Genome-wide analysis of Corsican population reveals a close affinity with Northern and Central Italy

Erika Tamm, Julie Di Cristofaro, Stéphane Mazières, Erwan Pennarun, Alena Kushniarevich, Alessandro Raveane, Ornella Semino, Jacques Chiaroni, Luisa Pereira, Mait Metspalu, et al.

SCIENTIFIC REPORTS
nature research

OPEN

Genome-wide analysis of Corsican population reveals a close affinity with Northern and Central Italy

Received: 24 April 2019
Accepted: 31 August 2019
Published online: 19 September 2019

Erika Tamm[1], Julie Di Cristofaro[2,3], Stéphane Mazières[2], Erwan Pennarun[1], Alena Kushniarevich[1,4], Alessandro Raveane[5], Ornella Semino[5], Jacques Chiaroni[2,3], Luisa Pereira[6,7], Mait Metspalu[1] & Francesco Montinaro[1,8]

Despite being the fourth largest island in the Mediterranean basin, the genetic variation of Corsica has not been explored as exhaustively as Sardinia, which is situated only 11 km South. However, it is likely that the populations of the two islands shared, at least in part, similar demographic histories. Moreover, the relative small size of the Corsica may have caused genetic isolation, which, in turn, might be relevant under medical and translational perspectives. Here we analysed genome wide data of 16 Corsicans, and integrated with newly (33 individuals) and previously generated samples from West Eurasia and North Africa. Allele frequency, haplotype-based, and ancient genome analyses suggest that although Sardinia and Corsica may have witnessed similar isolation and migration events, the latter is genetically closer to populations from continental Europe, such as Northern and Central Italians.

Corsica, located south of the shore of Côte d'Azur (France), and west of Tuscany (Italy), is separated from Sardinia to its south by the Strait of Bonifacio. It is the fourth largest Mediterranean island (8,680 km^2) and unlike most of them, its relief is very mountainous, with a mountain range bisecting the island. Nowadays, approximately ~339,000 people inhabit the island[1].

The understanding of the peopling of Corsica has remained incomprehensive. From a geological perspective, during the last glaciation, Corsica and Sardinia formed a single landmass and its distance to Italy was reduced, possibly increasing connections with mainland[2,3]. Furthermore, archaeological records suggest that the Southern part of the Sardinia-Corsica palaeo-island, characterised by milder climate and less geographical asperities, was settled at a first stage, with the area corresponding to modern day Corsica, characterised by harsher conditions, being colonised later. However, the acidity of deposits and submersion led to a scarce persistence of anthropological and archaeological remains, preventing the extensive characterization of its peopling dynamics. There is no clear evidence of human traces from the end of the Pleistocene and the beginning of the Holocene. The oldest human remains found so far in Corsica are from Campu Stefano and are dated 8,940 ^{14}C year BP (10,216–9,920 cal. BP, 95.4% range)[4]. Archaeological and genetic data suggest episodic and discontinuous settlements during Mesolithic and transitional phases between Mesolithic and Neolithic periods[5,6]. The permanent human presence in Corsica is attested in the Neolithic period since the 6th millennium BC and possible interactions with mainland and other islands are suggested by the wider appearance of non-local lithic resources and similar ceramic traditions over a larger Western Mediterranean region[7].

In the last three millennia, the Corsican population witnessed several dramatic demographic changes due to conquests, epidemics outbreaks and economic crises. The Greeks established the city of Alalia (today Aleria) in 565 BC, which is also the first mention of Corsica in historical records. Subsequently, it witnessed numerous intrusions and conquests by different populations. Carthaginians and Etruscans dominated the island until the Roman occupation in the third century BC. Successive invasions by the Vandals, Ostrogoths and Saracens, took

[1]Institute of Genomics, University of Tartu, Tartu, Estonia. [2]Aix Marseille Univ, CNRS, EFS, ADES, Marseille, France. [3]Etablissement Français du Sang PACA Corse, Biologie des Groupes Sanguins, Marseille, France. [4]Institute of Genetics and Cytology, National Academy of Sciences of Belarus, Minsk, 220072, Belarus. [5]Dipartimento di Biologia e Biotecnologie "L. Spallanzani" Università di Pavia, Via Ferrata 9, 27100, Pavia, Italy. [6]i3S - Instituto de Investigação e Inovação em Saúde, Universidade do Porto, 4200-135, Porto, Portugal. [7]Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), 4200-135, Porto, Portugal. [8]Department of Zoology, University of Oxford, Oxford, UK. Correspondence and requests for materials should be addressed to E.T. (email: erika.tamm@gmail.com) or F.M. (email: francesco.montinaro@gmail.com)

place until the beginning of the Byzantine influence in the 6th century, followed by Roman Papacy from the 8th century onwards. From the end of the 11th century Corsica was under the government of Pisa. Then, with the intermittent period dominated by the kings of Aragon, Corsica was under the rule of Genoa from the 13th to the 18th century, when it was incorporated into France in 1768[8].

Previous genetic surveys on Corsican population have shown regional differentiation[9–12], such as a marked North-South differentiation, confirmed by surname and linguistic studies[13], and mirroring the geographic features of the island.

Despite their geographic proximity, different investigations based on unilinear and autosomal markers have produced contradicting results in characterizing the relationship between Sardinia and Corsica. Some studies have found genetic affinity between the two islands[9,10,14–16], while conversely, some have suggested genetic distinction[11,12,17,18]. Furthermore, the extent of Corsica's genetic relationships with mainland populations is controversial. Close genetic connections have been observed between Corsica and continental Mediterranean populations[11,18–20], whereas other studies have reported limited genetic influences from mainland[10,12]. Furthermore, it has been previously suggested that different regions in Corsica have received genetic influences from distinct sources, showing diverse affinities with surrounding populations[9,12].

Regardless of its genetic affinity, it is unclear to what extent the Corsican population shows a signature of demographic bottleneck or population decrease, as expected in a long term isolation scenario. Previous studies have demonstrated that isolated populations are valuable resources in genome-wide association studies. In geographically and/or culturally isolated conditions evolutionary forces and population dynamics can lead to the high level of homozygosity, reduced genetic diversity and increase in peculiar allele frequencies, which makes easier to trace genetic variants affecting medical or phenotypic traits[21,22]. Sardinia is a well-known example (see for instance[23] and references therein). Preliminary studies have suggested Corsican potential in this context: an analysis on chromosome X microsatellite markers demonstrated an overall LD decrease in the innermost part of Corsica[24] and some medical and association studies have been conducted on Corsican population[25–27].

Despite its anthropological and epidemiological relevance, a genome-wide characterisation of Corsican population is not available so far. In order to evaluate the within-population genetic variation, and the affinity with other modern and ancient European populations, here we genotyped 16 Corsican samples collected from different locations of the island (Supplementary Fig. S1), together with 33 new samples from Portugal, France and Italy. We combined this newly generated data with 892 modern and 222 ancient Eurasian and African genomes from previously published sources (Supplementary Table S1) and applied a series of allele frequency and haplotype-based methods to unravel a multi-faceted genome-wide description of Corsica.

Results

Population structure. To explore whether Corsicans, as insular population, display characteristic traces of isolation and endogamy in their genomes, we assessed Runs of Homozygosity (RoH). RoH are uninterrupted segments of homozygous genotypes present in individuals. The number and extension of RoH stretches are correlated to the level of identity by descent and are largely affected by population's demographic history[28,29]. Corsicans show an excess of RoHs indicated by the high median values of total number and total length of RoHs, next to Sardinians and French Basques (Fig. 1). This homozygosity pattern is characteristic of isolated populations with relatively low effective population size (Ne) and high degree of endogamy[28–30]. Similar to Sardinians and Basques, Corsicans show high variability of long RoHs, possibly indicating recent ongoing admixture.

To position Corsicans in the context of their geographic neighbours we used an unsupervised clustering approach[31] implemented in the program ADMIXTURE[32]. At the level of lowest cross-validation index ($K = 6$), Corsicans and geographically related populations are composed mainly by three ancestry components - "*Sardinian*", "*Northern and Eastern Europe*", "*Caucasus and Middle East*" (dark blue, light blue, and lime green respectively; Fig. 2, Supplementary Figs S2, S3). Corsicans are most similar to North-Central Italian populations (Piedmont, Lombardy, Tuscany), displaying a slightly larger proportion of a modal component in Sardinians. This similarity in ADMIXTURE profiles remains throughout higher levels of K values, even when further components appear (Fig. 2, Supplementary Fig. S2). The closest geographic neighbours, Sardinians, and the other known isolate, French Basques, virtually lack Caucasus/Middle Eastern (lime green) ancestry. At $K = 8$ and $K = 11$ two components almost fixed in Sardinians (dark blue) and in French Basques (middle blue), respectively, emerge.

Analysis of average population pairwise F_{ST} distance confirms ADMIXTURE results showing shortest genetic distances between geographically close mainland populations and revealing larger distances with Sardinians and French Basques (Supplementary Table S2).

To evaluate the genetic affinity between Corsica and neighbouring populations we carried out f3-statistics analysis. Outgroup f3 statistics measures the amount of shared genetic drift between two populations from an outgroup[33]. The results showed high affinity of Corsicans with Sardinians and French Basques, followed, as for ADMIXTURE and F_{ST} analyses, by Northern Italians and a series of mainland European populations (Supplementary Fig. S4A). The qualitative discrepancy between F_{ST} and f3 statistics is probably due to the fact that the former is more affected by strong demographic changes (e.g. bottleneck) than the latter. However, the large standard errors associated to the f3 statistics suggest that the differences in affinity of Corsica with Sardinian/Basque or mainland populations are negligible or non-significant. The tests of admixture based on the f3 statistics revealed an overall complexity in the admixture history of Corsica. Out of 2,652 tests performed, 18 were statistically significant ($|Z| > 3$), although only 8 remained significant using a more conservative threshold ($|Z| > 4$) (Supplementary Fig. S4B). Regardless of the threshold considered, most of the significant tests described Corsica as a combination of Sardinian and Caucasus/Northern European contributions. Moreover, two tests including populations from North African/Arabia and one from Northern Europe were statistically significant ($|Z| > 4$).

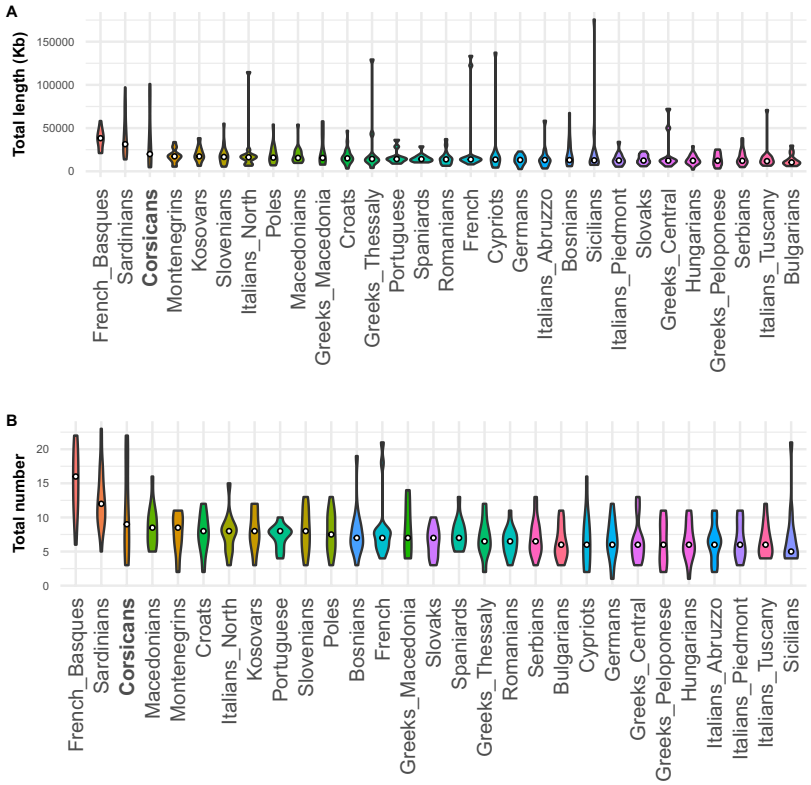

Figure 1. Violin plots of Runs of Homozygosity (RoH) estimations for Corsicans and comparative European populations. (**A**) Total length in kilobases (kb) of genome in RoH. (**B**) Total number of RoH segments. The violin plots show the distribution of RoH fragments with width indicating frequencies and the median as a white circle. Populations are ordered from left to right according to their median value.

Haplotype-based analyses. We inferred the fine-scale population structure harnessing the haplotype-sharing patterns among individuals. In detail, using ChromoPainter we reconstructed each analysed individual *j* as a mosaic of genomic fragments inherited by *n* donor samples[34], both in terms of total number (chunkcounts) or length of fragments (in cM, chunklengths). The resulting *j x n* coancestry chunkcount matrix was then employed to identify and characterise homogenous groups of individuals, in the form of a dendrogram. fineSTRUCTURE clustering results together with the ChromoPainter coancestry chunkcount matrix were subsequently used to identify, date and describe admixture events with GLOBETROTTER[35].

The complete fineSTRUCTURE dendrogram is represented in Supplementary Fig. S5, together with detailed chunkcount coancestry matrix (Supplementary Fig. S6) and pairwise coincidence matrix (Supplementary Fig. S7). For simplicity, we assigned a label to the inferred clusters summarising their composition, reported in Supplementary Table S3. All together, we identified 85 homogeneous groups, with a broad correlation with geographic origin of samples. Specifically, West Eurasian subjects grouped into 38 clusters, for which six macrogroups may be identified (Fig. 3A). Samples from the Levant area (Druze, Syrians_Lebanese and Lebanese clusters) grouped close to individuals from Cyprus and Armenia in the Caucasus (purple in Fig. 3). Other samples from Caucasus (light blue in Fig. 3) fell into a macrogroup that includes eight different clusters (Lezgins, Azeris, Turks, Georgians, Balkars_Adygei, Balkars, Adygei1, Adygei2). Samples from North and West Europe distributed into six clusters (FrenchBasques, French, Iberia2, Orcadians1, Orcadians2, Swedes), which grouped together and were related to a group of seven clusters comprising mostly individuals from the Balkans (CentralEastEurope, Balkan1, Kosovars, Montenegrins, Balkan2, Greeks, Bosnian_Greek). These two macrogroups were closest to the Central and East European group, which was composed by four clusters (Baltics, Germans, Estonians, EasternEurope). Finally, the only clusters solely represented by Corsicans (Supplementary Fig. S5) grouped with all the Italian samples and one Iberian cluster (Iberia1), establishing a Southern European macrogroup.

At a finer inspection, out of the 16 Corsican samples analysed, 12 formed a population specific cluster (Corsicans2) in the Southern European clade, related to North and Central Italian group. Three samples fell into three different clusters composed by North and Central Italians (NorthCentralItaly), French (French) and Iberians (Iberia2), respectively (Fig. 3, Supplementary Fig. S5), possibly reflecting recent relationship with mainland populations. The remaining Corsican individual formed a separate branch, closest to Portuguese and Spaniards (Iberia1). Sardinians and French Basques group with neighbouring populations: the former with Southern Europeans while the latter with Northern and Western Europeans.

In order to explore the relationships between populations, we performed a Principal Component Analysis (PCA) based on the chunkcount coancestry matrix. The first two PCs explain a large proportion of the total variance (55% and 19% respectively, Fig. 3B), confirming the efficacy in summarizing the genomic information

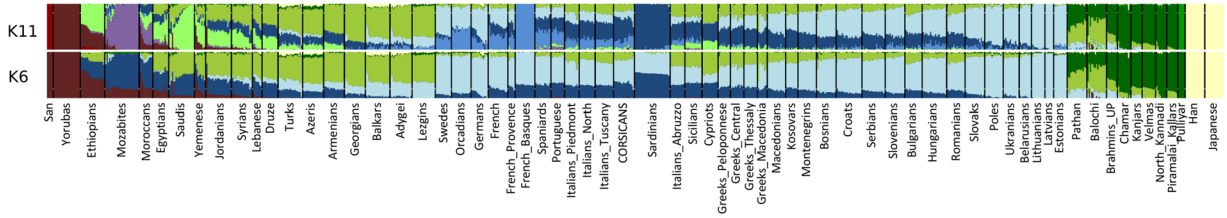

Figure 2. ADMIXTURE plot of individual ancestry proportions at K = 6 and K = 11.

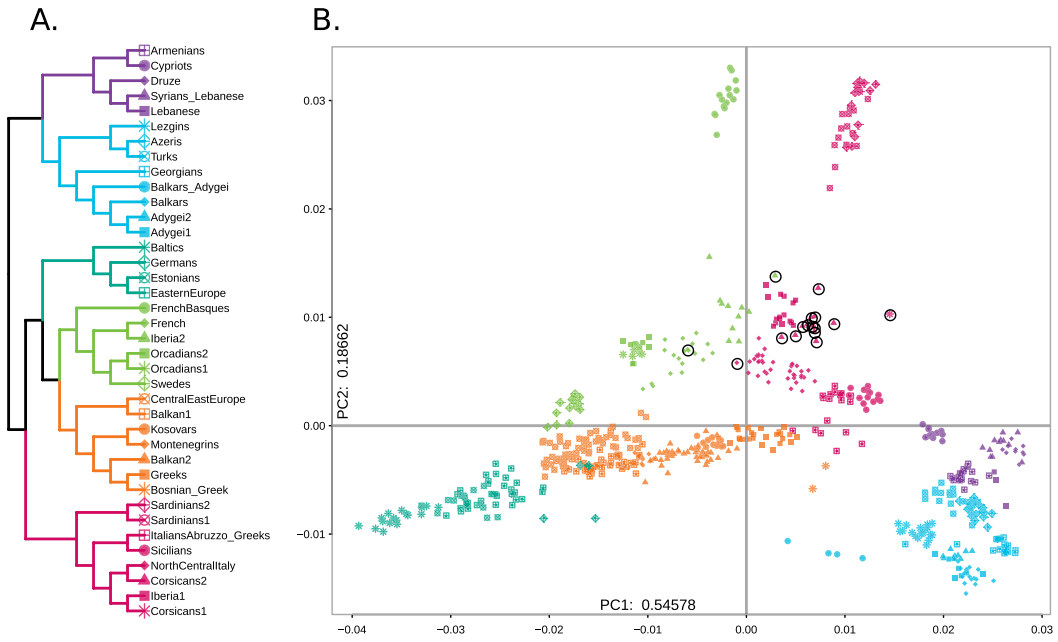

Figure 3. Genetic structure based on haplotype-sharing pattern. (**A**) fineSTRUCTURE tree of Eurasian populations. We applied the ChromoPainter/fineStucture pipeline in order to build a dendrogram showing the relationships among homogenous groups in our dataset. Only the portion of the tree including Western Eurasian and a sub-set of Levantine populations are shown. Colours of the tree branches indicate macroregional grouping discussed in the text. The full fineSTRUCTURE tree is shown in Supplementary Fig. S5. (**B**) Principal Component Analysis based on haplotype sharing. The chunkcount coancestry matrix of Western Eurasian and Levantine samples have been used to perform PCA analysis, and visualize the top two components in a scatterplot. Corsican samples are highlighted by black circles.

of the painting approach. The first principal component separates populations along a North-South axis, placing North-East Europeans on one side (left) and Near East/Caucasus populations to the opposite (right). The second component separates populations along a West-East axis, with Sardinians and French Basques being clear outliers. Most Corsican samples group together close to Italian and Spanish populations; the four samples that did not group with Corsican main cluster, occupy outlier positions also in the PCA.

To elucidate the admixture history of Corsican population, genetic clusters defined by fineSTRUCTURE were used to perform analysis with GLOBETROTTER[35]. We focused on Western Eurasian clusters composed by more than five individuals. As geographically close populations tend to share recent ancestry and distant genetic contacts may be masked, we performed two different GLOBETROTTER analyses, "full" and "non-local", as previously reported[36,37] (Fig. 4, Supplementary Fig. S8 and Table S4). The "full" analysis considers all samples as possible sources, while "non-local" excludes Southern European clusters as donors. In both the analysis, a single admixture involving more than one source was identified for the main Corsican cluster. This admixture involved North and West Europe and Levant/North African sources, and occurred between 37 and 76 generations ago, a time period spanning the fall of the Roman Empire and the invasions by Barbarians and Saracens, when considering a generation time of 30 years[38].

In the "full" analysis, the admixture sources included Sardinians, NorthCentral Italians, Spanish and Sicilians (Supplementary Table S4), possibly suggesting gene-flows from continental Europe, while the Levant/North African contribution inferred in the "non-local" analysis was most probably passed to Corsicans hitchhiking on mainland populations. Similar admixture profiles were observed for two Sardinian clusters, with central time estimates of 59 and 44 generations ago. These results suggest that similar admixture episodes affected both the

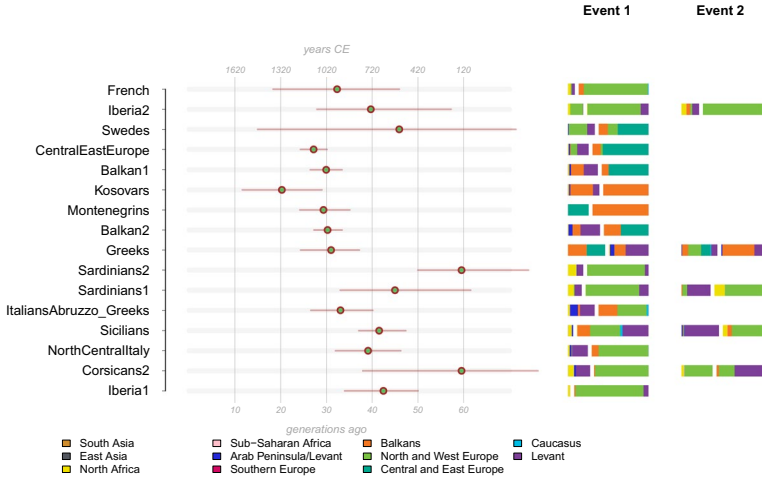

Figure 4. Admixture dates as inferred by GLOBETROTTER in the "non-local" analysis. We fit the painting profile of Western Eurasian populations into expected curves for different admixture models, as implemented in GLOBETROTTER. The estimated dates and sources composition are shown. When one date multiple way result was detected, two events are indicated. For each event the two putative source composition is separated by a white space in the barplot.

Mediterranean islands. The impact of North Africa and Levant is also evident in the remaining Italian and Iberian samples, highlighting the wide impact of the event.

Ancient contribution in Corsica. In order to understand how Corsica and neighbour populations are related to groups that occupied the continent in the last ~10k years, we have performed a Principal Component Analysis projecting ancient individuals onto PCs estimated on modern European allele-frequency variation (Fig. 5A). When the first two PCs are considered, Corsicans are close to European samples from Early, Middle Neolithic and Chalcolithic, together with Balkans Chalcolithic and Neolithic. Compared to other relevant modern European populations, Corsicans show their closeness to Central and Northern Italian rather than Sardinians, although they are scattered towards the latter, suggesting a larger affinity to Neolithic ancestry and Sardinian population than mainland European.

In order to better characterize the "ancestral" composition of Corsican and European populations, we inferred their genetic relationship with a set of ancient individuals using qpAdm[39]. In our analysis, the Corsican samples were characterised by a high ancestry of European Early Neolithic (56%), similarly to Italian, Spanish and Balkan populations (Fig. 5B). In addition, Corsica harboured a relatively high proportion related to Iranian Neolithic (22%), while the contribution of Western and Eastern Hunter Gatherer (WHG, EHG) was smaller, about 11%. According to previous research, a substantial proportion of the EHG and Iranian Neolithic (related to Caucasus Hunter Gatherer, CHG) trace back to Bronze Age movements from the Steppe, although, Iranian Neolithic could have arrived to Western Mediterranean with different migrations[37,39,40]. This seems to be supported also for Corsica when qpAdm analysis including Steppe_EMBA (Early Middle Bronze Age) and Iranian Neolithic (Iran_N) were considered in the same analysis (Anatolia Neolithic: 33%, Steppe_EMBA: 19%, Iran Neolithic: 14%, Europe Middle Neolithic/Chalcolithic: 34%). Compared to Corsicans, French and Spanish samples were characterised by a smaller proportion of Iranian Neolithic (13% and 15%, respectively), and a slightly higher contribution from Western Hunter Gatherers (14% and 17%).

Discussion

Despite being the fourth largest island in the Mediterranean basin, the genetic variation of Corsica has not been explored as exhaustively as Sardinia[41], which is situated only 11 km South. However, it is likely that the two populations have shared at least part of their demographic history, given their geographic proximity and similarities in the archaeological record. In addition, the relatively small size of Corsica may have contributed to create isolation conditions affecting the genetic variability of the autochthonous population[24].

Our analysis revealed that Corsican population shares several genomic features with Sardinia and North-Central Italy, creating a unique blend of genomic ancestries (Fig. 1–3).

The Corsican population shows a relatively high homozygosity characterised by high variance, suggesting that it witnessed both an isolation and migration phase. A similar pattern has been observed for Sardinians (Fig. 1). As isolation and extended homozygosity rates have been shown to be beneficial in gene-mapping and translational studies[22,30], relatively high homozygosity found in the current study confirms Corsican population as a potentially valuable resource for association studies, as also suggested by some previous surveys of autosomal markers[25–27] and chromosome X microsatellite LD[24] variation in Corsican population.

Allele frequency-based assignment algorithm and genetic distance methods showed that Corsica is more closely related to mainland populations from France, Italy, Spain and Greece rather than Sardinia (Fig. 2,

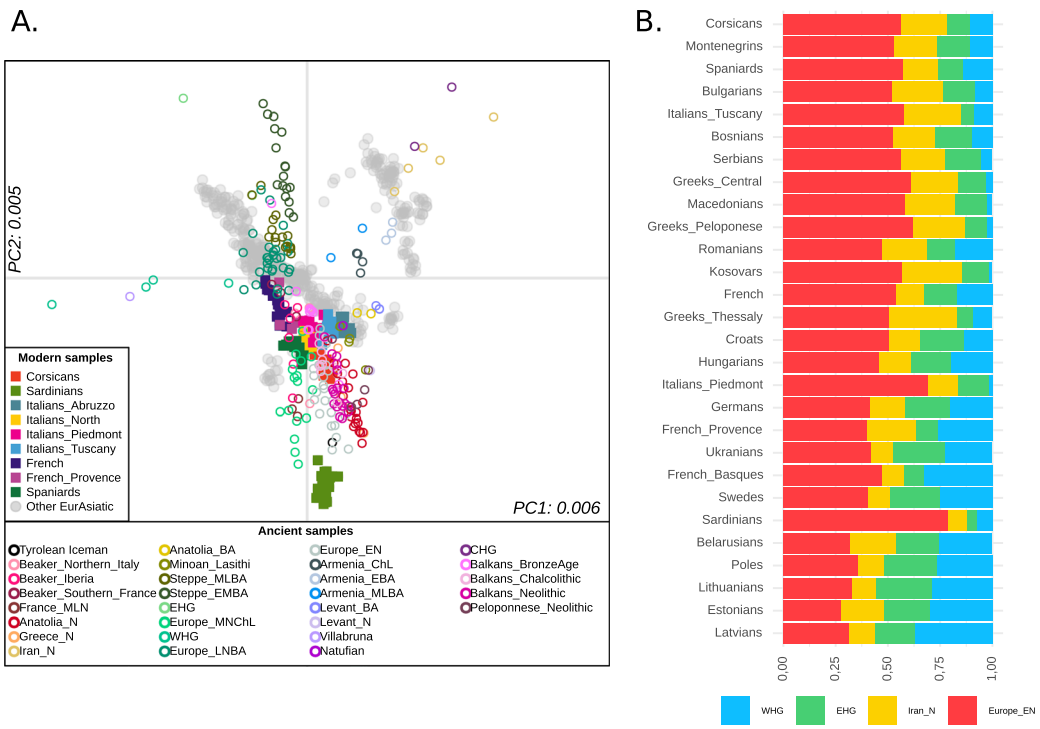

Figure 5. Genetic variability in the context of ancient individuals. (**A**) Principal Component Analysis of ancient individual genotypes projected onto the first two PCs estimated using modern West Eurasian populations. (**B**) Admixture profile of Western Eurasians as inferred by qpAdm. We have reconstructed the admixture profile of all the analysed populations with qpADM, which harnesses a combination of f4 describing the relationship of "target" and "sources" with a set of outgroups. A four-population model including WHG, EHG, European Early Neolithic (EN) and Iran Neolithic is supported in most of the tested populations. Populations are sorted according to Euclidean distances.

Supplementary Fig. S2, Supplementary Table S2). In contrast, the f3 outgroup analysis suggested that Corsica shares a high amount of genetic drift with Sardinia and Basque, followed by North Italy (Supplementary Fig. S4A).

To better characterise the genetic relationship of Corsica with other Mediterranean populations we have harnessed the information embedded in haplotype configurations. The analyses of the haplotype sharing patterns performed using ChromoPainter and fineSTRUCTURE provided substantial evidence of higher affinity between Corsica and Central or Northern Italian populations rather than Sardinian, French and Iberian populations (Fig. 3). In fact, most of the Corsican individuals tested fall in a homogenous cluster closely related to North and Central Italy. On the other hand, a small proportion of individuals are closer to French and Iberians, suggesting the existence of substantial heterogeneity possibly due to recent migration, as confirmed by RoHs analysis (Fig. 1).

When we investigated admixture evidence of Corsican population through f3 analysis, Corsicans could be described as a combination of allele frequencies from Sardinia, Northern Europe, Caucasus, North Africa and Arabian Peninsula (Supplementary Fig. S4B). A similar result was obtained when haplotypic patterns were explored (Fig. 4, Supplementary Fig. S8 and Table S4). Corsicans fit a scenario of admixture involving more than two sources related to Southern European, Western European and Levant Arabic populations which occurred ~60 generations ago, in a similar time frame inferred for Sardinians, although the large confidence intervals associated with the analysis make the overall interpretation challenging. Nevertheless, similar source compositions have been inferred not only for Sardinians, but also for North and Central Italians, suggesting that similar processes may have impacted the Mediterranean basin. In detail, we inferred that admixture events involving Southern European samples occurred about between 37 and 76 generations ago, a time period spanning the fall of the Roman Empire and the invasions from Barbarians and Saracens (Fig. 4). These estimates are in accordance with those inferred in previous investigations for a sub-sample of circum-Mediterranean populations[35,36].

Lastly, we have evaluated the genetic relationship of Corsica and other Mediterranean populations with ancient Eurasian individuals through PCA and qpAdm (Fig. 5). Corsicans show a high affinity to European ancient individuals characterized by high Neolithic ancestry such as European from Early, Middle and Late (Chalcolithic) Neolithic. Furthermore, most of the tested European populations can be modelled according to European Early Neolithic, Iranian Neolithic, Western and Eastern Hunter Gatherer contribution. The "ancient profile" of Corsica is characterised by a high proportion of ancestry related to European Early Neolithic, although lower than the one found in Sardinians (56% vs 79%), and similar to the one inferred for Tuscany and Spain. Interestingly, Corsica has a non-negligible fraction of ancestry related to Iranian Neolithic, which could be independent from the one brought to Europe through the Steppe related migration, as previously suggested[37,42]. In fact, we have found a similar proportion of Steppe Bronze Age (~19%) and Iranian Neolithic (~14%) in Corsicans.

In conclusion, the genetic characterisation of Corsica is consistent with a closer genetic affinity with Northern and Central Italian populations rather than Sardinians, although sharing with the latter a noteworthy proportion of ancestry and similar demographic and isolation processes. The analysis of larger sample sizes from different regions of the island and genetic material from ancient specimen may help to further evaluate the genetic structure in the island and its demographic history.

Materials and Methods

Sampling. A total of 49 DNA samples were genotyped for the current study. Sixteen Corsican samples were selected from a larger dataset[12] in order to maximize geographic coverage (Supplementary Fig. S1) and DNA quality criteria. All samples have grandparental origin in the geographic micro-regions of sampling locations. These Corsican samples have been analyzed previously in Y-chromosome surveys[12,43,44]. To extend the comparative dataset, additional samples were analysed: ten Portuguese, five French samples from Provence, eleven Italian samples from Piedmont and seven Italian samples from Tuscany. Samples from Piedmont and Tuscany have been studied for Y chromosome[45] and Tuscany samples also for mitochondrial DNA[46]. DNA samples have been collected from healthy unrelated individuals and all donors have provided informed consent. Experiments were carried out in accordance with the relevant guidelines and regulations of collaborative institutions. The research has been approved by the Research Ethics Committee of the University of Tartu.

Genome-wide SNP data. DNA was extracted from blood/saliva samples and genotyping was carried out on Illumina 660 K platform (Human660W-Quad BeadChip). New samples were combined with data from previous studies[47–55]. In total, 892 individuals from 67 populations were analysed (Supplementary Table S1). The merged dataset was preprocessed with PLINK v1.9[56] in order to include only autosomal SNPs with minor allele frequency $> 0.005\%$ and genotyping success $> 97\%$. The cryptic relationships between samples (relatives of 1st and 2nd degree) were controlled with software KING v1.4[57] and two samples (one Yemen and one North Kannadi) were randomly removed from detected relative pairs. For some analyses SNPs in strong linkage disequilibrium (pairwise genotypic correlation $R^2 > 0.4$) in a window of 1,000 SNPs, sliding the window by step of 25 SNPs, were excluded. Exact numbers of individuals, populations and markers used in each analysis are specified in Supplementary Table S1.

Runs of homozygosity. Runs of homozygosity (RoH) were inferred using PLINK v1.9[56], with sliding window of 50 SNPs (5,000 kb), allowing for one heterozygous and five missing calls per window. RoH were defined as regions of at least 50 consecutive homozygous SNPs spanning at least 1,500 kb, with a gap of less than 1,000 kb between adjacent regions. The required minimum density was set at 50 kb/SNP[29,58].

ADMIXTURE. Maximum likelihood unsupervised clustering algorithm implemented in ADMIXTURE[32] was used to infer putative ancestral components in the Corsican population in a worldwide context. Clustering was performed 100 times at $K = 2$ to $K = 15$ (Supplementary Fig. S2). Convergence between runs was assessed using log-likelihood scores (LL). According to a low level of variation in LL scores ($LLs < 1$) within the top 10% fraction of runs with the highest LLs, the global maximum was assumed to be reached at $K = 2$ to $K = 6$, $K = 9$, $K = 11$ and $K = 13$ (Supplementary Fig. S3B). The lowest cross-validation (CV) index, which points to the predictive accuracy of the model at a given K, was observed at $K = 6$ (Supplementary Fig. S3A).

F_{ST}. Mean population pairwise F_{ST} and standard deviation values were calculated using software EIGENSOFT v. 7.2[59,60].

f3 tests. f3 tests were performed with ADMIXTOOLS v. 4.1[33] on a subset of West Eurasian and North African populations. To remove outlier samples from included populations, a PC analysis was performed, and samples not clustering with their population on the PC plot were excluded (Supplementary Table S1). The outgroup f3 test of the form f3(Corsicans, X; Yoruba) was implemented to measure the amount of shared drift between Corsicans and other populations, with the Yoruba population from Nigeria being set as the outgroup. To test for evidence of admixture in the Corsican population as target, standard f3 statistics were computed using the test configuration f3(Corsicans; X, Y). Negative values of f3 statistics were considered statistically significant comparatively with two different Z-scores: $|Z| > 3$ and more stringent $|Z| > 4$.

ChromoPainter and fineSTRUCTURE. The haplotype-based structure of Corsicans and other European populations was explored by applying the ChromoPainter/fineSTRUCTURE pipeline[34]. First, genotype data was phased with SHAPEIT v.2[61] using default parameters and the HapMap phase II b37 genetic map. Subsequently, the painting profile of each individual was inferred using ChromoPainter. The nuisance parameters n and m were inferred by running ChromoPainter with -in -iM flags for 10 E-M iterations. Given the high computational requirements, the analysis has been carried out only on a subset of the data, using an approach similar to Montinaro *et al.* 2015[62]. In detail, five (where available, otherwise using all the possible samples) individuals were randomly selected from each population. The inferred parameters were finally used in ChromoPainter specifying all the available samples as donors and recipients. This resulted in two different matrices, the "chunkcount" and the "chunklength"; the former summarises, for each recipient, the number of fragments inherited from each donor, while in the latter the same statistics is expressed in total genomic (in centimorgan) length. The coancestry matrices based on the length (chunklength) and number (chunkcount) of fragments were then obtained by combining the different chromosomal outputs with ChromoCombine. The obtained chunkcount coancestry matrix was harnessed to identify homogeneous groups of individuals using fineSTRUCTURE. In detail, two different runs of 4,000,000 iterations were performed, discarding the first 1 million as burn-in and using a thin interval of 10,000. For each of the clustering approaches, a hierarchical tree was inferred by taking advantage of the "Tree"

method and the "maximum concordance state" approach, performing 1 million iterations. In order to assess the robustness of the clustering process, the pairwise coincidence statistics among individuals was evaluated (Supplementary Fig. S7).

PCA. Principal Component Analysis (PCA) was carried out using the coancestry matrix of coping vectors created with ChromoPainter. On the whole, 624 samples were used, including all the samples from Europe, the Caucasus and Anatolia and the sub-set of Near Easterners, that on fineSTRUCTURE tree (Supplementary Fig. S5), formed a sister-clade of the European cluster.

GLOBETROTTER. The time of admixture and mixture profile were estimated for all the previously inferred clusters (targets) using GLOBETROTTER[35]. In detail, the painting profile obtained by ChromoPainter was harnessed by testing for any evidence of admixture using the options null.ind = 1 prop.ind = 0, and performing 100 bootstrap iterations. For each of the inferred admixture events, only those characterised by bootstrap values for time of admixture between 1 and 400 were considered. Subsequently, the time of admixture was estimated by repeating the same steps with options null.ind = 0 and prop.ind = 1. For the "non-local" analysis the same procedure has been repeated excluding clusters from the Southern European group as possible sources of the target. We considered a generation time of 30 years[38].

Projecting ancient Europeans into modern Eurasian variation. In order to contextualise the genetic variation of modern individuals in a European pre-Iron Age context, a PCA analysis was performed in which the ancient genotype data were projected into the PCA inferred on modern Eurasians (Fig. 5A). For ancient samples, genotype data released by Lazaridis *et al*. 2017[63] and Olalde *et al*. 2018[64] were used, from which non relevant individuals and all the samples with more than 70% of missingness were removed. After the final filtering 222 samples were retained (Supplementary Table S1B). For modern samples, 568 West Eurasian Individuals were retained. In order to prevent the shrinkage bias in ancient individuals, the "autoshrink: YES" option was used.

qpADM. The admixture profile of all the analysed populations was reconstructed with qpADM[39], which harnesses a combination of f4 describing the relationship of "target" and "sources" with a set of outgroups. In detail, the reconstructed admixture profile of each target used any possible four-member combination drawn from a list of putative sources. As a preliminary step, qpWave[65] was used if: a) the tested sources were significantly different and b) the target may be reconstructed using a specific combination of sources. A p-value threshold of 0.01 was used, as well as the following set of sources (Supplementary Table S1B):

Anatolia_BA, Anatolia_N, Steppe_EMBA, WHG, EHG, Iran_N, Minoan_Lasithi, Mycenaean, Yorubas, Levant_N, CHG, Europe_EN, Europe_LNBA, Europe_MNChL

and the following set of Outgroups:

AfontovaGora3, EHG, ElMiron, GoyetQ116-1, Iran_N, Kostenki14, Levant_N, MA1, Mota, Natufian, Ust_Ishim, Vestonice16, CHG.

All the tests for which the fitted model was supported were shown in Supplementary Table S5.

Data Availability

The data for 49 sequences generated in the current study are available in the NCBI-GEO repository through accession nr. GSE129663 and on the Estonian Biocenter website ebc.ee/free_data.

References

1. Institut national de la statistique et des études économiques. *Estimation de la population au 1er janvier 2019*. (2019) Available at: https://www.insee.fr/fr/statistiques/1893198 (Accessed: 4th July 2019).
2. *The Oxford Illustrated Prehistory of Europe*. (Oxford University Press, 1994).
3. Benjamin, J. *et al*. Late Quaternary sea-level changes and early human societies in the central and eastern Mediterranean Basin: An interdisciplinary review. *Quat. Int.* **449**, 29–57 (2017).
4. Courtaud, P., Cesari, J., Leandri, F., Nebbia, P. & Perrin, T. *La sépulture mésolithique de Campu Stefanu* (Sollacaro, Corse-du-Sud, France) (2014).
5. Lugliè, C. Your path led trough the sea … The emergence of Neolithic in Sardinia and Corsica. *Quat. Int.* **470**, 285–300 (2018).
6. Modi, A. *et al*. Complete mitochondrial sequences from Mesolithic Sardinia. *Sci. Rep.* **7**, 1–10 (2017).
7. Le Bourdonnec, F. X. *et al*. Obsidians artefacts from Renaghju (Corsica Island) and the Early Neolithic circulation of obsidian in the Western Mediterranean. *Archaeol. Anthropol. Sci.* **7**, 441–462 (2015).
8. Arrighi, J.-M. *Histoire de la langue corse*. (Editions Jean-Paul Gisserot, 2002).
9. Grimaldi, M. C., Crouau-Roy, B., Contu, L. & Amoros, J. P. Molecular variation of HLA class I genes in the Corsican population: approach to its origin. *Eur. J. Immunogenet.* **29**, 101–107 (2002).
10. Vona, G., Moral, P., Memmì, M., Ghiani, M. E. & Varesi, L. Genetic structure and affinities of the Corsican population (France): classical genetic markers analysis. *Am. J. Hum. Biol.* **15**, 151–163 (2003).
11. Tofanelli, S., Taglioli, L., Varesi, L. & Paoli, G. Genetic history of the population of Corsica (western Mediterranean) as inferred from autosomal STR analysis. *Hum. Biol.* **76**, 229–251 (2004).
12. Di Cristofaro, J. *et al*. Prehistoric migrations through the Mediterranean basin shaped Corsican Y-chromosome diversity. *PLoS One* **13**, e0200641 (2018).
13. Morelli, L., Paoli, G. & Francalacci, P. Surname analysis of the Corsican population reveals an agreement with geographical and linguistic structure. *J. Biosoc. Sci.* **34**, 289–301 (2002).
14. Varesi, L., Memmi, M., Cristofari, M. C., Mameli, G., Calo, C. & Vona, G. Mitochondrial control–region sequence variation in the Corsican population, France. *Am. J. Hum. Biol.* **12**, 339–351 (2000).

15. Falchi, A. *et al*. Genetic history of some western Mediterranean human isolates through mtDNA HVR1 polymorphisms. *J. Hum. Genet.* **51**, 9–14 (2006).
16. Memmi, M. *et al*. Genetic structure of southwestern Corsica (France). *Am. J. Hum. Biol.* **10**, 567–577 (1998).
17. Scozzari, R. *et al*. Human Y-chromosome variation in the western Mediterranean area: implications for the peopling of the region. *Hum. Immunol.* **62**, 871–884 (2001).
18. Francalacci, P. *et al*. Peopling of three Mediterranean islands (Corsica, Sardinia, and Sicily) inferred by Y-chromosome biallelic variability. *Am. J. Phys. Anthropol.* **121**, 270–279 (2003).
19. Morelli, L. *et al*. Frequency distribution of mitochondrial DNA haplogroups in Corsica and Sardinia. *Hum. Biol.* **72**, 585–595 (2000).
20. Gonzalez-Perez, E. *et al*. The ins and outs of population relationships in west-Mediterranean islands: Data from autosomal Alu polymorphisms and Alu/STR compound systems. *J. Hum. Genet.* **52**, 999–1010 (2007).
21. Montinaro, F. *et al*. Using forensic microsatellites to decipher the genetic structure of linguistic and geographic isolates: A survey in the eastern Italian Alps. *Forensic Sci. Int. Genet.* **6**, 827–833 (2012).
22. Kristiansson, K., Naukkarinen, J. & Peltonen, L. Isolated populations and complex disease gene identification. *Genome Biol.* **9**, 109 (2008).
23. Lettre, G. & Hirschhorn, J. N. Small island, big genetic discoveries. *Nat. Genet.* **47**, 1224–1225 (2015).
24. Latini, V., Sole, G., Varesi, L., Vona, G. & Ristaldi, M. S. The value of some Corsican sub-populations for genetic association studies. *BMC Med. Genet.* **9**, 73 (2008).
25. Falchi, A. *et al*. Prevalence of genetic risk factors for coronary artery disease in Corsica island (France). *Exp. Mol. Pathol.* **79**, 210–213 (2005).
26. Falchi, A. *et al*. Cholesteryl ester transfer protein gene polymorphisms are associated with coronary artery disease in Corsican population (France). *Exp. Mol. Pathol.* **83**, 25–29 (2007).
27. Piras, I. S. *et al*. High frequencies of short alleles of NOS1 (CA)n polymorphism in β039 carriers from Corsica Island (France). *Exp. Mol. Pathol.* **86**, 136–137 (2009).
28. McQuillan, R. *et al*. Runs of Homozygosity in European Populations. *Am. J. Hum. Genet.* **83**, 359–372 (2008).
29. Kirin, M. *et al*. Genomic runs of homozygosity record population history and consanguinity. *PLoS One* **5**, e13996 (2010).
30. Ceballos, F. C., Joshi, P. K., Clark, D. W., Ramsay, M. & Wilson, J. F. Runs of homozygosity: Windows into population history and trait architecture. *Nat. Rev. Genet.* **19**, 220–234 (2018).
31. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
32. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
33. Patterson, N. *et al*. Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
34. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet.* **8**, 11–17 (2012).
35. Hellenthal, G. *et al*. A genetic atlas of human admixture history. *Science* **343**, 747–751 (2014).
36. Busby, G. B. J. *et al*. The role of recent admixture in forming the contemporary West Eurasian genomic landscape. *Curr. Biol.* **25**, 2518–2526 (2015).
37. Raveane, A. *et al*. Population structure of modern-day Italians reveals patterns of ancient and archaic ancestries in Southern Europe. *Sci. Adv.* **5**, eaaw3492 (2019).
38. Tremblay, M. & Vézina, H. New estimates of intergenerational time intervals for the calculation of age and origins of mutations. *Am. J. Hum. Genet.* **66**, 651–658 (2000).
39. Haak, W. *et al*. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207–211 (2015).
40. Lazaridis, I. *et al*. Genomic insights into the origin of farming in the ancient Near East. *Nature* **536**, 419–424 (2016).
41. Chiang, C. W. K. *et al*. Genomic history of the Sardinian population. *Nat. Genet.* **50**, 1426–1434 (2018).
42. Sarno, S. *et al*. Ancient and recent admixture layers in Sicily and Southern Italy trace multiple migration routes along the Mediterranean. *Sci. Rep.* **7**, 1984 (2017).
43. King, R. J. *et al*. The coming of the Greeks to Provence and Corsica: Y-chromosome models of archaic Greek colonization of the western Mediterranean. *BMC Evol. Biol.* **11**, 69 (2011).
44. Keller, A. *et al*. New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nat. Commun.* **3** (2012).
45. Grugni, V. *et al*. Reconstructing the genetic history of Italians: new insights from a male (Y-chromosome) perspective. *Ann. Hum. Biol.* **45**, 44–56 (2018).
46. Achilli, A. *et al*. Mitochondrial DNA Variation of Modern Tuscans Supports the Near Eastern Origin of Etruscans. *Am. J. Hum. Genet.* **80**, 759–768 (2007).
47. Behar, D. M. *et al*. No evidence from genome-wide data of a Khazar origin for the Ashkenazi Jews. *Hum. Biol.* **85**, 859–900 (2013).
48. Behar, D. M. *et al*. The genome-wide structure of the Jewish people. *Nature* **466**, 238–242 (2010).
49. Kovacevic, L. *et al*. Standing at the gateway to Europe - The genetic structure of Western Balkan populations based on autosomal and haploid markers. *PLoS One* **9**, e105090 (2014).
50. Kushniarevich, A. *et al*. Genetic heritage of the balto-slavic speaking populations: A synthesis of autosomal, mitochondrial and Y-chromosomal data. *PLoS One* **10**, e0135820 (2015).
51. Li, J. Z. *et al*. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100–1104 (2008).
52. Metspalu, M. *et al*. Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. *Am. J. Hum. Genet.* **89**, 731–744 (2011).
53. Raghavan, M. *et al*. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* **505**, 87–91 (2014).
54. Yunusbayev, B. *et al*. The caucasus as an asymmetric semipermeable barrier to ancient human migrations. *Mol. Biol. Evol.* **29**, 359–365 (2012).
55. Yunusbayev, B. *et al*. The genetic legacy of the expansion of Turkic-speaking nomads across Eurasia. *PLoS Genet.* **11**, e1005068 (2015).
56. Chang, C. C. *et al*. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* **4**, 1–16 (2015).
57. Manichaikul, A. *et al*. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
58. Joshi, P. K. *et al*. Directional dominance on stature and cognition in diverse human populations. *Nature* **523**, 459–462 (2015).
59. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
60. Price, A. L. *et al*. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
61. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6 (2013).
62. Montinaro, F. *et al*. Unravelling the hidden ancestry of American admixed populations. *Nat. Commun.* **6**, 6596 (2015).
63. Lazaridis, I. *et al*. Genetic origins of the Minoans and Mycenaeans. *Nature* **548**, 214–218 (2017).
64. Olalde, I. *et al*. The Beaker phenomenon and the genomic transformation of northwest Europe. *Nature* **555**, 190–196 (2018).
65. Reich, D. *et al*. Reconstructing Native American population history. *Nature* **488**, 370–374 (2012).

Acknowledgements

We thank all volunteers who donated their DNA samples. We would like to thank Bayazit Yunusbayev for helpful discussions, Viljo Soo for his help in genotyping and Tuuli Reisberg for assistance in data management. Computational analyses were performed at the High Performance Computing Center of the University of Tartu. This research was supported by institutional research funding IUT (IUT24-1) of the Estonian Ministry of Education and Research (ET, EP); the Estonian Research Council grants PUT (PRG243) (EP, MM) and PUT (PUT1339) (AK); the European Union through the European Regional Development Funds with projects No. 2014-2020.4.01.16-0030 (MM, FM) and No. 2014-2020.4.01.15-0012 (MM); the European Union through Horizon 2020 grant no. 810645 (MM); the University of Pavia strategic theme "Towards a governance model for international migration: an interdisciplinary and diachronic perspective" (MIGRAT-IN-G) (OS); the Italian Ministry of Education, University and Research (MIUR): Dipartimenti di Eccellenza Program (2018–2022), Dept. of Biology and Biotechnology "L. Spallanzani", University of Pavia (AR and OS).

Author Contributions

M.M. conceived the study. M.M., F.M. and E.T. designed the research. O.S., A.R., J.C., J.D.C., S.M., L.P. and M.M. contributed to sample collection. E.T. and F.M. performed the analyses. E.T, F.M., A.K., M.M. and E.P. interpreted the results. E.T. and F.M. wrote the manuscript with input from all the coauthors.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-49901-8.

Competing Interests: The authors declare no competing interests.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.