PROCEEDINGS OF
THE ROYAL SOCIETY B
BIOLOGICAL SCIENCES

Dissecting human North African gene-flow into its western coastal surroundings

Journal:	*Proceedings B*
Manuscript ID	RSPB-2019-0471.R1
Article Type:	Research
Date Submitted by the Author:	n/a
Complete List of Authors:	Arauna, Lara; Universitat Pompeu Fabra Departament de Ciences Experimentals i de la Salut Hellenthal, Garrett; University College London Genetics Institute Comas, David; Universitat Pompeu Fabra Departament de Ciences Experimentals i de la Salut,
Subject:	Evolution < BIOLOGY, Genetics < BIOLOGY, Genomics < BIOLOGY
Keywords:	North Africa, Canary Islands, Iberian Peninsula, gene-flow, haplotype-based methods, fineSTRUCTURE
Proceedings B category:	Genetics & Genomics

Author-supplied statements

Relevant information will appear here if provided.

Ethics

Does your article include research that required ethical approval or permits?:
This article does not present research with ethical considerations

Statement (if applicable):
CUST_IF_YES_ETHICS :No data available.

Data

It is a condition of publication that data, code and materials supporting your paper are made publicly available. Does your paper present new data?:
My paper has no data

Statement (if applicable):
CUST_IF_YES_DATA :No data available.

Conflict of interest

I/We declare we have no competing interests

Statement (if applicable):
CUST_STATE_CONFLICT :No data available.

Authors' contributions

This paper has multiple authors and our individual contributions were as below

Statement (if applicable):
LRA participated in data analysis, design of the study and manuscript writing. DC and GH contributed with manuscript writing and design of the study. All authors gave final approval for publication.

Dissecting human North African gene-flow into its western coastal surroundings

Lara R Arauna[1], Garrett Hellenthal[2], David Comas[1*]

1. Departament de Ciències Experimentals i de la Salut, Institute of Evolutionary Biology (CSIC-UPF), Universitat Pompeu Fabra, Barcelona, Spain.
2. UCL Genetics Institute, Department of Genetics, Evolution and Environment, University College London, London, UK.

*Corresponding author: david.comas@upf.edu

Keywords

North Africa; Canary Islands; Iberian Peninsula; gene-flow; haplotype-based methods; fineSTRUCTURE

Abstract

North African history and populations have exerted a pivotal influence on surrounding geographical regions, although scant genetic studies have addressed this issue. Our aim is to understand human historical migrations in the coastal surroundings of North Africa. We built a refined genome-wide dataset of North African populations to unearth the fine-scale genetic structure of the region, using haplotype information. The results suggest that the gene-flow from North Africa into the European Mediterranean coast (Tuscany and the Iberian Peninsula) arrived mainly from the Mediterranean coast of North Africa. In Tuscany, this North African admixture date estimate suggests the movement of peoples during the fall of the Roman Empire around the 4th century. In the Iberian Peninsula, the North African component likely reflects the impact of the Arab expansion since the 7th century and the subsequent expansion of the Christian Kingdoms. In contrast, the North African component in the Canary Islands has a source genetically related to present-day people from the Atlantic North African coast. We also find sub-Saharan gene-flow from the Senegambia region in the Canary Islands. Specifically, we detect a complex signal of admixture involving Atlantic, Senegambian and European sources intermixing around the 15th century, soon after the Castilian conquest. Our results highlight the differential genetic influence of North Africa into the surrounding coast and show that specific historical events have not only had a socio-cultural impact but additionally modified the gene pool of the populations.

Introduction

North Africa is a genetically diverse region from a human population perspective. North African populations show a complex and heterogeneous genetic structure that has been described as an amalgam of at least four different ancestral components: Middle Eastern, sub-Saharan African, European, and autochthonous North African (Henn et al., 2012). Most of the genetic studies about North Africa have focused on the inner relationships among populations, or the gene-flow from nearby populations (Arauna et al., 2017; Henn et al., 2012). However, there are scant studies that have focused on North African gene-flow into neighboring regions (Botigué et al., 2013). It is well known that the surrounding coast has been historically influenced by North African peoples (Plaza et al., 2003; Secher et al., 2014); however, the demographic impact of those contacts has not been properly addressed. Our aim is to assess the North African demographic and genetic influence in nearby regions outside the African continent by assessing the gene-flow in three geographical neighboring regions with documented contacts with North Africa: the Canary Islands, the Iberian Peninsula, and Tuscany.

The Canary Islands, located in the Atlantic coast of North Africa, have been inhabited since approximately 1,000 BCE (Navarro Mederos, 1997; Secher et al., 2014). The islands were known by the Phoenicians, Greeks, and Romans; however, it is thought that there was no contact with the autochthonous settlers of the islands since the 4th century until the Castilian conquest in the 15th century (Peña, 2013). By the time of this European conquest of the Islands, the aboriginal population size has been estimated around 100,000 individuals (Rodríguez-Martin and Martín-Oval, 2009). A Northwest African origin of the first settlers of the islands is consistent with patterns of uniparental and classical genetic markers in modern and ancient samples (Fregel et al., 2009a; Maca-Meyer et al., 2004). In particular, the presence of haplogroups in the Canary Islands that are only found in individuals of North African descent, such as mitochondrial (mtDNA) haplogroup U6 (Maca-Meyer et al., 2003) and Y chromosome haplogroup M81 (Solé-Morata et al., 2017), among some others considered founder lineages, support the North African origin of the islanders. The frequencies of these haplogroups in the extant population of the Canary Islands show a clear sexual bias: the percentage of the maternal North African component estimated through the analysis of mtDNA lineages is high, between 42 and 74% (Maca-Meyer et al., 2004); while the paternal component analyzed through the study of Y-chromosome lineages is lower, between 5 and 16% (Fregel et al., 2009a). Additionally, Botigué and colleagues (2013), analyzing genome-wide data, showed a higher IBD sharing between individuals from the Canary Islands and North Africa compared to individuals from continental Europe, suggesting a higher gene-flow from the African continent to the Islands. Finally, genome-wide analysis with ancient DNA from the Canary Islands have corroborated the North African origin of the autochthonous component and its presence in current Canary Islanders (Rodríguez-Varela et al., 2017). However, the exact dates of admixture from Europe and the precise geographical origin of the North African component in the Islands

have not been addressed.

The most well documented contact between North Africa and the Iberian Peninsula is the Arab expansion, which crossed the Mediterranean and arrived to Gibraltar in the 8th century. However, ~~genetic studies, mainly~~ genetic studies based on uniparental markers, together with archaeological and anthropological evidence, have suggested previous contacts across the Gibraltar Strait that date back to prehistoric times (Bosch et al., 2001; Currat et al., 2010; Maca-Meyer et al., 2003; Plaza et al., 2003; Secher et al., 2014). Recently, ancient DNA studies have supported prehistoric migrations from North Africa into the Iberian Peninsula since around 4,000 ya (years ago) (González-Fortes et al., 2019; Olalde et al., 2019; Valdiosera et al., 2018). Moreover, mtDNA, Y-chromosome, and STRs studies (Adams et al., 2008; Capelli et al., 2009; Casas et al., 2006; Plaza et al., 2003; Regueiro et al., 2015), as well as genome-wide and ancient DNA analyses (Botigué et al., 2013; Moorjani et al., 2011; Olalde et al., 2019), have also shown gene-flow in historical times. Moorjani et al. (2011) dated African gene-flow into Southern Europe around 55 generations ago, with the highest proportions in Iberia: 3.2 ± 0.3% in Portugal, and 2.4 ± 0.3% in Spain, which was related to a demographic impact either in Roman or Arab periods. Botigué et al. (2013) showed that the inclusion of North African populations in their analyses increased those estimated percentages of gene-flow, suggesting a higher North African gene-flow in Iberia, and that the sub-Saharan gene-flow detected entered with the North African wave, challenging the interpretation of a direct sub-Saharan influence in Southern Europe. Additionally, the North African gene-flow in the Iberian Peninsula was dated to 6-10 generations ago, although previous gene-flow was not discarded. In a large study of human populations admixture, Hellenthal and colleagues (2014) described a complex scenario with continuous gene-flow during the past 2,000 years in Iberia with North and sub-Saharan Africans. In sum, although all studies agree on the genetic influence of North Africa in Iberia, there is no clear consensus in the pattern of gene-flow and the estimated dates of the North African admixture.

The presence of the Etruscans, in what it is nowadays referred to as the Tuscany territory in the Italian Peninsula, has been largely documented. However, the genetic footprint of the Etruscans in current populations has been only claimed in some isolated populations, but not in the Tuscan general population (Ghirotto et al., 2013). Moreover, although a Middle Eastern or Anatolian origin has been hypothesized for the Etruscans (Achilli et al., 2007), recent studies analyzing mitochondrial DNA have rejected an origin outside Italy (Belle et al., 2006; Ghirotto et al., 2013). Recently, an exhaustive genome-wide study of the Italian population (Fiorito et al., 2016) has dated different admixture events in Italy coming from different sources, including old events dated around 3,000 ya that involved Caucasus, Middle Eastern, and Central Italian populations; whereas other more recent admixture processes involved gene-flow from North-Central Europe around the collapse of the Roman Empire, a period which has been associated with extensive human movements. This continuous gene-flow in multiple directions at different times has yielded a complex genetic structure in the Italian Peninsula shown in both uniparental and genomewide

analyses (Boattini et al., 2013; Fiorito et al., 2016), and traces of North African influences have also been detected, although the amount and timing of such contributions to Italy have not been assessed (Busby et al., 2015; Fiorito et al., 2016).

Our aim is to assess the impact of gene-flow from North Africa to surrounding populations for which there is documented evidence of contact between the populations, in particular the Canary Islands, Iberia, and Tuscany. Previous North African – European gene-flow analyses (Botigué et al., 2013; Moorjani et al., 2011) were limited by the scant geographic distribution of North African samples available. In order to overcome these issues, we use the largest genome-wide dataset of North African samples available, including different Berber groups that have not been included in previous studies of North African gene-flow, which allows us to describe detailed and complete genetic scenarios for North African admixture into the surrounding areas. The application of haplotype-based methods to a large dataset of samples and autosomal markers might refine our knowledge on the i) estimated dates of the admixture events; ii) the specific geographic sources of the gene-flow; and, iii) the quantification of the amount of gene-flow in the three targeted populations.

Material and methods

Building the dataset

We built a dataset of more than 1,200 samples that includes European and sub-Saharan African samples from the 1000 genomes project (The 1000 Genomes Project Consortium, 2015); Iberian, Basque and Canary Islands populations from Botigué et al. (2013); and a large and diverse dataset of North African populations (which includes both Arab and Berbers and covers a wide geographic extension) from Henn et al. (2012) and Arauna et al. (2017) (see Table S1 and Figure S1). For some of the Iberian samples from the 1000 genomes project, no geographical coordinates are available; and, therefore, for some analyses they were assigned as “Iberian” without specifying the location. Both Plink versions 1.07 and 1.9 have been used depending on the analyses (Chang et al., 2015; Purcell et al., 2007). SNPs missing in more than 10% of the individuals, those that failed Hardy-Weinberg test at 0.01 significance threshold, and those with a minor allele frequency (MAF) below 0.05 were discarded. After filters, 267,475 SNPs remained for analyses. Individuals sharing more than 85% of their genome identity by state (IBS) were removed, and remaining individuals with more than 10% of missing SNPs were also excluded. For the analyses that required linkage equilibrium, SNPs were pruned using a pairwise linkage disequilibrium maximum threshold of 0.5 using a windows size of 50 a shift step of 5, remining 149,956 SNPs.

Haplotype-based Methods

Phasing

The phasing of SNPs was performed with SHAPEIT (O'Connell et al., 2014), using the population-averaged genetic map from the HapMap phase II (The International HapMap Consortium, 2003) and the 1000 genomes dataset as a reference panel (The 1000 Genomes Project Consortium, 2015). This phasing step was performed after an alignment with the reference panel and the removal of SNPs that did not align.

Commented [MOU1]: Reviewer: Page 4 Line 134. It is not clear if the samples present in the dataset were discarded from the Imputation panel. If not, Is that expected to produce any bias? Authors should discuss about it and/or remove the samples from the imputation panel.

I don't understand what he means with: "samples present in the dataset", he means the reference panel? If he refers to that they are not included in any moment with our data, just used as a reference, but this is standard procedure I don't think we need to specify that.

ChromoPainter

ChromoPainter (Lawson et al., 2012) was run to infer the genome-wide number and proportion of haplotype segments for which each individual shared with every other individual, without population specification (i.e. using all sampled individuals as both recipients and donors, -a mode). We followed the protocol analogous to that outlined in Hellenthal et al (2014). In particular, first the global mutation probability and the switch rate parameters were estimated using the EM algorithm implemented in ChromoPainter with the following parameters: -i 10 -in -iM, in chromosomes 1, 7, 14 and 20 for all individuals. The mutation probability and the switch rate parameters estimated were averaged across these four chromosomes, weighting by the number of SNPs per chromosome. The average weighted values were 0.00017 and 208.30557 for the global mutation and switch rate, respectively. ChromoPainter was run afterwards for all chromosomes using these fixed global mutation and switch rates values. The final co-ancestry matrices (i.e. *.chuncklengths.out and *.chunckcounts.out files) were summed across chromosomes.

FineSTRUCTURE

We used ChromoCombine to estimate the fineSTRUCTURE C parameter (c=0.264). Then, following Leslie et al (2015), FineSTRUCTURE v.2.0.4 (Lawson et al., 2012) was run using 2 million iterations of MCMC, sampling values every 10,000 iterations following 1 million "burn-in" iterations (i.e. -x 1000000 -y 2000000 -z 10000). Finally, the FineSTRUCTURE tree was inferred using default parameters (i.e. –m T). Three seeds were estimated in order to check robustness of the analyses. Based on the FineSTRUCTURE results, we established genetic clusters to use as "populations" for subsequent Globetrotter analyses.

Globetrotter

We applied GLOBETROTTER to identify and date admixture events in each of our target populations using genome-wide linkage disequilibrium decay patterns, under a model that assumes instantaneous admixture involving 2 or more groups at 1 or 2 times in the past, followed by random mating among individuals from the admixed population. To do so, we followed the protocol of Hellenthal et al 2014. In particular, after defining genetic clusters based on fineSTRUCTURE results (see FineSTRUCTURE clusters in Figure 2 and Table S1), we performed a separate run of ChromoPainter painting each cluster using all other clusters as donors (i.e. disallowing "self-copying" from other members of the own cluster). The clusters are assigned geographical names in order to facilitate the comprehension, however the detailed information of

the distribution of the samples contributing to each cluster can be found in table S4 and figures S2 and S4. Then, we ran Globetrotter (Hellenthal et al., 2014) using the copy vectors (i.e. *chunklength.out file) from the first ChromoPainter run used in FineSTRUCTURE (i.e. that painted each individual using all other sampled individuals) and the painting (i.e. *samples.out files) from the second ChromoPainter run (i.e. that painted each individual using all other individuals outside of their cluster). The null.ind parameter was set to 1 for all the Globetrotter analyses, as recommended, to account for decay in linkage disequilibrium that may not be attributable to genuine admixture signals. Four different target groups were tested separately for admixture: Canary_Islands, Iberian_Peninsula, Tuscany, and Basque. For each of the four targets, other clusters were used as surrogates, except that the Canary_Islands cluster was not included as a surrogate when testing the Iberian_Peninsula cluster for admixture, due to the relatively high genetic similarity between Iberian_Peninsula and the Canary Islands. We performed 100 bootstrap iterations to infer confidence intervals for date estimates, for both one- and two-date models of admixture. As a result, for each target cluster we have 200 estimates of the fit of the model, combined across one and two dates of admixture, and the estimated dates for each bootstrap. We assumed a generation time of 25 years (Laval et al., 2010).

Admixture between more than two sources at a given time is inferred by GLOBETROTTER as multiway admixture, and described as two events that each involve two sources (where each such source may comprise some unknown mixture of the genuine admixing groups). To better interpret these events, in these multiway cases we manually reviewed the coancestry curves generated for each pair of surrogate populations to establish the sources participating in the admixture process, as illustrated in Fig S7. In all these cases we found evidence for three distinct sources intermixing. In particular we assumed the three surrogates (or groups of surrogates) demonstrating the patterns in Fig S7 represented three distinct admixing sources. We represent the genetic make-up of each of these three sources in Table S3 by decomposing the GLOBETROTTER proportions estimation considering only two sources and recalculating those proportions considering the three manually inferred sources.

Results

We have compiled a dataset of more than 1,200 samples that includes a large and diverse dataset of North African populations to study the influence of North African gene flow in neighboring populations. Principal Components Analysis (PCA) of all populations in the dataset differentiates sub-Saharan and European populations along the first PC (PC1) (Figure 1). The North African samples are widely spread along the first PC reflecting high heterogeneity, in accordance with the previously described differential admixture of the subpopulations (Arauna et al., 2017; Henn et al., 2012). PC2 further differentiates North African samples and highlights the genetic diversity within North Africa. On the first two PCs, the Canary Islands samples are placed close to the Iberian

212 samples but shifted towards the Middle East and North Africa. When focusing on the European
213 samples (inner PCA in Figure 1), three largely non-overlapping clusters can be observed: the Finns;
214 Northern and Western Europeans (Great Britain and CEU); and Southern Europeans (Tuscany,
215 Iberia, Basque Country, and also the Canary Islands).

216 We used haplotype-based methods to dissect the genetic structure of the studied populations and
217 understand their genetic relationships. We performed FineSTRUCTURE analyses (Figure 2) and
218 identified three major splits separating our data: North Africa, Europe, and sub-Saharan Africa.
219 Within these major geographical clusters, several sub-clusters can be identified that suggest a finer
220 resolution of genetic structure. For example, within the European cluster, six sub-clusters are found
221 that correlate with geography: Iberian Peninsula, Tuscany, Basque, Canary Islands, Northwest
222 Europe, and Finland (Figure S1). 13 Syrian samples clustered together with the Canary Islands
223 populations and were removed from further analyses. Similarly, within sub-Saharan African
224 samples we find four sub-clusters that correspond closely with sampling locations: Luhya (from
225 Kenya), sub-Saharan Atlantic (GWD and MSL), Guinean Gulf (YRI and ESN), and North
226 Africa_sub-Saharan_ancestry, which is composed of North African samples with substantial
227 sub-Saharan admixture (as previously described in Arauna et al., 2017). In contrast, sub-clusters
228 within North Africa do not show as precise a correlation with geography, with several sub-clusters
229 containing individuals that span broad geographical areas: East, West, Central, Atlantic,
230 Mediterranean, Tunisia Chenini, and Tunisa Sened (the last two have been already described as
231 drifted populations that show high levels of relatedness (Arauna et al., 2017; Henn et al., 2012))
232 (Figures 2 and S2). Finally, a dissection of the Iberian Peninsula sub-cluster shows four minor
233 clusters: NorthWest_Iberian, South_Iberian, and two clusters without clear geographic structure
234 (Iberian_Peninsula1 and Iberian_Peninsula2) (Figure S3). One Iberian individual was an outlier (did
235 not cluster), and therefore this individual was not included in further analyses.

236 We identified and dated admixture events with GLOBETROTTER using the clusters defined in Figure
237 2 (Figure 3). We focused on Tuscany, Iberia, and the Canary Islands, three populations that
238 surround North Africa for which there is documented contact with North Africa (Boattini et al.,
239 2013; Brett and Fentress, 1997; Camps, 1995; Camps and Vela i Aulesa, 1998; Fiorito et al., 2016;
240 Naylor, 2009), in order to dissect possible admixture events between these geographical areas. We
241 also tested admixture in the Basque population, but no admixture was detected. Assuming a single
242 date of admixture per group, different times of admixture were inferred for the three populations:
243 in Tuscany the mean estimated admixture time after 100 bootstrap iterations was 485±19 CE; in
244 the Iberian Peninsula the estimated gene-flow was dated to 1,000±9 CE; and, finally, in the Canary
245 Islands the estimated date of admixture with North Africa was 1,555±7 CE (Figure S4 and Table S2).
246 However, while the data strongly supports a single event of North African admixture in Tuscany; in
247 the Canary Islands and the Iberian Peninsula a history of multiple episodes of gene-flow cannot be
248 ruled out, according to the goodness of fit test for two admixture events (Figure S5). The

GLOBERTROTTER manual notes that the program concludes "multiple dates" of admixture when its goodness of fit score for two dates relative to the fit of one date is above 0.35 which are based on simulation results –(Hellenthal et al., 2014). In our dataset, 7% and 3.5% of the bootstraps exceed 0.30 for the Canary Islands and Iberian Peninsula, respectively (Fig S5).

The sources inferred in the admixture events are also different in each of these three populations. In Tuscany, GLOBETROTTER concludes a simple admixture event between two sources (Figure 3). The major source is inferred to be related to present-day European groups, with the largest component being Iberian-like but with an additional Northwestern European-like component. The minor contributing source inferred for Tuscany relates genetically to individuals from the Mediterranean shore of North Africa, though this minor source also contains an Iberian component. In contrast, in the Iberian Peninsula we detected a more complex pattern of gene-flow of a three-way admixture between a North African-like source from the Mediterranean shore, a Basque-like source, and a European-like source with Northwest and South (Tuscany) components, possibly at different times as noted above. Finally, in the Canary Islands, admixture is detected between a European-like source, mainly related to people from the Iberian Peninsula but with some relatedness to Northwest Europeans and Tuscans, and a second source of admixture representing a composite of present-day North Africans from the Atlantic and sub-Saharan Africans from the Senegambia region.

Since the Iberian Peninsula analysis showed a complex pattern of gene-flow that could be attributed to the presence of genetic substructure, we analyzed the genetic subclusters within Iberia. Four different minor genetic clusters could be identified, as described above. The analysis of these four minor clusters allowed us to dissect the sources and dates of admixture within the Iberian Peninsula (Figure 4). GLOBETROTTER infers a single pulse of admixture for each of the Iberian_Northwest and Iberian_Peninsula2 minor clusters, with overlapping dates of gene-flow related to North African sources occurring around the 8th century (717-759 CE and 734-778 CE respectively, 95% CI). In the Iberian_Peninsula1 minor cluster, the inferred date of North African related admixture is around the 11th century (1027-1058, 95% CI), while for the Iberian_South minor cluster, GLOBETROTTER dates admixture to the second half of the 14th century (1330-1356, 95% CI). However, in the last two cases, again multiple episodes of gene-flow cannot be ruled-out (Figure S6), and thus Figure 4 may reflect dates of more recent gene-flow and mask older gene-flow. In all Iberian clusters, GLOBETROTTER infers a North African-like source that mainly relates to our Mediterranean cluster. However, Iberian_Northwest and Iberian_Peninsula2 (which are the clusters for which GLOBETROTTER infers older, single pulses of admixture), also show a North African West-like component (Table S3).

In summary, the North African gene-flow detected in the three geographical areas analyzed (Tuscany, Iberia, and the Canary Islands) differ not only in the estimated dates of admixture, but also in the sources of admixture and amount of DNA inherited for each source. In particular

Tuscany and Iberia show admixture from a Mediterranean-like source, while the Canary Islands show admixture from an Atlantic North African-like source (Figure 2, Table S2).

Discussion

The aim of our study was to dissect gene-flow from North Africa to three surrounding coastal areas that have been documented to have had historical contact with North Africans: Tuscany, Iberia, and the Canary Islands. We applied haplotype-based methods on a large sample set using genome-wide markers in order to refine our knowledge of the gene-flow between these geographical areas, focusing on the following: i) the estimated dates of the admixture, ii) the geographical origins of the sources of the admixture events, and iii) the proportions of the gene-flow. The extensive dataset and the use of haplotype-based methods allowed us to estimate precise and narrow confidence intervals for admixture dates which we correlated with historical processes. Different estimated times, sources, and proportions of admixture were detected in each of the three populations analyzed.

While all three populations show evidence of admixture between European-like and North African-like source groups, the geographic characterization of the North African source varies across populations. In particular, the North African source in the Canary Islands is more genetically similar to populations along the Atlantic coast, while the North African source in Iberia and Tuscany is more genetically similar to populations along the Mediterranean Coast.

In the Canary Islands, our date of admixture corresponds to the time of the Castilian conquest (15th century). The European contribution is mainly Iberian, but it also shows a small amount of Northwest European genetic influence, which might be related to the presence of Normans involved in the first steps of the conquest (Reverón, 1944). The African source shows both a North African component from the Atlantic and a sub-Saharan component from Senegambia.

The mixture of the Atlantic and Senegambia components in the Canary Islands could be explained by admixture at different times prior to European contact. Our data suggest that the initial settlers of the Islands may have already been a composite of these two components. This scenario is supported by the presence of sub-Saharan mitochondrial lineages (i.e. L haplogroups) (Fregel et al., 2009b, 2015; Maca-Meyer et al., 2004; Ordóñez et al., 2017) in ancient Canary samples. Alternatively, admixture between the Atlantic and the Senegambia components could have occurred by gene-flow from Senegambia at different times after the initial settlement of the Islands and before their admixture with Europeans. However, the sub-Saharan gene-flow into North Africa is high and has been continuous through time, which makes it difficult to discern whether the Senegambia component was already present in North Africa before the first colonization of the Islands or whether it arrived later on. Moreover, the initial colonization of the Islands was very recent, making it difficult to ascertain how much of the North African component may be

attributable to the initial settlers versus potential gene-flow from North Africa after the initial colonization. Future studies including ancient DNA from North Africa could help resolve these issues.

Both the dates and the origin of the gene-flow from the North African Mediterranean coast suggest a genetic impact of the Arab expansion in the Iberian Peninsula. The Northwest of the Iberian Peninsula shows our oldest estimated date of North African admixture and is consistent with a single pulse of admixture around the time of the early arrival and conquest of Iberia by the Arabs. In contrast, our results suggest that the South of the Iberian Peninsula experienced more recent admixture and perhaps continuous gene-flow. In this case, the admixture is dated to the last periods of the Arab rule in the Peninsula in the second half of the 14th century. In 1212 , ~~when~~ the Christian Kingdoms became allies in the Battle of *Navas de Tolosa* and conquered all Southern territories except the Nasrid Kingdom of Granada, which was conquered at the end of the 15th century. The inferred continuous gene-flow suggests that contact between the Arab and Southern Iberian populations was not limited to that time period, and the estimated dates represent an upper bound on centuries of admixture (figures 4, S5 and S6). Collectively, we can identify at least two different gene-flow events in the Iberian Peninsula for which the inferred dates correlate with Arab rule in the territory: an early concentrated event in the Northwest of the Peninsula, and a continuous and more recent event in the South. Moreover, the North African populations that settled in the Peninsula during the Arab conquest may have had different origins (both in time and in geography), which could be indicative of different migration waves (table S3).

In three of the four minor genetic clusters identified for the Iberian Peninsula (Iberian_Peninsula1, Iberian_Peninsula2 and Iberian_South), three-way admixture was detected between European-like (mainly Iberian), North African-like, and Basque-like sources. Alternatively, in the case of the other minor cluster, Iberian_NorthWest, only two sources of admixture (North African-like and Iberian-like) were detected. This is in agreement with different admixture events occurring at different moments and in which different populations were involved. The fact that in the Northwest of Iberia the admixture does not involved a Basque-like component, while it participated in the admixture events detected in the rest of the Iberian Peninsula, suggests different Iberian populations participated in geographically separated admixture events. This may reflect different waves of the Christian Kingdoms expansion.

The genome-wide study of Fiorito et al., (2016) performed admixture analyses in a large-scale Italian dataset, and highlighted more complex events of admixture than the one described herein in Tuscany. Specifically, they described continuous gene-flow from different sources since 3,000 ya, which could be the result of their more geographically diverse sample set relative to our geographically localized sample of Tuscany. Perhaps because of this, we infer only a single pulse of admixture which coincides with the movement of people during the fall of the Roman Empire, which was just one of the multiple events detected by Fiorito and colleagues (2016). Nonetheless,

our focus on North African populations has allowed us to propose a more precise origin for the North African gene-flow into Tuscany, with our best surrogate group being comprised of present-day people living on the Mediterranean shores of North Africa.

Our study highlights the importance of including an extensive and diverse North African dataset in genetic studies. North Africa is a very heterogeneous region, with ample sociological, historical, and genetic diversity. Our use of an extensive dataset and the use of population clusters based on genetic homogeneity allowed us to detect and describe events of admixture with more precision than previous studies investigation the influence of North African gene-flow into surrounding regions. Recent methods based on haplotype information, such as those presented here, will illuminate the finer structure and genetic history of Iberian populations, particularly as sampling increases both in terms of numbers and geographic regions encompassed (Bycroft et al., 2018). In the case of the Canary Islands, ancient DNA studies might also help to better understand the origin of the first settlers of the islands and identify its influence in modern populations (Fregel et al., 2019).

Competing interests

The authors declare no competing interest.

Author's contributions

LRA participated in data analysis, design of the study and manuscript writing. DC and GH contributed with manuscript writing and design of the study. All authors gave final approval for publication.

Acknowledgments

We thank Francesc Calafell and Alex Mas-Sandoval for helpful discussion. We would like to thank all participants for collaborating in the present study.

Funding

This work was supported by the Spanish MINECO grants CGL2013-44351-P and CGL2016-75389-P (MINECO/FEDER, UE) and the "María de Maeztu" Program for Units of Excellence in R&D (MDM-2014-0370); and the Generalitat de Catalunya grant 2014SGR866. GH is supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (Grant Number 098386/Z/12/Z) and supported by the National Institute for Health Research University College London Hospitals Biomedical Research Centre.

References

Achilli, A., Olivieri, A., Pala, M., Metspalu, E., Fornarino, S., Battaglia, V., Accetturo, M., Kutuev, I., Khusnutdinova, E., Pennarun, E., et al. (2007). Mitochondrial DNA Variation of Modern Tuscans Supports the Near Eastern Origin of Etruscans. Am. J. Hum. Genet. *80*, 759–768.

Adams, S.M., Bosch, E., Balaresque, P.L., Ballereau, S.J., Lee, A.C., Arroyo, E., López-Parra, A.M., Aler, M., Grifo, M.S.G., Brion, M., et al. (2008). The Genetic Legacy of Religious Diversity and Intolerance: Paternal Lineages of Christians, Jews, and Muslims in the Iberian Peninsula. Am. J. Hum. Genet. *83*, 725–736.

Arauna, L.R., Mendoza-Revilla, J., Mas-Sandoval, A., Izaabel, H., Bekada, A., Benhamamouch, S., Fadhlaoui-Zid, K., Zalloua, P., Hellenthal, G., and Comas, D. (2017). Recent historical migrations have shaped the gene pool of Arabs and Berbers in North Africa. Mol. Biol. Evol. *34*, 318–329.

Belle, E.M.S., Ramakrishnan, U., Mountain, J.L., and Barbujani, G. (2006). Serial coalescent simulations suggest a weak genealogical relationship between Etruscans and modern Tuscans. Proc. Natl. Acad. Sci. *103*, 8012–8017.

Boattini, A., Martinez-Cruz, B., Sarno, S., Harmant, C., Useli, A., Sanz, P., Yang-Yao, D., Manry, J., Ciani, G., Luiselli, D., et al. (2013). Uniparental Markers in Italy Reveal a Sex-Biased Genetic Structure and Different Historical Strata. PLoS One *8*, e65441.

Bosch, E., Calafell, F., Comas, D., Oefner, P.J., Underhill, P. a, and Bertranpetit, J. (2001). High-resolution analysis of human Y-chromosome variation shows a sharp discontinuity and limited gene flow between northwestern Africa and the Iberian Peninsula. Am. J. Hum. Genet. *68*, 1019–1029.

Botigué, L.R., Henn, B.M., Gravel, S., Maples, B.K., Gignoux, C.R., Corona, E., Atzmon, G., Burns, E., Ostrer, H., Flores, C., et al. (2013). Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. Proc. Natl. Acad. Sci. *110*, 11791–11796.

Brett, M., and Fentress, E. (1997). The Empire and the Other: Romans and Berbers. In The Berbers, (Library of Congress Cataloging-in-Publication Data), pp. 50–80.

Busby, G.B.J., Hellenthal, G., Montinaro, F., Tofanelli, S., Bulayeva, K., Rudan, I., Zemunik, T., Hayward, C., Toncheva, D., Karachanak-Yankova, S., et al. (2015). The Role of Recent Admixture in Forming the Contemporary West Eurasian Genomic Landscape. Curr. Biol. *25*, 2518–2526.

Bycroft, C., Fernández-Rozadilla, C., Ruiz-Ponte, C., Quintela-García, I., Carracedo, Á., Donnelly, P., and Myers, S. (2018). Patterns of genetic differentiation and the footprints of historical migrations in the Iberian Peninsula. BioRxiv 250191.

Camps, G. (1995). Les Berbères : mémoire et identité (Errance).

Camps, G., and Vela i Aulesa, C. (1998). Los bereberes: de la orilla del Mediterráneo al límite meridional del Sáhara (Icaria).

425 Capelli, C., Onofri, V., Brisighelli, F., Boschi, I., Scarnicci, F., Masullo, M., Ferri, G., Tofanelli, S.,
426 Tagliabracci, A., Gusmao, L., et al. (2009). Moors and Saracens in Europe: estimating the
427 medieval North African male legacy in southern Europe. Eur. J. Hum. Genet. *17*, 848–852.

428 Casas, M.J., Hagelberg, E., Fregel, R., Larruga, J.M., and González, A.M. (2006). Human
429 mitochondrial DNA diversity in an archaeological site in al-Andalus: Genetic impact of
430 migrations from North Africa in Medieval Spain. Am. J. Phys. Anthropol. *131*, 539–551.

431 Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015).
432 Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*,
433 7.

434 Currat, M., Poloni, E.S., and Sanchez-Mazas, A. (2010). Human genetic differentiation across
435 the Strait of Gibraltar. BMC Evol Biol *10*, 237.

436 Fiorito, G., Di Gaetano, C., Guarrera, S., Rosa, F., Feldman, M.W., Piazza, A., and Matullo, G.
437 (2016). The Italian genome reflects the history of Europe and the Mediterranean basin. Eur. J.
438 Hum. Genet. *24*, 1056–1062.

439 Fregel, R., Gomes, V., Gusmão, L., González, A.M., Cabrera, V.M., Amorim, A., and Larruga, J.M.
440 (2009a). Demographic history of Canary Islands male gene-pool: replacement of native
441 lineages by European. BMC Evol. Biol. *9*, 181.

442 Fregel, R., Pestano, J., Arnay, M., Cabrera, V.M., Larruga, J.M., and González, A.M. (2009b). The
443 maternal aborigine colonization of La Palma (Canary Islands). Eur. J. Hum. Genet. *17*, 1314–
444 1324.

445 Fregel, R., Cabrera, V.M., Larruga, J.M., Hernández, J.C., Gámez, A., Pestano, J.J., Arnay, M.,
446 and González, A.M. (2015). Isolation and prominent aboriginal maternal legacy in the
447 present-day population of La Gomera (Canary Islands). Eur. J. Hum. Genet. *23*, 1236–1243.

448 Fregel, R., Ordóñez, A.C., Santana-Cabrera, J., Cabrera, V.M., Velasco-Vázquez, J., Alberto, V.,
449 Moreno-Benítez, M.A., Delgado-Darias, T., Rodríguez-Rodríguez, A., Hernández, J.C., et al.
450 (2019). Mitogenomes illuminate the origin and migration patterns of the indigenous people of
451 the Canary Islands. PLoS One *14*, e0209125.

452 Ghirotto, S., Tassi, F., Fumagalli, E., Colonna, V., Sandionigi, A., Lari, M., Vai, S., Petiti, E., Corti,
453 G., Rizzi, E., et al. (2013). Origins and Evolution of the Etruscans' mtDNA. PLoS One *8*, e55519.

454 González-Fortes, G., Tassi, F., Trucchi, E., Henneberger, K., Paijmans, J.L.A., Díez-del-Molino, D.,
455 Schroeder, H., Susca, R.R., Barroso-Ruíz, C., Bermudez, F.J., et al. (2019). A western route of
456 prehistoric human migration from Africa into the Iberian Peninsula. Proc. R. Soc. B Biol. Sci.
457 *286*, 20182288.

458 Hellenthal, G., Busby, G.B.J., Band, G., Wilson, J.F., Capelli, C., Falush, D., and Myers, S. (2014).
459 A Genetic Atlas of Human Admixture History. Science (80-.). *343*, 747–751.

460 Henn, B.M., Botigué, L.R., Gravel, S., Wang, W., Brisbin, A., Byrnes, J.K., Fadhlaoui-Zid, K.,
461 Zalloua, P. a, Moreno-Estrada, A., Bertranpetit, J., et al. (2012). Genomic Ancestry of North
462 Africans Supports Back-to-Africa Migrations. PLoS Genet. *8*, e1002397.

Laval, G., Patin, E., Barreiro, L.B., and Quintana-Murci, L. (2010). Formulating a Historical and Demographic Model of Recent Human Evolution Based on Resequencing Data from Noncoding Regions. PLoS One *5*, e10284.

Lawson, D.J., Hellenthal, G., Myers, S., and Falush, D. (2012). Inference of Population Structure using Dense Haplotype Data. PLoS Genet. *8*, e1002453.

Leslie, S., Winney, B., Hellenthal, G., Davison, D., Boumertit, A., Day, T., Hutnik, K., Royrvik, E.C., Cunliffe, B., Lawson, D.J., et al. (2015). The fine-scale genetic structure of the British population. Nature *519*, 309–314.

Maca-Meyer, N., González, A.M., Pestano, J., Flores, C., Larruga, J.M., and Cabrera, V.M. (2003). Mitochondrial DNA transit between West Asia and North Africa inferred from U6 phylogeography. BMC Genet. *4*, 15.

Maca-Meyer, N., Arnay, M., Rando, J.C., Flores, C., González, A.M., Cabrera, V.M., and Larruga, J.M. (2004). Ancient mtDNA analysis and the origin of the Guanches. Eur. J. Hum. Genet. *12*, 155–162.

Moorjani, P., Patterson, N., Hirschhorn, J.N., Keinan, A., Hao, L., Atzmon, G., Burns, E., Ostrer, H., Price, A.L., and Reich, D. (2011). The history of African gene flow into Southern Europeans, Levantines, and Jews. PLoS Genet. *7*, e1001373.

Navarro Mederos, J.F. (1997). Arqueología de las Islas Canarias. Espac. Tiempo y Forma, Ser. I, Prehist. y Arqueol. *10*, 447–478.

Naylor, P.C. (2009). North Africa: a history from antiquity to the present (University of Texas Press).

O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., Traglia, M., Huang, J., Huffman, J.E., Rudan, I., et al. (2014). A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. PLoS Genet. *10*, e1004234.

Olalde, I., Mallick, S., Patterson, N., Rohland, N., Villalba-Mouco, V., Silva, M., Dulias, K., Edwards, C.J., Gandini, F., Pala, M., et al. (2019). The genomic history of the Iberian Peninsula over the past 8000 years. Science (80-.). *363*, 1230–1234.

Ordóñez, A.C., Fregel, R., Trujillo-Mederos, A., Hervella, M., de-la-Rúa, C., and Arnay-de-la-Rosa, M. (2017). Genetic studies on the prehispanic population buried in Punta Azul cave (El Hierro, Canary Islands). J. Archaeol. Sci. *78*, 20–28.

Peña, P.A. (2013). Consideraciones en relación con la colonización protohistórica de las Islas Canarias. Anu. Estud. Atlánticos 519–562.

Plaza, S., Calafell, F., Helal, A., Bouzerna, N., Lefranc, G., Bertranpetit, J., and Comas, D. (2003). Joining the pillars of hercules: mtDNA sequences show multidirectional gene flow in the Western Mediterranean. Ann. Hum. Genet. *67*, 312–328.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

Regueiro, M., Garcia-Bertrand, R., Fadhlaoui-Zid, K., Álvarez, J., and Herrera, R.J. (2015). From Arabia to Iberia: A Y chromosome perspective. Gene *564*, 141–152.

Reverón, B.B. (1944). Las canarias y la conquista Franco-Normanda: Juan de Bethencourt.

Rodríguez-Martin, C., and Martín-Oval, M. (2009). Guanches, una historia bioantropológica (Museo Arqueológico de Tenerife).

Rodríguez-Varela, R., Günther, T., Krzewińska, M., Storå, J., Gillingwater, T.H., MacCallum, M., Arsuaga, J.L., Dobney, K., Valdiosera, C., Jakobsson, M., et al. (2017). Genomic Analyses of Pre-European Conquest Human Remains from the Canary Islands Reveal Close Affinity to Modern North Africans. Curr. Biol. *27*, 3396–3402.e5.

Secher, B., Fregel, R., Larruga, J.M., Cabrera, V.M., Endicott, P., Pestano, J.J., and González, A.M. (2014). The history of the North African mitochondrial DNA haplogroup U6 gene flow into the African, Eurasian and American continents. BMC Evol. Biol. *14*, 109.

Solé-Morata, N., García-Fernández, C., Urasin, V., Bekada, A., Fadhlaoui-Zid, K., Zalloua, P., Comas, D., and Calafell, F. (2017). Whole Y-chromosome sequences reveal an extremely recent origin of the most common North African paternal lineage E-M183 (M81). Sci. Rep. *7*, 15941.

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.

The International HapMap Consortium (2003). The International HapMap Project. Nature *426*, 789–796.

Valdiosera, C., Günther, T., Vera-Rodríguez, J.C., Ureña, I., Iriarte, E., Rodríguez-Varela, R., Simões, L.G., Martínez-Sánchez, R.M., Svensson, E.M., Malmström, H., et al. (2018). Four millennia of Iberian biomolecular prehistory illustrate the impact of prehistoric migrations at the far end of Eurasia. Proc. Natl. Acad. Sci. *115*, 3428–3433.

Figure captions

Figure 1. Principal Component Analysis (PCA). The larger PCA shows all the samples included in the study, whereas the inner PCA only includes the European samples.

Figure 2. FineStructure clustering shown as a dendrogram and its correspondence in a map. The filled rectangles are the North African samples, and the proportion of individuals from each of the clusters in each geographical sampled population is shown in pie-charts. Clusters containing European and sub-Saharan African individuals are denoted by non-filled rectangles colored blue and yellow, respectively, and are labeled primarily according to geography.

Figure 3. Globetrotter admixture results for the three geographical regions analyzed (Tuscany, Iberia, and the Canary Islands). The mean admixture date and confidence intervals for each admixture event are shown above the graphs. The geographical locations of surrogates that contribute more than 2.5% are colored in the maps, with circle sizes showing the proportion of contribution. Colored areas boundaries are defined by the genetic clusters' geographic distribution. Each different shade of grey corresponds to a different admixing source group, with the surrogates

538 representing that source group linked via a continuous or dashed line. The pie in each graph shows
539 the proportion inferred from each admixing source for the given target population (Tuscany, Iberia,
540 or the Canary Islands, respectively).

541 **Figure 4.** Density plot for the admixture dates estimates after 100 bootstrap iterations of
542 Globetrotter. The x-axis shows the date of admixture in years. On the top left the FineStructure
543 dendrogram and the geographical distribution of minor clusters for the Iberian samples are shown,
544 with each pie showing the proportion of individuals from that sampling location that were assigned
545 to each of the four minor clusters (colors). The size of each circle corresponds to the number of
546 sampled individuals. One cluster was formed by only one individual and therefore is not
547 considered.

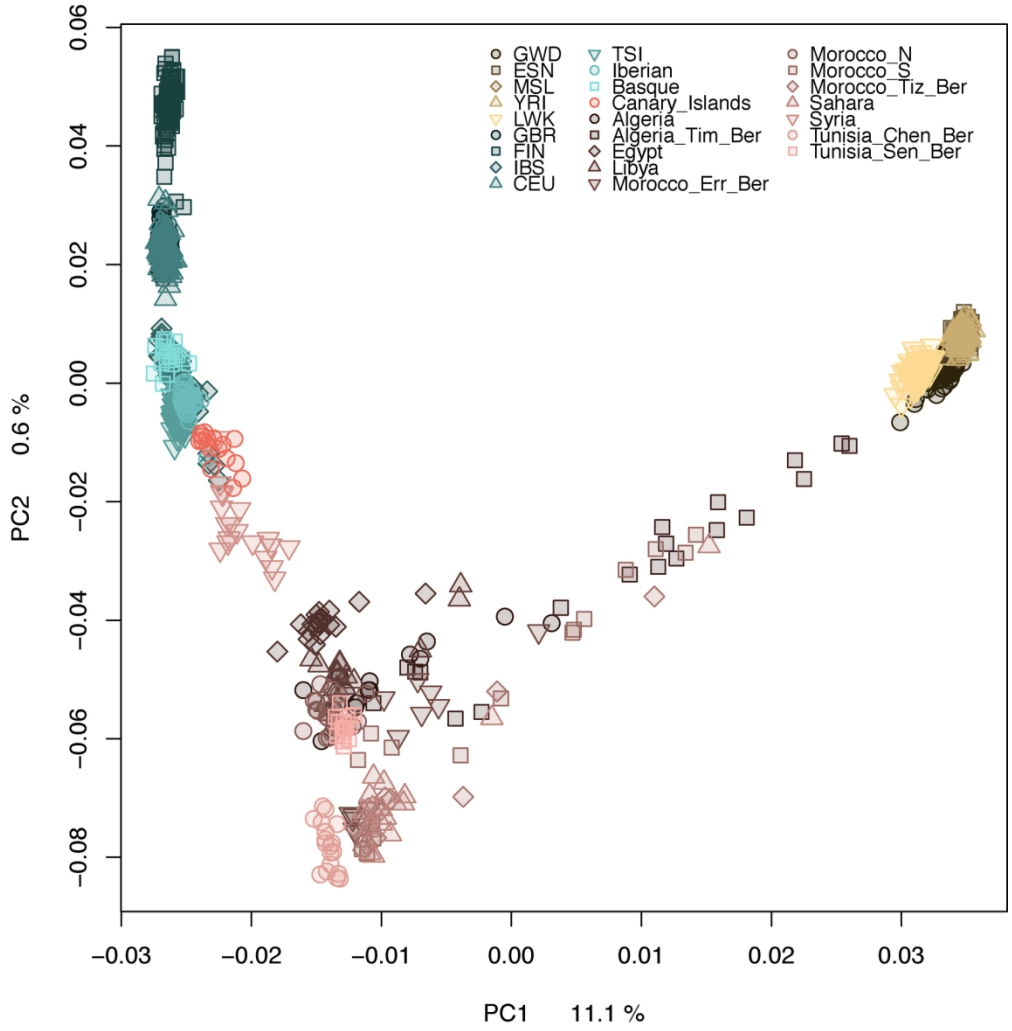

Figure 1. Principal Component Analysis (PCA). The larger PCA shows all the samples included in the study, whereas the inner PCA only includes the European samples.

153x157mm (300 x 300 DPI)

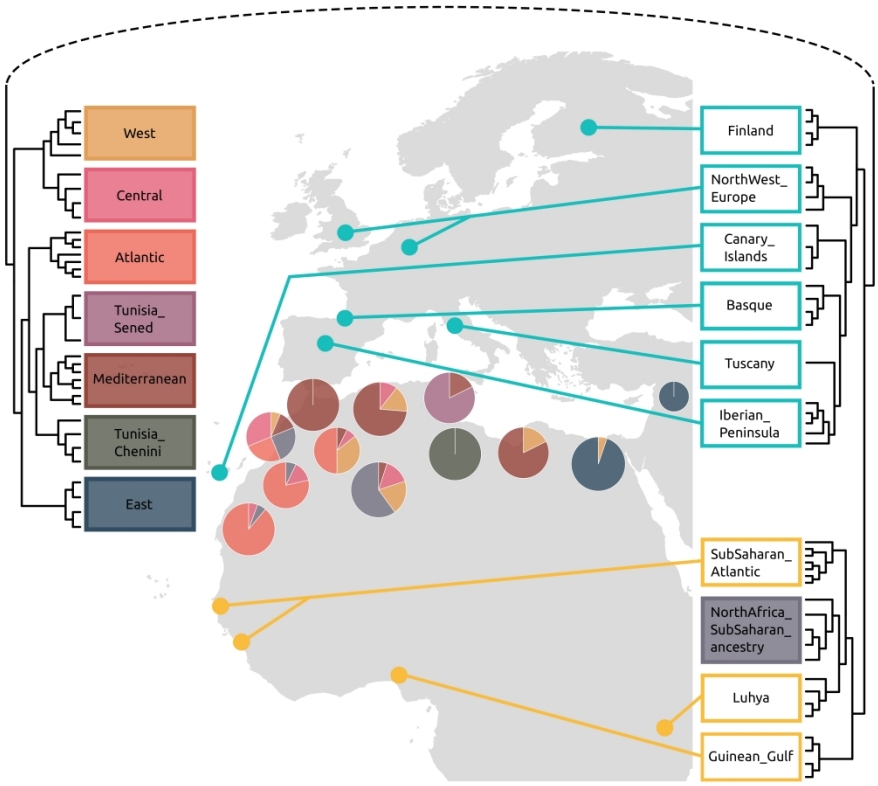

Figure 2. FineStructure clustering shown as a dendrogram and its correspondence in a map. The filled rectangles are the North African samples, and the proportion of individuals from each of the clusters in each geographical sampled population is shown in pie-charts. Clusters containing European and sub-Saharan African individuals are denoted by non-filled rectangles colored blue and yellow, respectively, and are labeled primarily according to geography.

496x512mm (299 x 299 DPI)

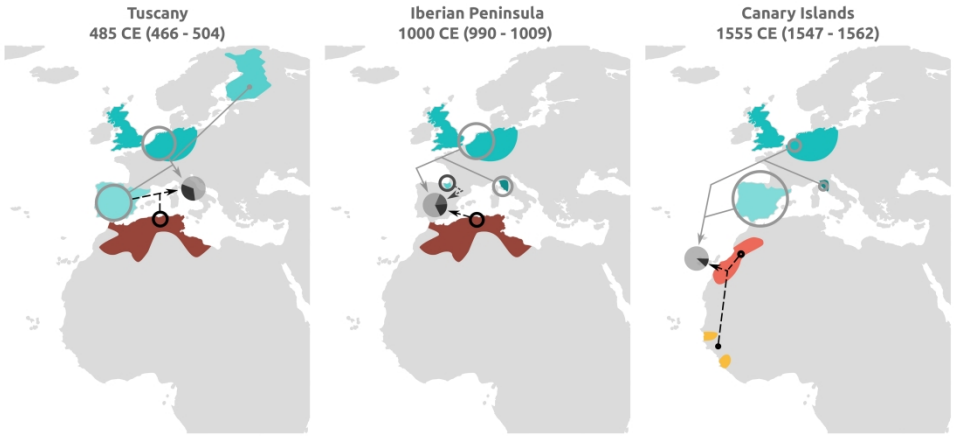

Figure 3. Globetrotter admixture results for the three geographical regions analyzed (Tuscany, Iberia, and the Canary Islands). The mean admixture date and confidence intervals for each admixture event are shown above the graphs. The geographical locations of surrogates that contribute more than 2.5% are colored in the maps, with circle sizes showing the proportion of contribution. Colored areas boundaries are defined by the genetic clusters' geographic distribution. Each different shade of grey corresponds to a different admixing source group, with the surrogates representing that source group linked via a continuous or dashed line. The pie in each graph shows the proportion inferred from each admixing source for the given target population (Tuscany, Iberia, or the Canary Islands, respectively).

199x100mm (600 x 600 DPI)

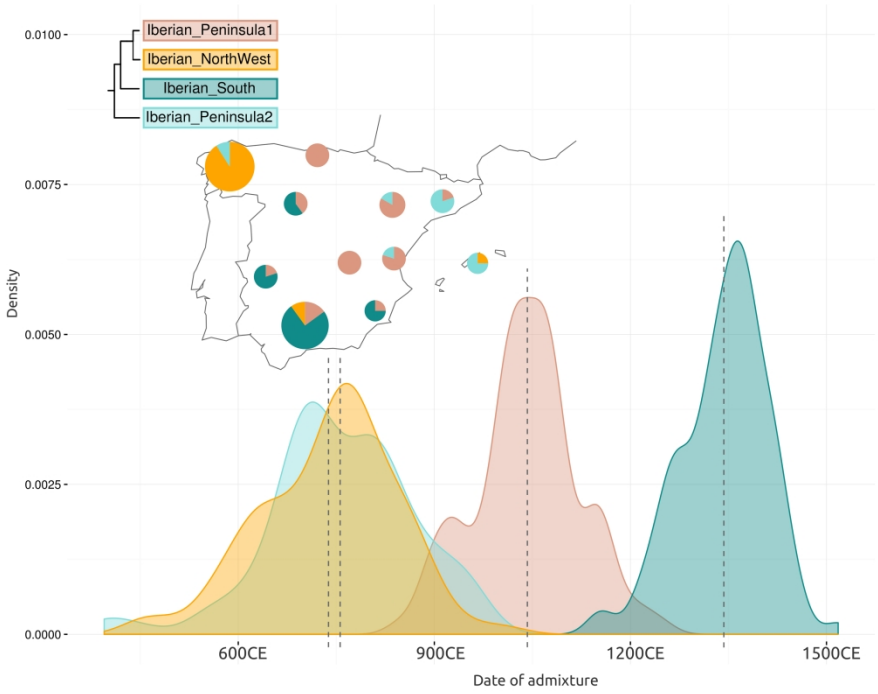

Figure 4. Density plot for the admixture dates estimates after 100 bootstrap iterations of Globetrotter. The x-axis shows the date of admixture in years. On the top left the FineStructure dendrogram and the geographical distribution of minor clusters for the Iberian samples are shown, with each pie showing the proportion of individuals from that sampling location that were assigned to each of the four minor clusters (colors). The size of each circle corresponds to the number of sampled individuals. One cluster was formed by only one individual and therefore is not considered.

296x209mm (300 x 300 DPI)