BMC Genomics

RESEARCH

Open Access

The genetic scenario of Mercheros: an under-represented group within the Iberian Peninsula

André Flores-Bello†, Neus Font-Porterias†, Julen Aizpurua-Iraola, Sara Duarri-Redondo and David Comas*

Abstract

Background: The general picture of human genetic variation has been vastly depicted in the last years, yet many populations remain broadly understudied. In this work, we analyze for the first time the Merchero population, a Spanish minority ethnic group that has been scarcely studied and historically persecuted. Mercheros have been roughly characterised by an itinerant history, common traditional occupations, and the usage of their own language.

Results: Here, we examine the demographic history and genetic scenario of Mercheros, by using genome-wide array data, whole mitochondrial sequences, and Y chromosome STR markers from 25 individuals. These samples have been complemented with a wide-range of present-day populations from Western Eurasia and North Africa. Our results show that the genetic diversity of Mercheros is explained within the context of the Iberian Peninsula, evidencing a modest signal of Roma admixture. In addition, Mercheros present low genetic isolation and intrapopulation heterogeneity.

Conclusions: This study represents the first genetic characterisation of the Merchero population, depicting their fine-scale ancestry components and genetic scenario within the Iberian Peninsula. Since ethnicity is not only influenced by genetic ancestry but also cultural factors, other studies from multiple disciplines are needed to further explore the Merchero population. As with Mercheros, there is a considerable gap of underrepresented populations and ethnic groups in publicly available genetic data. Thus, we encourage the consideration of more ethnically diverse population panels in human genetic studies, as an attempt to improve the representation of human populations and better reconstruct their fine-scale history.

Keywords: Merchero population, Haplotype-based methods, Genome-wide autosomal data, Genetic origin, Minority ethnic group, Iberian Peninsula, Uniparental markers

Background

Genetic studies focused on human populations have been intensely boosted in the last decades owing to the increasingly affordable genome-wide platforms, expanding the amount of high-quality genetic material considered in the studies, and allowing their availability in public genomic repositories. This has enabled to vastly depict the human genetic landscape both from an anthropological and biomedical point of view [1–3]. Yet, many human groups have been systematically underrepresented in such studies, leading to an incomplete picture of the human genetic diversity. Recently, new efforts to include minority ethnic groups represent a step forward to reduce these disparities and towards the understanding of the fine-scale human population history [4, 5].

*Correspondence: david.comas@upf.edu
†André Flores-Bello and Neus Font-Porterias contributed equally to this work.
Departament de Ciències de la Salut i de la Vida, Institut de Biologia Evolutiva (CSIC-UPF), Universitat Pompeu Fabra, 08003 Barcelona, Spain

The Iberian Peninsula genetic scenario is the complex result of several demographic processes related to different historical events characterized by diverse population contacts, including multiple ethnic and geographic origins [6, 7]. The early history of the Iberian Peninsula was characterized by the presence of the Celtiberians, Iberians, Lusitanians, and Tartessians, together with a high influence of Mediterranean cultures, such as Phoenicians, Greeks, and especially the Roman Empire [8]. This was followed by the settlement of the Germanic tribes, before the Iberian Islamic periods which were characterized by a relevant contact with North African populations [9, 10]. These processes led not only to a genetic conglomerate, but also to wealthy cultural and linguistic influences that shape the present Iberian population. The Iberian Peninsula population has been analysed in many genetic studies, showing an internal heterogeneity with genetic geographic patterns following an east-to-west axis [6]. In addition, several population groups have stood out from this general context due to their particular demographic histories and genetic landscapes, as it is the case of Basque [11], Eivissa [12], and Spanish Roma populations [13, 14].

Despite the exhaustive genetic coverage of the Iberian Peninsula, being also part of the 1000 genomes project [1], some population groups have been neglected, which precludes depicting and understanding the demographic history of the entire region. One of these underrepresented groups is the Merchero population, a minority ethnic group that has been historically persecuted and socially invisibilized. They have an itinerant history, share common traditional occupations and speak *quinqui*, an unclassified language. They are nowadays distributed across the whole Iberian Peninsula, although the total number of Mercheros has not been properly estimated [15, 16]. Little is known about the origins of the Mercheros, although several untested hypotheses have been proposed. It has been suggested that Mercheros and Iberian Roma besides having a common nomadic history and cultural characteristics, share the same origin in South Asia. Other literature points to a Moorish origin, due to the presence of Muslim groups in the Iberian Peninsula from the 8th to 15th centuries. Some written records, on the other hand, mention that their origin might be found in other European nomadic groups [15, 17, 18]. Finally, Mercheros could also be the descendants of multiple non-related groups of Spanish land workers that abandoned the feudalism system during medieval times [15, 17, 18]. However, none of these contrasting hypotheses have been formally tested and a genetic approach would provide a reference landmark to the characterization of this underrepresented population group.

In the present study, we aim to genetically characterize the Merchero population in order to complete the genetic landscape of Iberian Peninsula, by analysing genome-wide autosomal and uniparental data from 25 samples. These are the first available Merchero genetic samples, which will reduce the gap of the underrepresented populations and ethnic groups in the reference panels. In order to assess the genetic origin of the Merchero population, their ancestry profiles and admixture levels with external groups are examined with allele-frequency and haplotype-based methods. In addition, we describe their demographic history and genetic substructure, through the inference of runs of homozygosity (ROHs), identity by descent (IBD) segments, and effective population size (Ne) dynamics.

Results

Mercheros fall within the genetic context of the Iberian Peninsula with modest evidence of Roma admixture

To assess the genetic origin of Mercheros, a large genome-wide database including West Eurasian and South Asian populations individuals was analyzed (see Material and Methods). We first computed a Principal component analysis (PCA) (Fig. 1A), where Merchero samples cluster together with other Spanish individuals, suggesting that they share a similar genetic profile. In the UMAP analysis (Fig. S1), individuals establish stronger local clustering within discrete populations, while Merchero individuals also overlap with the Spanish genetic samples.

The global ancestry of the Merchero population was examined with an ADMIXTURE analysis (Fig. 1B, Fig. S2). At $K=6$ (lowest cross-validation error), Mercheros and the rest of Spanish individuals share similar ancestry proportions, with the exception of the Roma-related component (in pink), which is significantly higher in Mercheros ($2.380\% \pm 2.373$ and $0.826\% \pm 1.048$, respectively, Wilcoxon test p-value $= 0.0018$). The higher Roma proportion together with a large standard deviation of this component in the Merchero's groups suggest that few Merchero individuals might have experienced gene flow from Spanish Roma. This result is also supported by the UMAP analysis of the Iberian Peninsula context, where some Merchero samples are placed between the Roma and the rest of the Spanish samples (Fig. S3).

We further explored the Merchero's genetic composition using the fine-scale haplotype-based methods ChromoPainter and fineSTRUCTURE. As previously shown in the PCA and UMAP analyses, Mercheros are genetically close to other Spanish samples and they consequently cluster in the same branch of the fineSTRUCTURE dendrogram (Fig. 2A, Fig. S4 and Table S2). However, moderate genetic substructure is observed

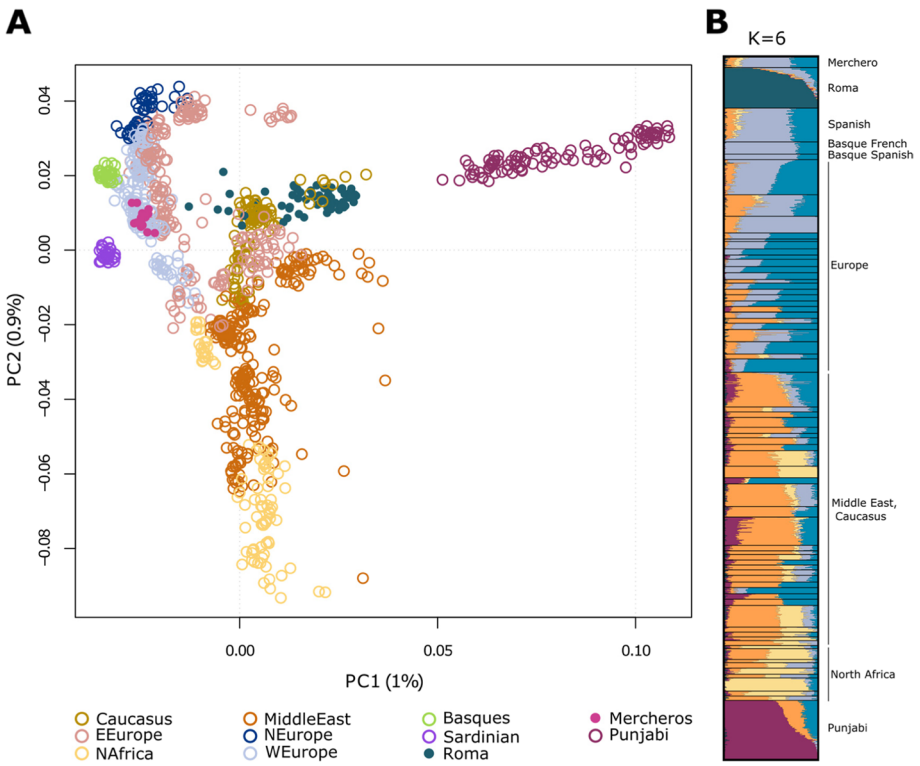

Figure 1 Genetic contextualization of Mercheros. Principal Component **(A)** and Admixture **(B)** analyses in a wide genetic context including the whole dataset are shown. Only K = 6, with the lowest cross-validation error, is shown for the ADMIXTURE analysis. Ks from 2 to 5 and the complete list of populations are presented in Fig. S2

within this group, where three genetically homogeneous clusters can be differentiated: Merchero1 in the same branch as Spanish1, Merchero2 clustering with Spanish2, and Merchero3 as an independent branch (Fig. 2A). The ancestry profiles of these clusters were obtained from the ChromoPainter coancestry matrix with the "non-negative least squares" (NNLS) method. Both Merchero1 and Merchero2 groups show a similar ancestry composition as Spanish1 and Spanish2, respectively (Fig. 2B). However, Merchero3 profile reveals a non-negligible contribution of Roma-related ancestry (~4.5%), confirming the presence of Roma gene flow in some Merchero samples (Fig. 2B). In fact, an AMOVA analysis shows the higher heterogeneity of the Merchero3 group (Table S3). In order to test for an admixture event, GLOBETROTTER was applied to the Merchero3 cluster; however, the result provides an *unclear signal*, and no estimates of an admixture event were available.

The uniparental lineages identified in the Merchero samples are similar to those found in other European populations [19, 20], without distinctive influence from South Asian or North African contributions. The most common mitochondrial haplogroup is H, (present in almost 40% of the Merchero sequences), with H1 being the most common H sublineage. For the Ychr, R1b haplogroup is the most frequent lineage within the dataset (up to 61.5%), consistent with previous observations of uniparental markers in Europe and the Iberian Peninsula [19, 20] (Table S1, S4-S5, Fig. S5).

These results show that Mercheros present a genetic ancestry profile similar to the general Spanish population. A modest and uneven Roma gene flow into Mercheros is detected, suggesting that the genetic admixture between both groups is subtle and not a widespread and inherent characteristic of these populations.

Mercheros show limited genetic isolation coupled with intrapopulation heterogeneity

In order to describe the population-specific genetic patterns of Mercheros and assess their demographic history, ROHs and IBD segments were explored.

High ROHs values of total number and length (>4 Mb) are observed in Mercheros (Fig. S6), which are larger than those previously observed in Roma [13, 21]. This suggests signals of recent inbreeding in the Merchero population. However, it is noteworthy that few samples (18%) are responsible for these high ROHs values as shown by the standard deviations and the proportion of

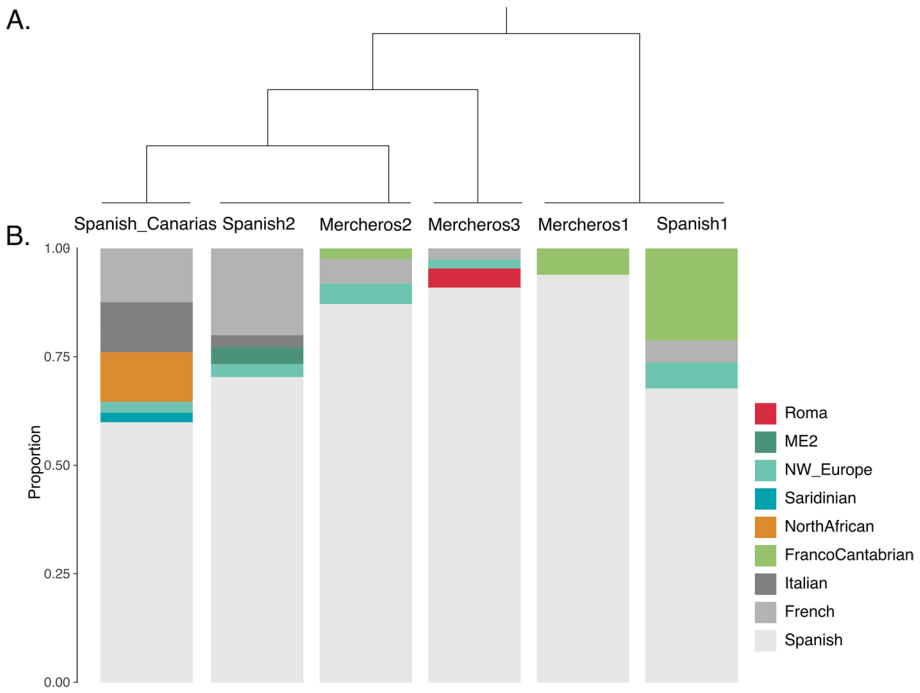

Fig. 2 Intrapopulation genetic structure and ancestry components. **A**. Spanish branch of the global fineSTRUCTURE dendrogram. The complete dendrogram is shown in Fig. S4. **B**. Ancestry profile proportions inferred from the NNLS analysis

Table 1 Intrapopulation genetic patterns. Sum of ROH length in Mb and IBD in cM for each population cluster. Only ROHs > 1 Mb and IBD segments > 3 cM were included. IBD pairs summing more than 72 cM or less than 12 cM were excluded [22]

Population	Median sum ROHs (Mb)	Median sum IBD (cM)
Merchero1	19.726	13.587
Merchero2	17.05	14.971
Merchero3	36.906	12.235
Basque	44.348	35.327
Roma	63.198	57.181
Spanish	21.918	15.438
CEU	19.657	15.179
YRI	8.196	14.862

individuals represented in the longest ROH categories. In fact, when dissecting the ROH values by the genetic clusters inferred from fineSTRUCTURE, Merchero3 shows the highest, but not statistically significant, median ROH length, immediately below Roma and Basques, whereas Merchero1 and Merchero2 clusters show values close to Spanish and CEU (Table 1, Table S6). Nevertheless, the median cumulative IBD length within groups are similar in all Merchero clusters (without significant differences), and close to Spanish and CEU values (Table 1, Table S6). Therefore, these results show that the observed signals of recent inbreeding come from sample-specific values within the Merchero3 cluster. However, this result should be taken with caution due to the low sample size of Merchero3 cluster.

The global pairwise sharing of IBD segments was explored to test for the genetic relationship among populations. These analyses mirror the results presented above within the Iberian Peninsula context (Fig. S3). Spanish Roma and Basques show large amounts of IBD segments shared internally, in contrast to what is shown in non-Roma Spanish and Mercheros (Fig. S7). Interestingly, the probabilities of sharing IBD between all Merchero clusters and Spanish are similar to the probability of Spanish samples internally sharing IBD segments. This result rejects the hypothesis that Mercheros originated from a founder event from a subset of Spanish individuals. Moreover, the probabilities involving Roma are higher for the Merchero3 cluster, suggesting a genetic closeness between these groups (Table S7). Since the length of the IBD fragments is inversely related to age due to recombination processes, several time depths were explored. Total IBD was decomposed by different length intervals to examine the genetic scenario along the corresponding time periods (Fig. 3). A stable scenario is observed over time regarding short and medium bins of IBD length

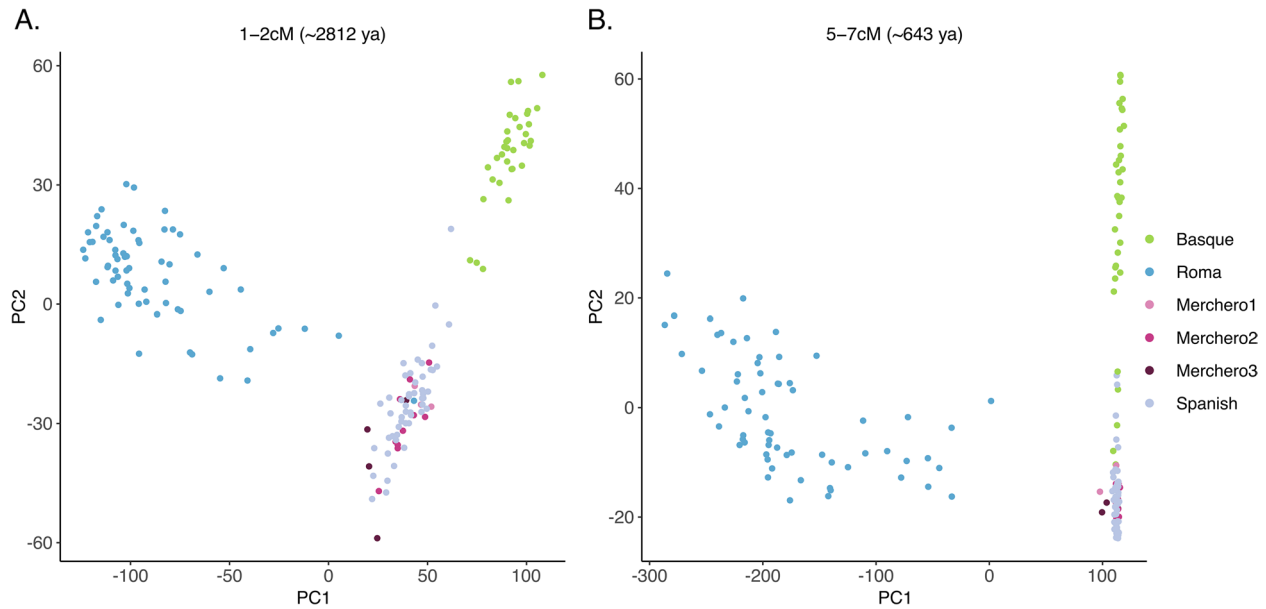

Fig. 3 The genomic scenario in the Spanish context over time. The most informative results are shown in a PCA for the IBD length bins of 1–2 cM (**A**) and 5-7 cM (**B**), corresponding to ~ 2812 and ~ 643 years ago, respectively

(< 5 cM), where the Merchero samples fall homogeneously within the Spanish context, whereas Spanish Roma and Basques depict their singular genetic differentiation (Fig. 3A). However, considering long bins (5–7 cM) a slight affinity to Spanish Roma is observed in some Merchero samples, corresponding to a period of time around six centuries ago (Fig. 3B).

These results, in agreement with the previous section, suggest a very recent scenario of genetic inbreeding and substructure in Mercheros, that might be the result of a subtle admixture with an already inbred population such as Roma (Fig. 2B).

We further examined the genetic demographic history of Mercheros focussing on their Ne dynamics. Long-term Ne (200 to 2000 generations ago) was estimated from genome-wide patterns of linkage disequilibrium, showing slightly lower values in Mercheros than in the Spanish general population (Fig. S8A), although sharing a similar trend through time (Fig. S8B). Recent Ne dynamics (0 to 200 generations ago) were next investigated from IBD segments. Mercheros show a Ne decrease contemporaneous to the Black Death (20 generations ago) without recovery, whereas the Spanish group experiences a population reduction around the same age followed by an increase in their Ne (Fig. S9). Although this result might point to a continuous decline in Mercheros Ne, it is important to note that these values are estimated assuming an homogeneous population [23], and as mentioned above, Mercheros present a varying amount of IBD segments, especially in the largest categories. Thus, this result is probably reflecting the Ne trend of some individuals, not the entire population. However, no estimates of the recent Ne can be tested separately for each Merchero group inferred with fineSTRUCTURE due to the low sample size of the genetic clusters.

Discussion

Several hypotheses have been suggested regarding the origin of Mercheros. However, these hypotheses have not been supported by robust data and they have mostly spread by word of mouth [15]. Our study provides for the first time genetic data from Merchero volunteers and suggests that a common genetic origin with other Spanish groups is the most plausible hypothesis as shown by their genetic affinity and the presence of uniparental lineages commonly found in western European populations (Fig.1, Fig.2, Fig. S1, Fig. S2, Fig. S3 and Fig. S4, Table S1, S4-S5.). Nevertheless, the scarce literature and the lack of other genetic studies focused on the Merchero population challenges the contextualization and interpretation of the analyses and results. Thus, no details can be reported about the specific period and source through which the present Merchero population originated (e.g. landworkers that abandoned the feudalism system during medieval times [15, 17, 18]).

The common origin between Mercheros and Spanish Roma has been suggested due to their cultural similarities, such as a nomadic history and common traditional

occupations [15, 17, 18]. Moreover, in spite of being an unclassified language, the *quinqui* exhibits an influence from Spanish and Portuguese Romani (*Caló*) and Basque Romani (*Erromintxela*) [15, 16]. However, our study shows no evidence of a common origin between these groups. Instead, moderate and uneven admixture with Roma is suggested regarding the ancestry proportions inferred in the ADMIXTURE and NNLS analyses (Fig. 1, Fig. 2, Fig. S2, Fig. S3B). No clear admixture events were detected with GLOBETROTTER, supporting that the level of gene flow is too limited to provide robust estimates. Nonetheless, the study of IBD segments shows that those Merchero samples with Roma-related ancestry (Mercheros3; $n=4$) present a higher probability of randomly sharing segments (Table S7), and the analysis of bins pointed out a very subtle genetic affinity between Mercheros and Spanish Roma in the longest segments of IBDs (Fig. 3). These IBD lengths suggest a contact period which is consistent with the arrival of the Roma people to the Iberian Peninsula, circa 1425 CE [24].

The Merchero population does not represent a genetically isolated and homogenous group within the Iberian Peninsula, which is consistent with previous literature about the social aspects of this population [15, 17, 18]. We show that their origin cannot be explained by a founder effect from the Spanish general population and we describe multiple evidences of intrapopulation substructure. Merchero samples are scattered within the rest of Spanish samples, both in allele-frequency and haplotype-based methods, where different genetic clusters can be identified (Fig. 1 and Fig. 2). These Merchero subgroups have different levels of individual inbreeding: those samples with traceable Roma ancestry display higher number and lengths of ROHs than the rest (Table 1). However, all groups show similar IBD values, lower than those of Roma and Basque populations and comparable with the general Spanish population (Table 1). These results suggest that Mercheros, although being a minority ethnic group and historically persecuted [15, 17, 18], can be defined as a genetically heterogeneous and open population. However, they show a slightly lower genetic diversity compared to the rest of the Spanish population, as shown by the lower Ne values.

Conclusions

The present study represents the first comprehensive genetic analysis of the Spanish minority ethnic group known as Mercheros. Since ethnicity is influenced by both genetic ancestry and cultural identity, further studies are needed to reconstruct the history of the Merchero population from different scientific disciplines, such as linguistics, anthropology, and genetics. Despite the large amount of publicly available genetic data, there is a considerable gap regarding underrepresented populations and ethnic groups. Overcoming this limitation is pivotal to properly expand the diversity frame of human population genetic studies. This will broaden the knowledge about the complete human population history and the demographic processes that have shaped the genetic variation in fine-scale details, and thus, to improve and facilitate the interpretation of biomedical studies as well [4, 5]. Therefore, we encourage the inclusion of more underrepresented populations and ethnic groups to expand our knowledge about human populations and thus, reduce the overgeneralization, especially regarding the current European ascertainment bias.

Materials and methods

Samples, genotyping, and quality control

Twenty five samples were collected from self-reported Merchero volunteers, whose written informed consent was obtained. The parents and grandparents of all volunteers self-identified as Mercheros from several locations in Spain. Moreover, eleven new Spanish Roma samples were collected in order to test for a common origin between Mercheros and Roma, as previously suggested [15]. DNA extraction was performed from saliva samples using a standard protocol, and they were genotyped with the Axiom™ Genome-Wide Human Origins Array (~629,443 variants). Genotype calling was performed with the software Axiom™ Analysis Suite 4.0 following the Affymetrix Best Practices Workflow. Twenty samples passed the process with an averaged quality control call rate of 99.7%. Furthermore, 600,225 autosomal single-nucleotide polymorphisms (SNPs) passed the recommended thresholds and were exported to perform the quality control filtering by using PLINK v1.9 software [25]. A filter for more than 10% missing SNPs per individual was applied, resulting in no sample exclusion. Then, SNPs that were missing in more than 5% of the samples, with an extreme deviation from Hardy-Weinberg equilibrium ($p<10^{-5}$), and a minor allele-frequency (MAF) below 0.05 were filtered out, remaining a total of 599,176 SNPs. Linkage disequilibrium (LD) pruning was performed with a window size of 200 SNPs, a sliding shift of 25 SNPs, and an r^2 of 0.5 keeping 231,582 SNPs. Three Merchero samples were removed from the dataset due to high level of relatedness (third degree; PI_HAT > 0.125).

To have a reference panel, our dataset was merged with the West Eurasian and North African samples from Lazaridis et al. 2016 study [3]; Utah residents with Northern and Western European ancestry (CEU), Punjabis (PJL), and Yoruba (YRI) from 1000G [1]; and 55 additional Spanish Roma samples [26]. The final dataset includes 1349 samples, 449,160 linked SNPs, and 195,541

unlinked SNPs, for haplotype- and allele-frequency based analyses, respectively.

Wide- and fine-scale population structure analysis

A PCA was performed using the SmartPCA program from the EIGENSOFT 6.0.1 package [27]. Uniform manifold approximation and projection (UMAP) analysis was run from the first ten PCs with the *umap* method [28] in R [29] with combinations of different "number of neighbors" (2, 5, 10, 20, 50, 100, 200) and "minimum distance" (0.1, 0.25, 0.5, 0.8, 0.99). Model-based individual ancestries were explored with ADMIXTURE v1.3 software [30]. The unsupervised method was run for K ancestral components from 2 to 10 using random seeds in ten independent iterations. Pong v1.4.7 software [31] was used with default parameters to obtain the major modes in the ADMIXTURE results.

In order to better assess population structure and detect fine-scale ancestry patterns, haplotype-based methods were applied. SHAPEIT v2 [32] was first used to perform the phasing of the data, using the HapMap GRCh37 genetic map [2] and the 1000G dataset as a reference panel [1]. The data was aligned to the reference and the mismatched SNPs were removed, then the proper phase inference was performed. Secondly, ChromoPainter v2 was used to infer the total length and count of haplotype fragments shared between individuals [33]. ChromoPainter was first run to estimate the global mutation probability and the switch rate parameters by running 15 iterations of the expected-maximization (EM) algorithm over chromosomes 1, 4, 17, and 20. These inferred values were averaged across the four chromosomes and samples. The parameters were used to run ChromoPainter for all chromosomes and individuals in order to obtain the final coancestry matrices of count and length sharing. Next, matrices across all chromosomes were summed by using ChromoCombine to obtain the copying profile for each individual and the C parameter required for running fineSTRUCTURE [33]. FineSTRUCTURE v2.1.0 was launched in order to cluster the data obtained from ChromoPainter into homogeneous genetic groups. The analysis included 2 million MCMC iterations, with 1 million burn-in iterations and sampling values from the posterior probability every 10,000 iterations. Three different runs of the analysis were performed for three different seeds to check the consistency of the analysis. An analysis of molecular variance (AMOVA) was performed with the *poppr* v2.1.0 R package [34] to analyse the homogeneity of these groups. *P*-values were obtained through the empirical distribution under the null hypothesis from 1000 permutations with the *ade4* v2.0.1 R package [35]. After grouping all individuals into genetically homogeneous clusters, ChromoPainter was run for these clusters and ancestry profiles were estimated using the NNLS method implemented in *nnls* v1.4 R package [36]. GLOBETROTTER [37] was used to test for admixture events. Following the recommended procedure, null.ind parameter was set to 1 in order to test for plausible admixture events using 100 bootstrap resamples. Then, null.ind parameter was set to 0 to characterize the event and estimate the admixture sources, proportions and dates. Finally, confidence intervals for the dates were inferred through 100 bootstrap iterations, considering one- and two-date admixture models.

Inbreeding estimation: ROHs and IBD segments

ROHs were identified using PLINK v1.9b software [25] with the following non-default parameters: maximum gap between SNPs of 100 kb and a minimum threshold of 500 kb and 50 SNPs. This analysis was performed using a set of reference populations to enable an informative comparison: Spanish as representative of the general Iberian Peninsula context; Basque and Roma as examples of ancient and recent inbreeding, respectively [11, 13, 21]; CEU and YRI which represent the genetic diversity with and without the Out-of-Africa bottleneck, respectively. We performed Wilcoxon tests to assess whether the sum of ROH lengths between each pair of populations was statistically significant.

IBD segments were identified using IBDSeq (version r1206) software [38] with default parameters. HapMap GRCh37 genetic map [2] was used to convert base pairs to genetic positions in cM. To construct an IBD heatmap, we first excluded IBD segments shorter than 3 cM [39], we summed the IBD pairwise lengths between individuals and removed those individual pairs whose IBD sharing was lower than 12 cM and higher than 72 cM [39]. We then constructed a heatmap with the *ggplot2* R package [40] showing theIBD sharing between pairs of individuals. Wilcoxon tests were performed to statistically assess the significance of the differences between "within-population" IBD sharing lengths of each pair of populations. We next performed PCA with *prcomp* function in R base [29] with the sum of IBD pairwise lengths computed separately using IBD segments with 1–2 cM and 5–7 cM to gain insights on the temporal population structure. To approximate the expected age of the IBD segments in these two bins (1-2 cM; 5-7 cM), we applied the formula described previously by Byrne et al. [41], assuming a generation time of 25 years. Although the length of IBD segments represents wide time distributions, these expected values are point estimates that can be used as a reference. In addition, we computed the probability that an individual selected at random from one population shares an IBD pairwise length greater than 7 cM with an individual selected at random from another population, after

excluding IBD segments lower than 3 cM, as previously described [42].

Effective population size (ne) inference

To estimate the Ne trajectories (0 to 200 generations ago), we used IBDNe (version 23Apr20.ae9) software [23] with default parameters with the IBD segments identified with IBDSeq [38]. In addition, long-term Ne values (200 to 2000 generations ago) were estimated using the NeON v1.0 R package [43] with default parameters.

Uniparental markers analyses

Mitochondrial DNA (mtDNA) was PCR-amplified in four fragments using 4 pairs of primers under identical conditions [44]. Genetic libraries were prepared and sequencing was performed by following the *Illumina mtDNA Genome* [45] and *Illumina MiSeq* guidelines [46], respectively. Mapping was performed according to GATK best practices [47] and individual haplogroups were assigned with HaploGrep [48]. Only twenty three out of twenty five mtDNA sequences passed the following quality thresholds: a minimum of 15X of coverage in all four amplified regions and a HaploGrep quality score of at least 80% [49] (Table S1).

Seventeen STRs present in the commercially available AmpFLSTR Yfiler PCR Amplification Kit (Thermo Fisher Scientific) were genotyped following manufacturer's recommendations. Y-chromosome haplogroups were predicted from YSTRs using Whit Atheys' haplogroup predictor (http://www.hprg.com/hapest5/) [50].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-021-08203-y.

Additional file 1. Supplementary Information file. Additional Figures (Fig. S1-S9) and Tables (Table S1- S7) with their references.

Acknowledgements

The authors acknowledge the participation of the volunteers who have been involved in the sampling process. We want to specially mention the invaluable help from María Remedios García Grande and Benigno Varillas. We also thank Mònica Vallés for her technical support.

Authors' contributions

AF-P, NF-P and DC conceptualized and designed the study. AF-P, NF-P, JA-I, and SD processed the data and conducted the analysis. AF-P and NF-P interpreted the results and wrote the manuscript. All authors read and approved the final manuscript.

Funding

This study was supported by the Spanish Ministry of Science, Innovation and Universities (MCIU) and the Agencia Estatal de Investigación (AEI) grant number PID2019-106485GB-I00/AEI/https://doi.org/10.13039/501100011033, and "Unidad de Excelencia María de Maeztu" (AEI,CEX2018–000792-M). NF-P was supported by a FPU17/03501 fellowship.

Availability of data and materials

Genome-wide array data and whole mtDNA sequences have been deposited at the European Genome-phenome Archive (EGA), under accession number EGAS00001005360.

Declarations

Ethics approval and consent to participate

Written informed consent was also obtained from these samples, and they were interviewed to confirm that their four grandparents were self-reported Mercheros. This procedure was supported by the corresponding IRB approvals (***Comitè Ètic d'Investigació Clínica*** (CEIC)-***Parc de Salut Mar***, ***Institut Hospital del Mar d'Investigacions Mèdiques***, Barcelona, 2016/6723/I and 2019/8900/I) and all methods adhered to the tenets of the Declaration of Helsinki.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 19 July 2021 Accepted: 18 November 2021
Published online: 15 December 2021

References

1. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56–65.
2. International HapMap Consortium. The international HapMap project. Nature. 2003;426(6968):789–96.
3. Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, et al. Genomic insights into the origin of farming in the ancient Near East. Nature. 2016 25;536(7617):419–24.
4. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet. 2019 Apr;51(4):584–91.
5. Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature. 2016 13;538(7624):161–4.
6. Bycroft C, Fernandez-Rozadilla C, Ruiz-Ponte C, Quintela I, Carracedo Á, Donnelly P, et al. Patterns of genetic differentiation and the footprints of historical migrations in the Iberian Peninsula. Nat Commun. 2019 Feb 1;10(1):551.
7. Olalde I, Mallick S, Patterson N, Rohland N, Villalba-Mouco V, Silva M, et al. The genomic history of the Iberian Peninsula over the past 8000 years. Science. 2019 Mar 15;363(6432):1230–4.
8. Institución Fernando el Católico. Acta Paleohispánica IX. Paleohispánica Rev Sobre Leng Cult Hisp Antig. 2005;5.
9. Melo Carrasco M, Vidal CF. A 1300 años de la conquista de al-Andalus (711–2011): Historia, cultura y legado del Islam en la Península Ibérica. Centro Mohammed VI para el diálogo de civilizaciones: Coquimbo-Chile; 2012. 569 p.
10. Orlandis RJ. Historia del Reino Visigodo Español. 2nd ed. Rialp: Madrid (España); 2006.
11. Flores-Bello A, Bauduer F, Salaberria J, Oyharçabal B, Calafell F, Bertranpetit J, et al. Genetic origins, singularity, and heterogeneity of Basques. Curr Biol. 2021;31:1–11.
12. Biagini SA, Solé-Morata N, Matisoo-Smith E, Zalloua P, Comas D, Calafell F. People from Ibiza: an unexpected isolate in the Western Mediterranean. Eur J Hum Genet. 2019 Jun;27(6):941–51.
13. Font-Porterias N, Arauna LR, Poveda A, Bianco E, Rebato E, Prata MJ, et al. European Roma groups show complex west Eurasian admixture footprints and a common South Asian genetic origin. PLoS Genet. 2019 Sep 23;15(9):e1008417.
14. Gómez-Carballa A, Pardo-Seco J, Fachal L, Vega A, Cebey M, Martinón-Torres N, et al. Indian signatures in the westernmost edge of the European Romani diaspora: new insight from Mitogenomes. PLoS One. 2013 Oct 15;8(10):e75397.

15. García-Egocheaga J. Minorías malditas: La historia desconocida de otros pueblos de España. 1st ed. Susaeta: Barcelona; 2003.
16. Eberhard DM, Simons GF, Fennig CD. Ethnologue: languages of the world. SIL international: Dallas; 2020.
17. Cavendish M. Peoples of Europe: Marshall Cavendish Corporation; 2002.
18. Bonilla K. Las minorías étnicas. Doc Soc - Rev Estud Soc Sociol Apl. 1977;28.
19. Simoni L, Calafell F, Pettener D, Bertranpetit J, Barbujani G. Geographic patterns of mtDNA diversity in Europe. Am J Hum Genet. 2000 Jan;66(1):262–78.
20. Jobling MA, Tyler-Smith C. The human Y chromosome: an evolutionary marker comes of age. Nat Rev Genet. 2003;4(8):598–612.
21. Moorjani P, Patterson N, Loh P-R, Lipson M, Kisfali P, Melegh BI, et al. Reconstructing Roma history from genome-wide data. PLoS One. 2013;8(3):e58633.
22. Dai CL, Vazifeh MM, Yeang C-H, Tachet R, Wells RS, Vilar MG, et al. Population histories of the United States revealed through fine-scale migration and haplotype analysis. Am J Hum Genet. 2020;106(3):371–88.
23. Browning SR, Browning BL. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. Am J Hum Genet. 2015;97(3):404–18.
24. Leblon B. Les Gitans d'Espagne (the gypsies of Spain). Paris Press Univ Fr. 1985.
25. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007 Sep;81(3):559–75.
26. Font-Porterias N, Caro-Consuegra R, Lucas-Sánchez M, Lopez M, Giménez A, Carballo-Mesa A, et al. The counteracting effects of demography on functional genomic variation: the Roma paradigm. Mol Biol Evol. 2021.
27. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006;2(12):e190.
28. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv180203426 Cs Stat [Internet]. 2020 17 [cited 2020 Dec 9]; Available from: http://arxiv.org/abs/1802.03426
29. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013;
30. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19(9):1655–64.
31. Behr AA, Liu KZ, Liu-Fang G, Nakka P, Ramachandran S. pong: fast analysis and visualization of latent clusters in population genetic data. Bioinforma Oxf Engl. 2016 15;32(18):2817–23.
32. O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, et al. A general approach for haplotype phasing across the full spectrum of relatedness. PLoS Genet. 2014;10(4):e1004234.
33. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. PLoS Genet. 2012;8(1):e1002453.
34. Kamvar ZN, Tabima JF, Grünwald NJ. Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. PeerJ. 2014;2:e281.
35. Dray S, Dufour A-B. The ade4 package: implementing the duality diagram for ecologists. J Stat Softw. 2007;22(1):1–20.
36. M. Mullen K, H. M. van Stokkum I. nnls: The Lawson-Hanson algorithm for non-negative least squares (NNLS) [Internet]. 2012. Available from: https://CRAN.R-project.org/package=nnls
37. Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, et al. A genetic atlas of human admixture history. Science. 2014;343(6172):747–51.
38. Browning BL, Browning SR. Detecting identity by descent and estimating genotype error rates in sequence data. Am J Hum Genet. 2013;93(5):840–51.
39. Han E, Carbonetto P, Curtis RE, Wang Y, Granka JM, Byrnes J, et al. Clustering of 770,000 genomes reveals post-colonial population structure of North America. Nat Commun. 2017;8(1):14238.
40. Wickham H. ggplot2: elegant graphics for data analysis. Springer-Verlag: New York; 2016.
41. Byrne RP, van Rheenen W, van den Berg LH, Veldink JH, McLaughlin RL. Dutch population structure across space, time and GWAS design. Nat Commun. 2020;11(1):4556.
42. Ioannidis AG, Blanco-Portillo J, Sandoval K, Hagelberg E, Miquel-Poblete JF, Moreno-Mayar JV, et al. Native American gene flow into Polynesia predating Easter Island settlement. Nature. 2020;583(7817):572–7.
43. Mezzavilla M, Ghirotto S. Neon: an R package to estimate human effective population size and divergence time from patterns of linkage disequilibrium between SNPS. J Comput Sci Syst Biol. 2015;8.
44. García A, Nores R, Motti JMB, Pauro M, Luisi P, Bravi CM, et al. 15 [cited 2021 Apr 22];(ddab105). Available from. 2021. https://doi.org/10.1093/hmg/ddab105.
45. Illumina. Human mtDNA Genome Guide (15037958) [Internet]. 2016. Available from: https://emea.support.illumina.com/downloads/human_mtdna_genome_guide_15037958.html
46. Illumina. Legal information and guidelines [Internet]. 2018. Available from: https://www.illumina.com/company/legal.html
47. der Auwera GAV, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Curr Protoc Bioinforma. 2013;43(1):11.10.1–11.10.33.
48. Weissensteiner H, Pacher D, Kloss-Brandstätter A, Forer L, Specht G, Bandelt H-J, et al. HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic Acids Res. 2016;44(W1):W58–63.
49. Seo S, Mourad Assidi, Mourad Assidi, Mohamed H Al-Qahtani, Antti Sajantila, Bruce Budowle. Underlying Data for Sequencing the Mitochondrial Genome with the Massively Parallel Sequencing Platform Ion TorrentTM PGMTM. BMC Genomics. 2015;16((Suppl 1)):S4.
50. Athey TW. Haplogroup prediction from Y-STR values using a Bayesian-allele- frequency approach. J Genet Geneal. 2006;2:34–9.

Publisher's Note