently detected in Iron Age Levant, appearing and disappearing from the archaeological record within a range of few centuries[13]. Our results offer a solution to this disappearance, given that their signal may have become erased as a consequence of major warfare after 1000 BCE[14], bringing these genetic components towards Ethiopia and North Africa.

In conclusion, our work shows that when the mixing components are deeply differentiated, such as in the case of contemporary Ethiopians, ancestry deconvolution increases the sensitivity of allele sharing tests and enables to fully exploit the high quality of modern genomes.

Acknowledgments

The authors would like to thank Dr. Doron Behar and Dr. Iosif Lazaridis for fruitful discussion on an early version of this manuscript. This work was supported by the European Union through the European Regional Development Fund Project No. 2014-2020.4.01.16-0024, MOBTT53 (LM, LP).

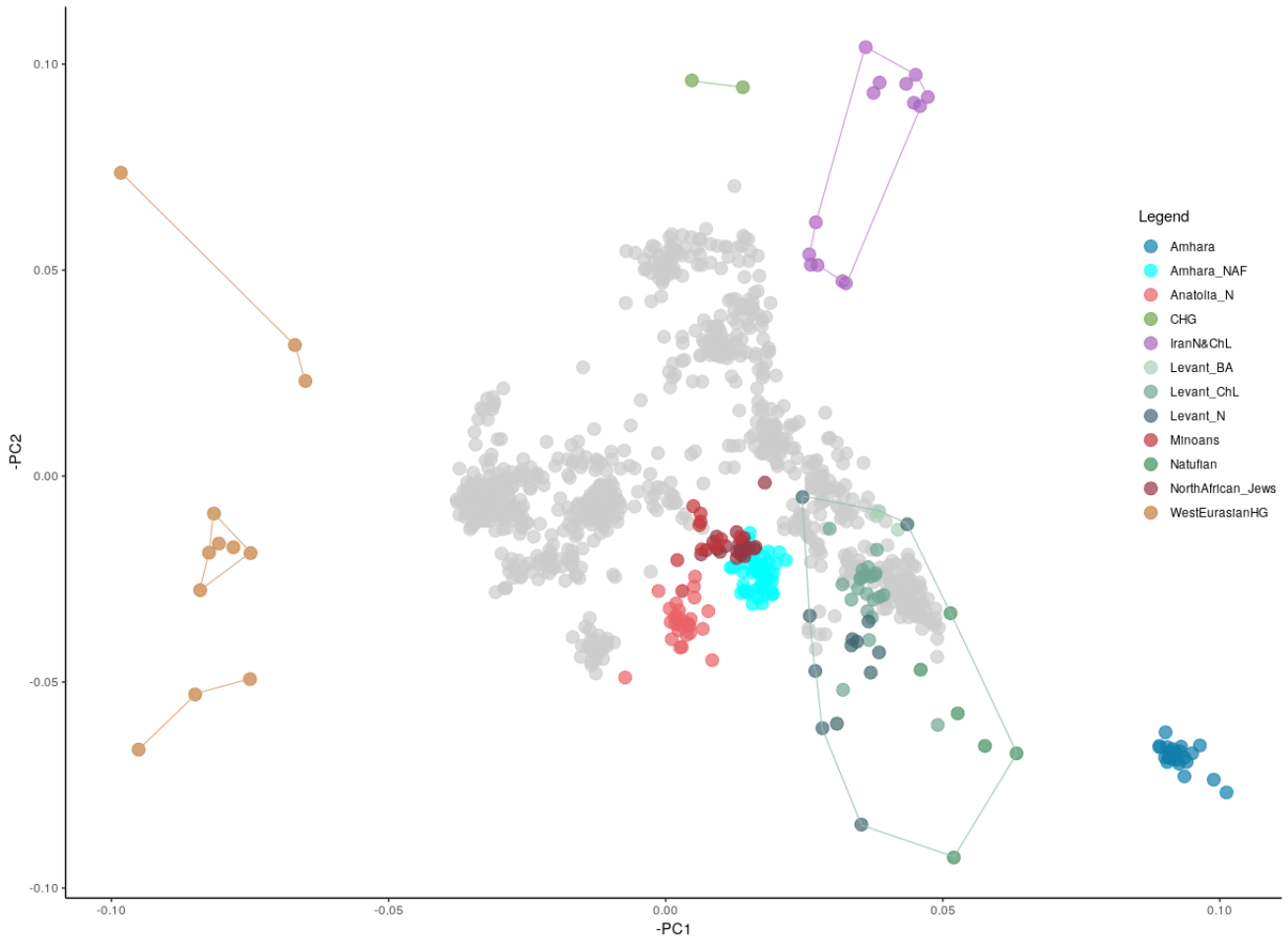

Figure 1: Principal component analysis of modern West Eurasian populations used as a scaffold (grey points) on which we projected ancient and ancestry deconvoluted genomes. To highlight the populations studied we coloured European hunther-gatheres, ancient genomes from Anatolia and Levant areas, Jews from North Africa and Amhara whole and NAF genomes. Variance explained by PC1 is 0.9% and PC2 is 0.3%

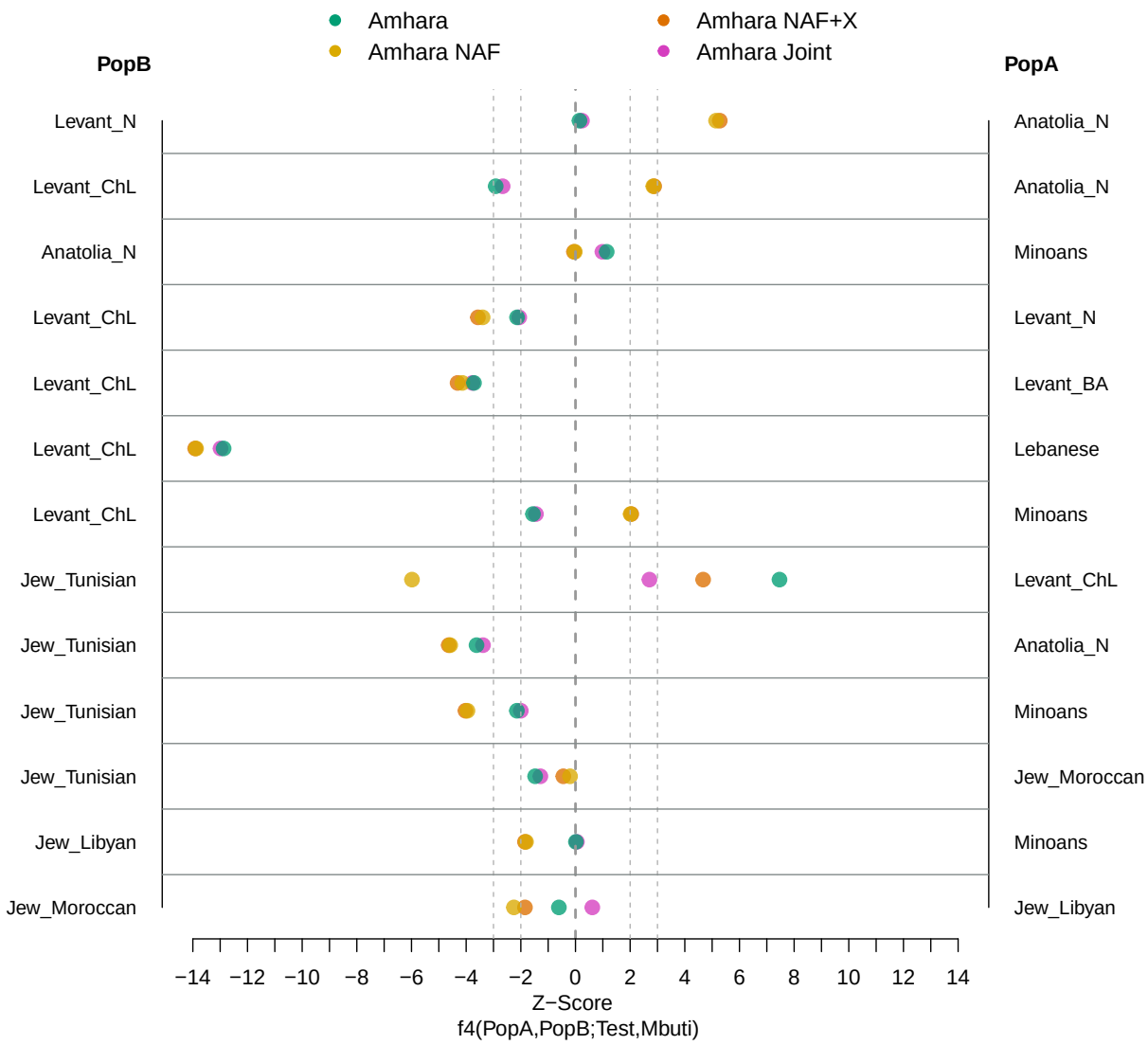

Figure 2: *f4* statistic test on Amhara in form of (PopA,PopB;Test,Mbuti) to assess genetic similarity between Amhara and respective NAF genomes to pairs of several Near Eastern populations. A and B populations are listed in the left and right side of the plot, respectively. Values in x axis indicate the Z-Scores, we draw two lines to highlight |z-Scores|= 2 and 3. Points with |z-Score| $>$ 3 indicate a clear affinity of the test population towards one of the other population. Amhara's segments tested: Amhara whole-genome (Amhara, in blue), the Non African component (Amhara NAF, in yellow), Amhara African and Non African components together (Amhara Joint, in violet) and Amhara NAF with X component (Amhara NAF+X, in orange).

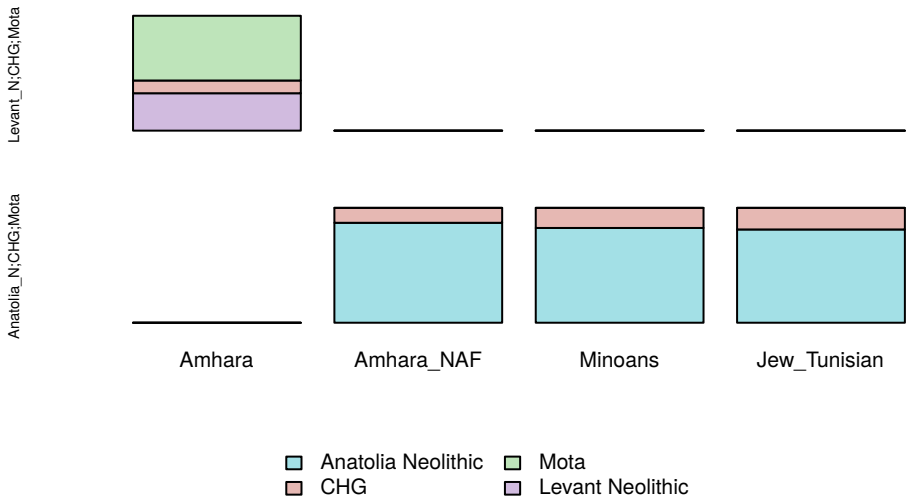

Figure 3: Modelling Amhara, Amhara_NAF, Minoans and Jews from Tunisia as a mix of Mota and Near Eastern populations, with 2 and 3 ways admixtures. Violet indicates the Levantine component, pink the Caucasus Hunter-Gatherers, light green the African component and light blue highlights the Anatolian ancestry. The left side of the graph lists the sources used to model the populations in the x axis; unfilled boxes indicate unfeasible results or p-value < 0.01.

STAR Methods

Dataset and Samples

We merged different datasets available containing both ancient and modern DNA, African and Eurasian populations from the following publications[15,16,17,3,18,19,20,21,5,22,23,24,25,26,27,28]. Northeast African populations whole-genome sequences were taken from Pagani 2015[7], and included 5 modern Ethiopian populations: Amhara, Gumuz, Oromo, Somali and Wolayta. We chose to focus on the whole genome sequence data rather than on SNP arrays[1] to increase the number of available SNPs to be compared with aDNA and other references. To maximize the number of individuals typed at each SNP, we downsampled the dataset to 1037084 markers to match the ones of Human Origin Array on which most of the ancient DNA samples were typed. For ease of exposition we chose Amhara, the population with the highest Eurasian fraction among the available ones[7], to represent all main results. We provide full description of all other Ethiopian populations in Supplementary Material. Similarly, we chose not to group all the available samples within a single "Ethiopian" population, to allow for group-specific stories to emerge.

Ancestry Deconvolution

Subsetting Modern Genomes

From phased genomes, we refined the ancestral components identification in Eastern Africans individuals provided by Pagani 2015 with PCAdmix[29]. For every 20 SNPs window of the genome, there is a probability for the window to have a source of African (AF) ancestry or Non African (NAF) ancestry (in which case the probability is 1 – AF), which is given by fbk values and refined with Viterbi algorithm[30]. We set a fbk threshold of 0.9 probability in order to assign every window to either one layer of ancestry or the other. If a window did not reach the threshold for any component, it would have been labeled as unassigned. CEU (Utah residents with ancestry from northern and western Europe) were used as a proxy for the Non African component, and Gumuz (the Ethiopian population showing minimal introgression) were used as a proxy for the African component following Pagani et al. 2015. Once the ancestral components were detected, we created the "Genomes Subsets" using the windows that reached the threshold. The "Genomes Subsets" are genomes in which for every haplotype only the confidently assigned African or Non African component is retained, while the rest is assigned as "missing data". Therefore, they are partial genomes in which only the sequences derived from a specific ancestry (either African or Non African) are present (see Yelmen et al. 2019 for further details). The ancestry deconvolution process has been applied to East African populations only from Pagani 2015 populations, namely: Amhara, Gumuz, Oromo, Ethiopian Somali and Wolayta.

Sifting through all possible ancestry fractions

To test for possible biases introduced by using CEU as proxy for the Non African component, we further divided the deconvolution results into different segments to investigate specifically the parts of the genome that were not assigned to either ancestry. We retrieved the different components from the fbk values alone, without refining them with the Viterbi algorithm, to maintain all possible segments information. For each of the two ancestries we obtained two components: X and Y, which held the sequences assigned with 51-90% and 10-50% respectively, representing the unassigned sequences in the masking process. The component X is made of sequences that were not assigned to NAF, representing the unassigned segments that we expect to bear Eurasian traces along with spurious African ones; the component Y is made of segments which we expect to be characterized mainly by African traces. The X and Y segments correspond each for 7% of the genome, and we expect their contribution to the final the results to be minimal.

Principal Component and ADMIXTURE Analyses

We performed PCA as an initial screening method on the dataset with smartpca from EIGENSOFT[31,32], using the lsqproject option and autoshrink:YES. We used modern European and Near Eastern populations with minimal missingness (–geno 0.1 with PLINK[33]) to compute PCs and projected the rest of the samples included the ancient samples and the Ethiopian NAF genomes. We used ADMIXTURE[34] software to perform supervised clustering of ancient and decolvoluted genomes using as reference modern European and Near Eastern genomes along with Yoruba as African, Gumuz as East African and Han as East Asian. We used R and ggplot2 package for visualization[35,36].

Frequency-Based Allele-Sharing Analyses

We used POPSTATS[37] to calculate Outgroup *f3* statistic in the form of *f3*(Test, A, Mbuti) with Test being the Ethiopian whole-genome sequences and the NAF individuals, and A being the same set of all possible chronological and geographical proxies for the admixture. To further infer the Non-African component we used Admixtools 4.1[26]. We performed *f4* analyses using qpDstat along with the option F4:YES with this format: A,B;Test,O. As Test populations we used Ethiopian populations with non-zero contribution from the Non-African component (namely: Amhara, Somali, Wolayta and Oromo). With Admixtools we performed qpWave and qpAdm with the set of Right populations firstly defined by Lazaridis 2016, with the exception of Onge, which is not present in our analyses. Right populations used: Ust_Ishim, Kostenki14, MA1, Han, Papuan, Chukchi, Karitiana, EHG, Natufian, Switzerland_HG, WHG. We reported qpAdm results that show significance < 0.001 in qpWave, which was performed with the set of Left populations, without the Test population. We used for every analysis a custom list of Left populations to test a two-ways or a three-ways admixture. The Left populations used to perform qpAdm were selected in this order: the Test population, A and Mota for the two-ways

admixture; the Test population, A, B and Mota for the three-ways admixture. Where A stands for the top scoring populations in the Outgroup *f3* analyses and B for CHG. We reported both significative and non significative results as they might be both indicative for the purpose of our analyses. We set our threshold to accept a result as significant at 0.01. We then used the information gathered from qpAdm to build a qpGraph model. We proceeded modelling qpGraph tree starting from a simple tree topology, then adding populations of interest at each step and modifying the topology to minimize the *f2* and *f4* Z-Score values.

Bias Testing

We performed further analyses in order to detect in the unassigned sequences (X and Y components) whether important signal were lost in the deconvolution process. We compared our test populations with the *f4* statistic using this format: A,B,Test,O. As Test populations we used: Ethiopians whole genome sequences, NAF genomes, Ethiopians_J, where "J" stands for "Joint". The Joint individuals, created for each ethnic group with Eurasian contribution (Amhara, Oromo, Somali and Wolayta), are build as a synthetic population made of the NAF and AF sequences refined by the Viterbi algorithm that passed the fbk 90% threshold, and thus not yielding the unassigned segments. To the NAF and the Ethiopians_J individuals, we added the X segments, to test if the unassigned component would give different results from the Non-African component NAF alone, which would indicate presence of biases in the deconvolution step. To the Ethiopians_J individuals along with the X component we then added the Y component as well to mimic the whole-genome. As A and B we used the possible proxy populations that may have contributed to the admixture: Levant_N, Anatolia_N, Levant_ChL. We modelled the NAF along the X component with qpAdm, using the same Left and Right populations used for the main analyses to investigate how the X component can be modelled and if the NAF with the addition of X could be modelled as the Non African component, which could indicate no bias.

References

Pagani, Luca et al. (2012). "Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool". In: *American Journal of Human Genetics* 91.1, pp. 83–96. DOI: 10.1016/j.ajhg.2012.05.015.

Pickrell, Joseph K. et al. (2014). "Ancient west Eurasian ancestry in southern and eastern Africa". In: *Proceedings of the National Academy of Sciences of the United States of America* 111.7, pp. 2632–2637. DOI: 10.1073/pnas.1313787111.

Lazaridis, Iosif et al. (2016). "Genomic insights into the origin of farming in the ancient Near East". In: *Nature* 536.7617, pp. 419–424. DOI: 10.1038/nature19310.

Hodgson, Jason A. et al. (2014). "Early Back-to-Africa Migration into the Horn of Africa". In: *PLoS Genetics* 10.6, e1004393. DOI: 10.1371/journal.pgen.1004393.

Harney, Éadaoin et al. (2018). "Ancient DNA from Chalcolithic Israel reveals the role of population mixture in cultural transformation". In: *Nature Communications* 9.1, p. 3336. DOI: 10.1038/s41467-018-05649-9.

Yelmen, Burak et al. (2019). "Ancestry-specific analyses reveal differential demographic histories and opposite selective pressures in modern South Asian populations". In: *Molecular Biology and Evolution*, pp. 1–15. DOI: 10.1093/molbev/msz037.

Pagani, Luca et al. (2015). "Tracing the Route of Modern Humans out of Africa by Using 225 Human Genome Sequences from Ethiopians and Egyptians". In: *American Journal of Human Genetics* 96.6, pp. 986–991. DOI: 10.1016/j.ajhg.2015.04.019.

Moore, Frank and Lawrence E Stager (2006). "Cypro-Minoan Inscriptions Found in Ashkelon". In: *Israel Exploration Journal* 56.2, pp. 129–159.

David Ben-Shlomo, Eleni Nodarou, and Jeremy B. Rutter (2011). "Transport Stirrup Jars from the Southern Levant: New Light on Commodity Exchange in the Eastern Mediterranean". In: *American Journal of Archaeology* 115.3, p. 329. DOI: 10.3764/aja.115.3.0329.

Rüden, Constance von (2013). "Beyond the East-West Dichotomy in Syrian and Levantine Wall Paintings". In: *Critical Approaches to Ancient Near Eastern Art* Taraqji 1999. DOI: 10.1515/9781614510352.55.

Bryce, Trevor (2005). *The Kingdom of the Hittites. New Edition.* Oxford University Press. ISBN: ISBN 0-19-928132-7.

D'Amato, Raffaele and Andrea Salimbeti (2015). *Sea Peoples of the Bronze Age Mediterranean c.1400 BC–1000 BC*. Osprey Publishing. ISBN: ISBN 978-1472806816.

Feldman, Michal et al. (2019). "Ancient DNA sheds light on the genetic origins of early Iron Age Philistines". In: *Science Advances* 5.7. DOI: 10.1126/sciadv.aax0061.

Hayes, John H. and Jeffrey K. Kuan (1991). "The Final Years of Samaria (730-720 BC)". In: *Biblica* 72.2, pp. 153–181.

Behar, Doron M. et al. (2010). "The genome-wide structure of the Jewish people". In: *Nature* 466.7303, pp. 238–242. DOI: 10.1038/nature09103.

Behar, Doron M. et al. (2013). "No Evidence from Genome-Wide Data of a Khazar Origin for the Ashkenazi Jews". In: *Human Biology* 85.6, pp. 859–900. DOI: 10.3378/027.085.0604.

Lazaridis, Iosif et al. (2014). "Ancient human genomes suggest three ancestral populations for present-day Europeans". In: *Nature* 513.7518, pp. 409–413. DOI: 10.1038/nature13673.

Haber, Marc et al. (2017). "Continuity and Admixture in the Last Five Millennia of Levantine History from Ancient Canaanite and Present-Day Lebanese Genome Sequences". In: *American Journal of Human Genetics* 101.2, pp. 274–282. DOI: 10.1016/j.ajhg.2017.06.013.

Lazaridis, Iosif et al. (2017). "Genetic origins of the Minoans and Mycenaeans". In: *Nature* 548.7666, pp. 214–218. DOI: 10.1038/nature23310.

Mathieson, Iain et al. (2015). "Genome-wide patterns of selection in 230 ancient Eurasians". In: *Nature* 528.7583, pp. 499–503. DOI: 10.1038/nature16152.

The 1000 Genomes Project Consortium (2015). "A global reference for human genetic variation". In: *Nature* 526.7571, pp. 68–74. DOI: 10.1038/nature15393.

Fu, Qiaomei et al. (2016). "The genetic history of Ice Age Europe". In: *Nature* 534.7606, pp. 200–205. DOI: 10.1038/nature17993.

Haak, Wolfgang et al. (2015). "Massive migration from the steppe was a source for Indo-European languages in Europe". In: *Nature* 522.7555, pp. 207–211. DOI: 10.1038/nature14317.

Mallick, Swapan et al. (2016). "The Simons Genome Diversity Project: 300 genomes from 142 diverse populations". In: *Nature* 538, p. 201.

Lipson, Mark et al. (2017). "Parallel palaeogenomic transects reveal complex genetic history of early European farmers". In: *Nature* 551, p. 368.

Patterson, Nick et al. (2012). "Ancient admixture in human history". In: *Genetics* 192.3, pp. 1065–1093. DOI: 10.1534/genetics.112.145037.

Raghavan, Maanasa et al. (2013). "Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans". In: *Nature* 505, p. 87.

Jones, Eppie R et al. (2015). "Upper Palaeolithic genomes reveal deep roots of modern Eurasians". In: *Nature Communications* 6, p. 8912.

Brisbin, Abra et al. (2012). "PCAdmix: Principal Components-Based Assignment of Ancestry Along Each Chromosome in Individuals with Admixed Ancestry from Two or More Populations". In: *Human Biology* 84.4, pp. 343–364. DOI: 10.3378/027.084.0401.

Ryan, Matthew S. and Graham R. Nudd (1993). *The Viterbi Algorithm*. Tech. rep. Coventry, UK, UK.

Patterson, Nick, Alkes L Price, and David Reich (2006). "Population structure and eigenanalysis." In: *PLoS genetics* 2.12, e190. DOI: 10.1371/journal.pgen.0020190.

Price, Alkes L. et al. (2006). "Principal components analysis corrects for stratification in genome-wide association studies". In: *Nature Genetics* 38.8, pp. 904–909. DOI: 10.1038/ng1847.

Purcell, Shaun et al. (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses." In: *American journal of human genetics* 81.3, pp. 559–75. DOI: 10.1086/519795.

Alexander, D H, J Novembre, and K Lange (2009). "Fast model-based estimation of ancestry in unrelated individuals". In: *Genome Research* 19.9, pp. 1655–1664. DOI: 10.1101/gr.094052.109.

Wickham, Hadley (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN: 978-3-319-24277-4.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria.

Skoglund, Pontus et al. (2015). "Genetic evidence for two founding populations of the Americas". In: *Nature* 525.7567, pp. 104–108. DOI: 10.1038/nature14895.