ARTICLES
https://doi.org/10.1038/s41588-018-0215-8
nature genetics

Genomic history of the Sardinian population

Charleston W. K. Chiang[1,2,3*], Joseph H. Marcus[4], Carlo Sidore[5,6], Arjun Biddanda[4], Hussein Al-Asadi[7], Magdalena Zoledziewska[5], Maristella Pitzalis[5], Fabio Busonero[5,6], Andrea Maschio[5], Giorgio Pistis[5,6], Maristella Steri[5], Andrea Angius[5], Kirk E. Lohmueller[3], Goncalo R. Abecasis[6], David Schlessinger[8], Francesco Cucca[5,9] and John Novembre[4*]

The population of the Mediterranean island of Sardinia has made important contributions to genome-wide association studies of complex disease traits and, based on ancient DNA studies of mainland Europe, Sardinia is hypothesized to be a unique refuge for early Neolithic ancestry. To provide new insights on the genetic history of this flagship population, we analyzed 3,514 whole-genome sequenced individuals from Sardinia. Sardinian samples show elevated levels of shared ancestry with Basque individuals, especially samples from the more historically isolated regions of Sardinia. Our analysis also uniquely illuminates how levels of genetic similarity with mainland ancient DNA samples varies subtly across the island. Together, our results indicate that within-island substructure and sex-biased processes have substantially impacted the genetic history of Sardinia. These results give new insight into the demography of ancestral Sardinians and help further the understanding of sharing of disease risk alleles between Sardinia and mainland populations.

How complex traits change through time is a central question in evolutionary biology and genetics. Human genetics provides a compelling context for studying this process but requires populations where it is possible to integrate trait mapping with a detailed knowledge of population history. The people of the Mediterranean island of Sardinia are particularly well suited for genetic studies as evident from a number of successes in complex trait and disease mapping[1]. For example, early studies illuminated the genetic basis of thalassemia and more recent studies have mapped novel quantitative trait loci for traits such as hemoglobin levels[2], inflammation levels[2,3], height[4], and diseases such as multiple sclerosis[5] and type 1 diabetes[6]. These autoimmune diseases and hematological diseases like beta-thalassemia show unique incidences in Sardinia (for example, see Marrosu et al.[7], Pugliatti et al.[8], and Cao and Galanello[9]). Understanding how and why these conditions reached their frequencies in Sardinia would provide valuable insights into the dynamics of complex trait evolution. Yet, to empower such studies, a detailed understanding of the population history of Sardinia is needed.

One key characteristic of Sardinia is its differentiation from mainland populations, as evidenced by a distinctive cultural, linguistic, and archaeological legacy[10,11]. Early genetic studies made clear that Sardinia has also been a genetically isolated population on the basis of classical autosomal markers, uniparental markers, and elevated linkage disequilibrium (LD)[12–17]. Partly on this basis, Sardinia was included in the Human Genome Diversity Project (HGDP; see Cann[18]), which has been used as a reference sample in many studies, including recent ancient DNA (aDNA) studies of Europe[19–23]. Despite substantial research and interest in Sardinia, genetic studies of its demographic history are still incomplete.

One of the most remarkable findings to date regarding Sardinia's demographic history is that it has the highest detected levels of genetic similarity to ancient Neolithic farmer peoples of Europe[19–21,23–27]. This result is currently interpreted in a model with three ancestral populations that contribute ancestry to modern European populations[19,21,24,26]. The model postulates that early Neolithic farmers from the Near East and Anatolia expanded into Europe ~7,500–8,000 years ago and mixed in varying proportions with the existing pre-Neolithic hunter-gatherers in Europe. Then, a substantial post-Neolithic expansion of steppe pastoralists (associated with the Yamnaya culture) in the Bronze Age ~4,500–5,000 years ago introduced a third major component of ancestry across Europe. This model has been useful in explaining patterns observed in ancient and modern DNA data throughout Europe; here we look closely at how well it explains Sardinian population history.

In this model, Sardinia is effectively colonized by early Neolithic farmers during the European Neolithic, with minor contributions from pre-Neolithic hunter-gatherer groups. Sardinia then remained largely isolated from subsequent migrations on the mainland[27], including the Bronze Age expansions of the steppe pastoralists[19,21,24]. Support for this model is based on single nucleotide polymorphism (SNP) array data and has additional support from an ancient mitochondrial DNA (mtDNA) study in Sardinia, which showed relative isolation from mainland Europe since the Bronze Age, particularly for the Ogliastra region[28]. There is also support for this model in the relatively low frequency in Sardinia of U haplogroups that are markers of hunter-gatherer ancestry[19,29–32]. The archaeological record in Sardinia is also broadly supportive of such a model; there are few notable sites from the pre-Neolithic, followed by an expansion of sites in the Neolithic and subsequent development of a unique local

[1]Center for Genetic Epidemiology, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA. [2]Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Behavior, University of California, Los Angeles, Los Angeles, CA, USA. [3]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, Los Angeles, CA, USA. [4]Department of Human Genetics, University of Chicago, Chicago, IL, USA. [5]Istituto di Ricerca Genetica e Biomedica, Consiglio Nazionale delle Ricerche, Monserrato, Cagliari, Italy. [6]Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA. [7]Committee on Evolutionary Biology, University of Chicago, Chicago, IL, USA. [8]Laboratory of Genetics, National Institute on Aging, US National Institutes of Health, Baltimore, MD, USA. [9]Dipartimento di Scienze Biomediche, Università degli Studi di Sassari, Sassari, Italy. *e-mail: charleston.chiang@med.usc.edu; jnovembre@uchicago.edu

cultural assemblage (Nuragic culture) by the Bronze Age in Sardinia (see Francalacci et al.[33] and Caramelli et al.[34]).

The conclusion that Sardinia effectively descended from early Neolithic farmers is not without question though. The first human remains on Sardinia date to the Upper Paleolithic and flint-stone instruments are found in the Lower Paleolithic[13,35], so the potential earliest residents arrived in Sardinia ~14,000–18,000 years ago, much earlier than the Neolithic. Studies using Y-chromosome haplotypes (particularly haplotype I2a1a1) have found Mesolithic or Paleolithic dates for the common ancestor of Sardinian-specific haplotypes and have interpreted these results as evidence for a strong pre-Neolithic component of Sardinian ancestry[13,36–39], although this interpretation is controversial[28,29,34,40,41]. Moreover, the relatively high prevalence of haplogroup R1b1a2 (R-M269) in Sardinia (~18%) has been interpreted to reflect a large component of pre-Neolithic hunter-gatherer ancestry in Sardinia (see Contu et al.[36] and Morelli et al.[36]), although recent modern and aDNA studies support a recent coalescent time for the haplogroup throughout Europe, in line with an expansion during the late Neolithic and Bronze Age[19,42,43]. Sex-biased migration processes[44,45] are a common occurrence in humans, and such processes can give rise to differing patterns on autosomal versus uniparental markers.

Here, to provide novel insight into the peopling of Sardinia and its relationship to mainland populations, we analyze a collection of 3,514 individual whole-genome sequences (WGS) sampled as part of the SardiNIA project[2] (see URLs). We specifically assess whether the isolation of Sardinia is consistent with a dominantly early Neolithic farmer or hunter-gatherer peopling and whether there is evidence for sex-biased processes. As part of our analysis we address the hypothesis of an ancestral connection between Sardinia and the Basque populations of Spain and France. (The Basque also show high affinity with early Neolithic farmer aDNA[21,46].) We also address whether gene flow from North African populations to Sardinia[33,47–49] has been substantial, since Sardinia may be the result of a recent multiway admixture involving sources from around the Mediterranean[47].

A key factor in our analysis is the evaluation of the internal substructure within Sardinia. Numerous studies have noted relative heterogeneity of small subpopulations within Sardinia[28–30,34,50,51]. Our dataset, including broadly dispersed samples from across Sardinia[52,53] and a deep sampling of several villages from the Lanusei Valley region of the Ogliastra province[2,54,] is especially well suited to systematically assess the extent and source of such heterogeneity (Fig. 1). We show that samples from the Ogliastra province and the broader, mountainous Gennargentu region have signs of elevated isolation. Thus, we frame our analyses of the demographic history of Sardinia by contrasting results for sampled individuals from the Gennargentu region with those outside of the region.

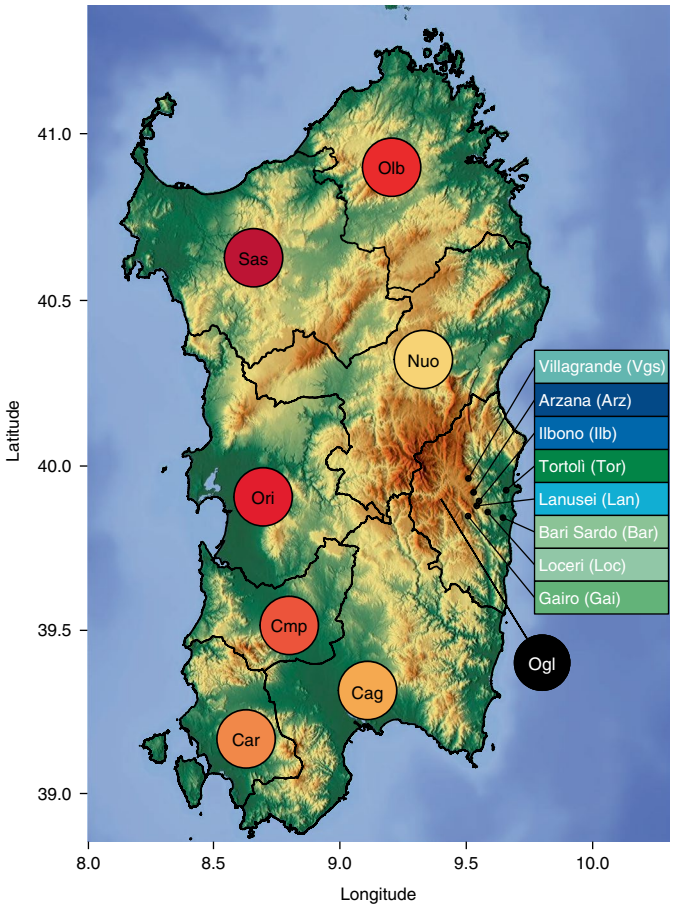

Fig. 1 | Geographical map of Sardinia. The provincial boundaries are given as black lines. The provinces are abbreviated as Cag (Cagliari), Cmp (Campidano), Car (Carbonia), Ori (Oristano), Sas (Sassari), Olb (Olbia-tempio), Nuo (Nuoro), and Ogl (Ogliastra). For sampled villages within Ogliastra, the names and abbreviations are indicated in the colored boxes. The color corresponds to the color used in the PCA plot (Fig. 2a). The Gennargentu region referred to in the main text is the mountainous area shown in brown that is centered in western Ogliastra and southeastern Nuoro.

Results

Before addressing broader-scale questions regarding Sardinian demographic history, we first examine the population structure within Sardinia. To lessen the confounding of recent internal migrations, we focus on a subset (N=1,577) of unrelated individuals with at least three grandparents originating from the same geographical location.

We find that the strongest axis of genetic variation is between individuals from Ogliastra (Ogl) and those outside of Ogliastra (Fig. 2a,b). A notable exception is the subpopulation from Tortolì, a recently developed coastal city and the main seaport of the Ogliastra region. Samples from Tortolì show closer affinity in the principal component analysis (PCA) to samples from the western part of the island. Samples from outside Ogliastra show lower levels of differentiation by fixation index (F_{ST}) (Fig. 2c) and higher levels of allele sharing among themselves (Fig. 3), despite the greater geographical distance between populations. When we use a spatially explicit statistical method (estimated effective migration surface (EEMS), see Petkova et al.[55]) for visualizing genetic diversity patterns, the resulting effective migration surface (Fig. 2d) is consistent with high effective migration in the western regions of Sardinia connecting the major populations centers of Cagliari (Cag), Oristano (Ori), and Sassari (Sas). Low effective migration rates separate these provinces from a broad area that extends to the mountainous Gennargentu massif region, including inland Ogliastra to the west. The Gennargentu region is also where some of the Sardinian individuals in the HGDP originate (A. Piazza, personal communication). We find that the HGDP Sardinian individuals partially overlap with our dataset and include a subset that clusters near the Ogliastra subpopulation (Supplementary Figs. 1, 2, Supplementary Tables 1, 2). Thus, we use the term 'Gennargentu region' to describe this ancestry component (red component in Fig. 2b). Based on these results, and to simplify analyses going forward, we use individuals from the town of Arzana (Arz) as a representative of the Gennargentu region ancestry component and Cagliari as a representative of ancestry outside the Gennargentu region.

Sardinia as an isolated Mediterranean population. To assess Sardinian variation in a regional context, we created a merged dataset of Sardinian with Mediterranean populations from the Human Origins Array (HOA)[21]. PCA of this shows a one-dimensional isolation-by-distance pattern around the Mediterranean, from

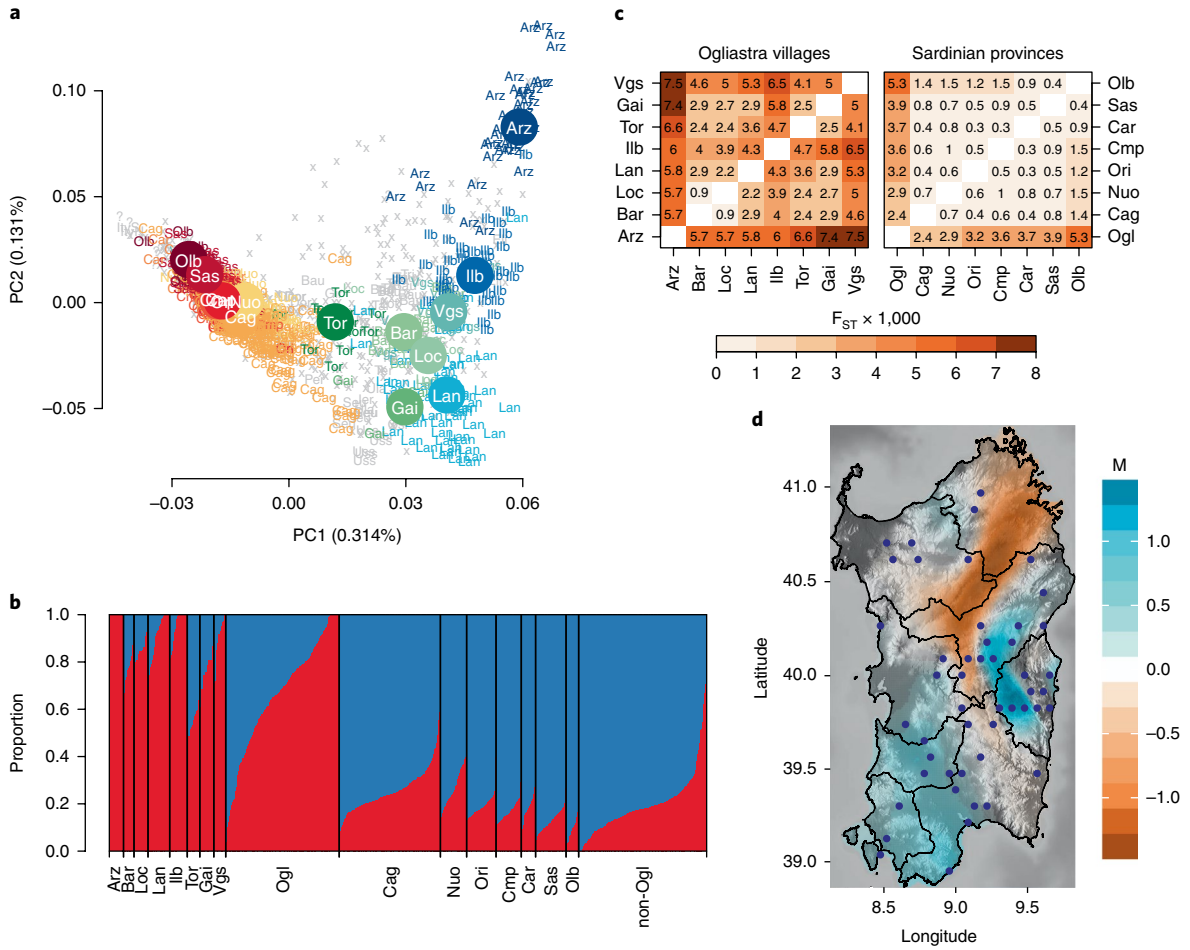

Fig. 2 | Within-island population structure. a, Top two principal components of PCA based on 1,577 unrelated Sardinians. Each individual is labeled by geographical origin as defined by grandparental birthplaces (see Methods); otherwise they are labeled with 'x' or '?' for mixed grandparental origin or missing information, respectively. Subpopulations with fewer than eight individuals are displayed in the background in gray color. **b**, Admixture result at $K=2$, which had the lowest cross-validation errors from $K=2$ to $K=7$ (not shown). Individuals from Ogliastra and outside Ogliastra who were not assigned to a major location or had mixed grandparental origins are grouped under Ogl and non-Ogl, respectively. Bars for locations with individuals fewer than 40 (Bar, Gai, Loc, Olb, Tor, Vgs) were expanded to a fixed minimum width to aid visualization. **c**, Genetic differentiation among Ogliastra villages (left) or among Sardinian provinces (right) as measured by Weir and Cockerham's unbiased estimator of F_{ST}. Ogliastra appears to be the most differentiated from other provinces; within Ogliastra, the level of differentiation between villages is substantial (reaching as high as 0.0075 between Villagrande and Arzana), with Arzana being consistently well differentiated from other villages. **d**, EEMS plot within Sardinia based on 181 Sardinians with all four grandparents born in the same location. Arz, Arzana; Bar, Bari Sardo; Cag, Cagliari; Car, Carbonia; Cmp, Campidano; Gai, Gairo; Ilb, Ilbono; Lan, Lanusei; Loc, Loceri; Ogl, Ogliastra; Olb, Olbia-tempio; Ori, Oristano; Nuo, Nuoro; Sas, Sassari; Tor, Tortolì; Vgs, Villagrande.

North Africa through the Near East and then toward Iberia[56–59], with Sardinian samples clustering offset from southern European samples (Fig. 4a). The effective migration surface shows the Mediterranean Sea isolating Sardinia from neighboring mainland populations, with stronger isolation between Sardinia and North Africa than Sardinia and mainland Europe (Fig. 4b). An analysis with ADMIXTURE further supports this isolation of Sardinian populations (Fig. 4c, Supplementary Fig. 3). Across analyses of varying number of population clusters, Sardinians tend to form a distinct cluster with all individuals near 100% ancestry (Supplementary Fig. 3); this is consistent with relatively high levels of differentiation (F_{ST} ~0.023–0.037 between the 'blue' component and other ancestral components in Supplementary Fig. 3), which may results from extended divergence and/or elevated rates of drift.

Timescale of divergence and population size history. While the relative isolation of Sardinia is apparent, the timescale of the divergence is unclear. We used a recent approach that leverages information from both sequential Markovian coalescent and site frequency spectra-based frameworks (SMC++, see Terhorst et al.[60]) to infer an approximate divergence time and population size history. SMC++ infers a divergence time reflecting the time point in an idealized two-population split model after which effective migration between populations becomes negligible; thus, it should be expected to underestimate divergence times when post-divergence gene flow has taken place. For the mainland European ancestry CEU and TSI populations, SMC++ infers a divergence time of 14.4 ± 3.5 generations (or ~430 years ago). Sardinia is estimated as having a deeper divergence time with each of these populations, with an estimated divergence time of 143.3 ± 1.3 generations (~4,300 years ago) between Sardinia and TSI and 231.7 ± 12.9 generations (~7,000 years ago) between Sardinia and CEU (Fig. 5a). We complemented this approach with another commonly used method (the multiple sequentially Markovian coalescent (MSMC), see Schiffels and Durbin[61]). Consistently, MSMC estimates Sardinia as more deeply diverged from the CEU and TSI populations than CEU

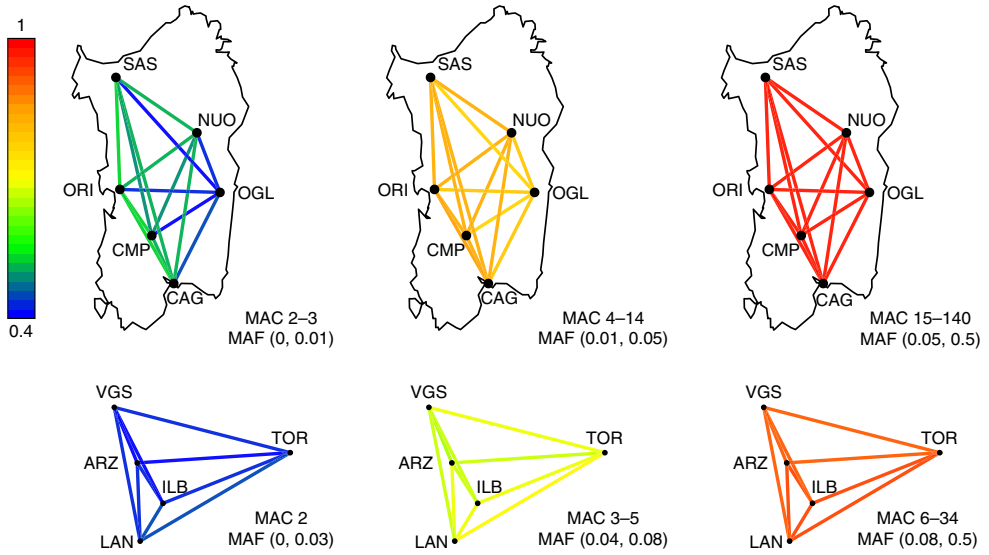

Fig. 3 | Allele sharing across the island and within Ogliastra. As a function of allele frequencies, allele sharing across the island (top) and within Ogliastra (bottom) are shown. Allele sharing between a pair of populations is defined as the probability that two randomly drawn carriers of the allele of a given MAF are from different populations, normalized by the panmictic expectation. Allele sharing is visualized here by the color of the lines connecting two populations. Island-wide analysis used subpopulations with at least 70 individuals and the minor allele counts (MACs) of each of 1 million randomly selected variants were downsampled to 140 chromosomes. Within Ogliastra, the analysis used subpopulations with at least 17 individuals and the MACs were downsampled to 34 chromosomes. ARZ, Arzana; CAG, Cagliari; CMP, Campidano; ILB, Ilbono; LAN, Lanusei; OGL, Ogliastra; ORI, Oristano; NUO, Nuoro; SAS, Sassari; TOR, Tortolì; VGS, Villagrande.

and TSI are to each other (Supplementary Fig. 4a). Both methods also show that Sardinia has had lower long-term effective population sizes and lacks the signature of strong population growth typical of mainland European populations (Fig. 5b, Supplementary Fig. 4b). Sardinian populations from the Ogliastra province (Arzana, Lanusei, and Ilbono) showed consistently lower population size, while Sardinian populations from outside of Ogliastra (Cagliari) showed a pattern of growth more similar to that observed in CEU and TSI (Supplementary Fig. 5a,b). Sardinian populations from the Ogliastra province showed a more ancient split time with mainland European populations, while Cagliari showed a more recent split time (Supplementary Fig. 5c).

Sardinia in relation to other Mediterranean populations. Due to its smaller long-term effective population size (Fig. 5b), Sardinia is expected to have undergone accelerated genetic drift. To correct for this when measuring similarity to other mainland populations, we used the 'shared drift' outgroup-f_3 statistics[62], which measures the length of shared branch length between two populations relative to an outgroup. Using this metric, we find that the Basque are the most similar to Sardinia, even more so than mainland Italian populations such as Tuscan and Bergamasque (Supplementary Fig. 6a,b). We also tested the affinity between Sardinians and Basque with the *D*-statistics of the form *D(Outgroup, Sardinia; Bergamo or Tuscan, Basque)*. In this formulation, significant allele sharing between Sardinia and Basque, relative to sharing between Sardinia and Italian populations, will result in positive values for the *D*-statistics. Sardinia consistently showed increased sharing with the Basque populations compared to mainland Italians ($Z > 4$; Supplementary Fig. 6c). The result was stronger when using the Arzana than the Cagliari sample ($D(Outgroup, Basque, CAG, ARZ) = 0.0020$ and 0.0021 for French Basque and Spanish Basque, respectively; $Z > 3.2$). In contrast, sharing with other Spanish samples was generally weaker and not significant (Supplementary Fig. 6c), suggesting the shared drift with the Basque is not mediated through modern Spanish ancestry.

The admixture and PCA analyses described earlier (Fig. 4) suggest that Sardinian samples, particularly outside of Ogliastra, may be admixed with mainland sources, as suggested previously[47–49]. For example, Cagliari individuals demonstrated ~10% of a non-Sardinian component ('green' in Fig. 4c) that is found among extant individuals from Southern Europe, the Middle East, Caucasus, and North Africa. To assess this further, we used the f_3 test for admixture[63]. Contrary to mainland Europeans, we found none of the Sardinian populations showed evidence of admixture (Supplementary Fig. 7). Because f_3-based tests may lose power when applied to populations that have experienced extensive drift post-admixture[63], we also tested for admixture using a complementary LD-based approach (Admixture-induced Linkage Disequilibrium for Evolutionary Relationships (ALDER), see Loh et al.[48]). Using this approach, a number of Sardinian populations outside Ogliastra are inferred to be admixed (Table 1, Supplementary Table 3). The inferred source populations are typically a mainland Eurasian population and a sub-Saharan African population. The admixture proportions range from 0.9 to 5% of sub-Saharan ancestry by the f_4-ratio estimator[63] with estimated admixture dates of approximately 62–101 generations (Table 1, Supplementary Table 3).

Elevated Neolithic and pre-Neolithic ancestry. Ancient DNA studies have shown that across the autosome, Sardinians exhibit higher levels of Neolithic farmer ancestry compared to mainland Europeans[19,21]. However, because previous samples from Sardinia have been limited in sample size, we revisited the question using our dataset and addressing within-island variation.

We confirm that Sardinians have the highest observed levels of shared drift with early Neolithic farmer cultures (represented by the LBK380 sample from Stuttgart, Germany[21]; hereafter referred to as 'Stuttgart') and relatively low levels of shared drift with earlier hunter-gatherer cultures (represented by an aDNA sample from Loschbour rock shelter in Luxembourg[21]; hereafter referred to as 'Loschbour') (Fig. 6a, Supplementary Fig. 8). As expected, the Neolithic farmer ancestry component is more abundant than the

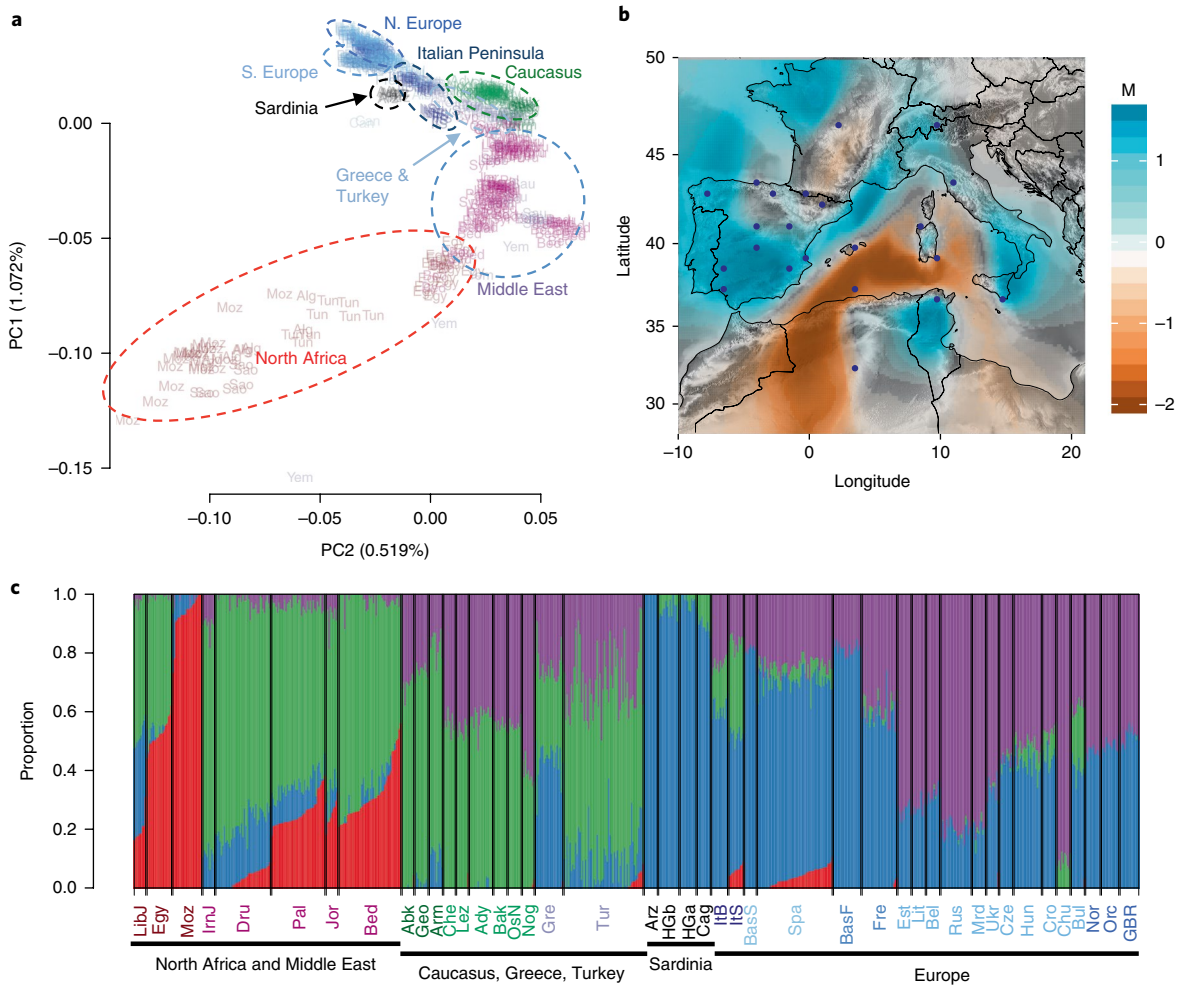

Fig. 4 | Population structure relative to mainland Europeans. a, Top two principal components of PCA of the merged dataset of Sardinia and HOA data. Populations are enclosed in dashed ellipses by major subcontinents. **b**, EEMS result for the pan-Mediterranean analysis. **c**, Admixture results at $K=4$, which has the lowest cross-validation error in analysis from $K=2$ to $K=15$. For clarity, only populations with a sample size > 8 are visualized. Arzana and Cagliari contained 100% and 89% of the European-dominant, 'blue', ancestry. Populations are ordered by subcontinental regions and then by population median values in PC1. (See Supplementary Fig. 3 for the full result.) Population labels are color-coded by major subcontinental regions. North Africa: LibJ, Jewish in Libya; Egy, Egyptian; Moz, Mozabite; Middle East: IrnJ, Jewish in Iran; Dru, Druze; Pal, Palestinian; Jor, Jordanian; Bed, Bedouin. Caucasus: Abk, Abkhasian; Geo, Georgian; Arm, Armenian; Che, Chechen; Lez, Lezgin; Ady, Adygei; Bak, Balkar; OsN, North Ossetian; Nog, Nogai. Turkey and Greece: Gre, Greece; Tur, Turkey. Europe: Arz, Arzana; HGb, HGDP Sardinian; HGa, HGDP Sardinian; Cag, Cagliari; ItB, Bergamasque; ItS, Sicilians; BasS, Spanish Basque; Spa, Spanish; BasF, French Basque; Fre, French; Est, Estonian; Lit, Lithuanian; Bel, Belarusian; Rus, Russian; Mrd, Mordovian; Ukr, Ukranian; Cze, Czech Republican; Hun, Hungarian; Cro, Croatian; Chu, Chuvash; Bul, Bulgarian; Nor, Norwegian; Orc, Orcadian; GBR, British (Great Britain).

hunter-gatherer ancestry component across all Sardinian populations (Supplementary Table 4). Surprisingly though, using supervised estimation of ancestry proportions[19] based on aDNA, we found an indication of higher levels of Neolithic and pre-Neolithic ancestries in the Gennargentu region, and higher levels of steppe pastoralist ancestry outside the region (Supplementary Fig. 9, Supplementary Table 5). Investigating this further, we found that shared drift with Neolithic farmers and with pre-Neolithic hunter-gatherers are significantly correlated with the proportion of Gennargentu region ancestral component estimated from admixture analysis, while that shared with steppe pastoralists is weakly negative and non-significantly correlated with Gennargentu region ancestry ($Z>6$ for Neolithic farmers and pre-Neolithic hunter-gatherers, $Z<2$ for steppe pastoralists; Fig. 6b, Supplementary Table 6). Moreover, the *D*-statistics of the form *D(Outgroup, Ancient, Ogliastra, Non-Ogliastra)* also support increased sharing with Neolithic and pre-Neolithic individuals, but not post-Neolithic individuals from the steppe, in the Ogliastra samples ($D=-0.0029$ and -0.0035, $Z=6.1$ and 6.8 when aDNA sample = Stuttgart and Loschbour, respectively; $D=-0.0002$, $Z=0.7$, when aDNA sample = Yamnaya).

Together, these results confirm that relative to the mainland, Sardinia appears to harbor the highest amount of Neolithic farmer ancestry and very little of the pre-Neolithic hunter-gatherer or Bronze Age pastoralists ancestries. We further found within-island variation of ancestry. Specifically, we found that with increasing levels of isolation (represented by increasing levels of Gennargentu ancestry), there is greater Neolithic farmer and pre-Neolithic hunter-gatherer ancestry, while steppe ancestry generally showed no significant correlation.

Sex-biased demography in prehistoric Sardinia. The relatively high frequencies and low divergences within two particular Y-chromosome haplogroups[33,36–39] (I2a1a1 at ~39% and R1b1a2 at ~18%) in Sardinia are a notable feature of Sardinian genetic variation. Neither haplogroup is typically affiliated with Neolithic ancestry in aDNA data, raising the potential of sex-biased processes in the history of Sardinia.

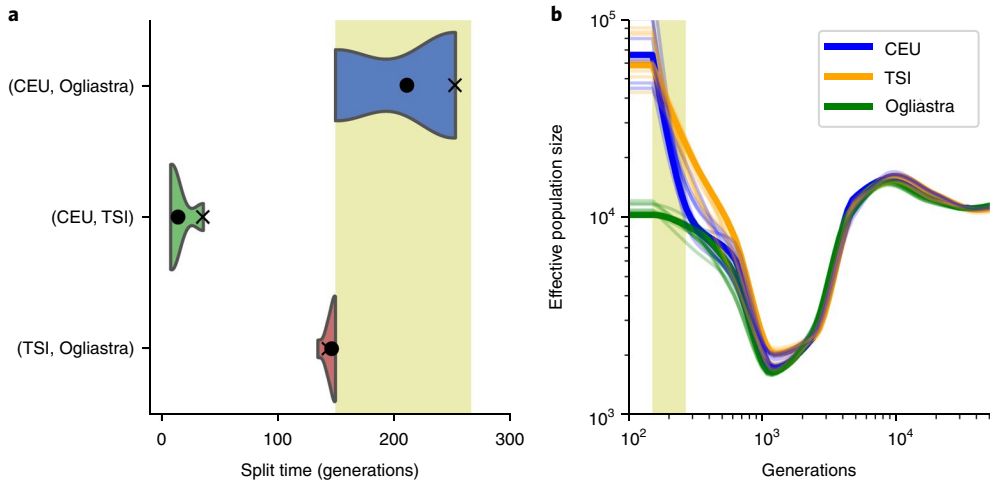

Fig. 5 | Coalescent-based inference of demographic history using SMC++. a, Inference of population divergence times. **b**, Population size history. Analysis based on 4 focal individuals and 90 low-coverage samples from the combined dataset of Lanusei and Arzana individuals (Ogliastra) and 1000 Genomes CEU and TSI. The uncertainty and mean point estimates of population divergence time are shown using ten bootstrap samples (the violin plot and the black dot, respectively). We also show the point estimate using all the data by the black cross. For population size trajectories, we estimated the size until 150 generations in the past and used 10 internal spline knots when running SMC++. Uncertainty reflected through ten bootstrap samples is also shown in the same but lighter colors. The orange shaded box denotes the Neolithic period, ~4500–8000 years ago, converted to units of generations assuming 30 years per generation.

To investigate further, we first use ADMIXTURE and contrast the inferred Gennargentu region ancestry on the X chromosome versus the autosome. Intriguingly, on average, we find a higher proportion of the Gennargentu region ancestry ('red' component in Supplementary Fig. 10) on the X chromosome (37%) than on the autosome (30%, $P < 1 \times 10^{-6}$ by permutation). The Gennargentu region ancestry is correlated with Neolithic or pre-Neolithic ancestries rather than more recent Bronze Age steppe ancestry (Fig. 6b), suggesting this result may be due to sex-biased processes in which more females than males carried the non-steppe ancestries. We also examined relative levels of nucleotide diversity on the X chromosome versus the autosome. Doing so, we found that Sardinia shows a high ratio of X-to-A diversity, particularly when compared to most mainland European populations (Supplementary Fig. 11), suggesting that Sardinian demographic history has had a relatively low male effective size.

Discussion

We investigated the fine-scale population structure and demography of the people of Sardinia using the WGS of 3,514 Sardinians with detailed self-reported ancestry that goes back two generations. The genotype calling leveraged extensive haplotype sharing to produce a high-quality call set[2], and we integrated the data with the 1000 Genomes, HGDP, and HOA reference sets. From our analyses, we could confirm a number of major features of previous analyses and provide more detail regarding the isolation between Sardinia and the mainland.

Our analysis of divergence times suggests the population lineage ancestral to modern-day Sardinia was effectively isolated from the mainland European populations ~140–250 generations ago, corresponding to ~4,300–7,000 years ago assuming a generation time of 30 years and a mutation rate of 1.25×10^{-8} per basepair per generation. However, these quantitative estimates should be treated with caution, since the SMC++ model assumes an idealized model of homogeneous ancestries with no post-divergence gene flow. Nevertheless, in terms of relative values, the divergence time between Northern and Southern Europeans is much more recent than either is to Sardinia, signaling the relative isolation of Sardinia from mainland Europe.

We documented fine-scale variation in the ancient population ancestry proportions across the island. The most remote and interior areas of Sardinia—the Gennargentu massif covering the central and eastern regions, including the present-day province of Ogliastra—are thought to have been the least exposed to contact with outside populations[28,30,51]. We found that pre-Neolithic hunter-gatherer and Neolithic farmer ancestries are enriched in this region of isolation. Under the premise that Ogliastra has been more buffered from recent immigration to the island, one interpretation of the result is that the early populations of Sardinia were an admixture of the two ancestries, rather than the pre-Neolithic ancestry arriving via later migrations from the mainland. Such admixture could have occurred principally on the island or on the mainland before the hypothesized Neolithic era influx to the island. Under the alternative premise that Ogliastra is simply a highly isolated region that has differentiated within Sardinia due to genetic drift, the result would be interpreted as genetic drift leading to a structured pattern of pre-Neolithic ancestry across the island, in an overall background of high Neolithic ancestry.

We found Sardinians show a signal of shared ancestry with the Basque in terms of the outgroup f_3 shared-drift statistics. This is consistent with long-held arguments of a connection between the two populations, including claims of Basque-like, non-Indo-European words among Sardinian placenames[64]. More recently, the Basque have been shown to be enriched for Neolithic farmer ancestry[21,46] and Indo-European languages have been associated with steppe population expansions in the post-Neolithic Bronze Age[19,24]. These results support a model in which Sardinians and the Basque may both retain a legacy of pre-Indo-European Neolithic ancestry[46]. To be cautious, while it seems unlikely, we cannot exclude that the genetic similarity between the Basque and Sardinians is due to an unsampled pre-Neolithic population that has affinities with the Neolithic representatives analyzed here.

We also examined possible sources of African admixture to Sardinia. Before our studies, there have been reports of a minor proportion (0.6–2.9%) of sub-Saharan admixture[48,49] and a multi-way admixture involving an African source[47] in HGDP Sardinians. In light of the close geographical proximity of Sardinia and North Africa, as well as the substantial admixture proportion from North Africa in Southern Europe[57], we tested for admixture using modern North African reference populations included in the HOA data (Tunisian, Algerian, Mozabite, Egyptian, and Saharawi). We found

Table 1 | Evidence of admixture as inferred by ALDER

Test population	Source 1	Source 2	Admixture date (generation)		Fitted amplitude ($\times 10^{-5}$)		*P* value	Admixture proportion	
			Mean	s.e.m.	Mean	s.e.m.		Mean	s.e.m.
	Outside Ogliastra								
Cagliari	Wambo	Spanish (Castilla y León)	62.57	6.61	2.70	0.310	1.28×10^{-13}	0.0039	0.0036
Campidano	Luhya	Tuscan	63.56	12.80	2.59	0.506	0.0285	0.0086	0.0030
Carbonia	No successful fit								
Nuoro	No successful fit								
Olbia-tempio	No successful fit								
Oristano	Wambo	Estonian	101.05	20.08	3.51	0.638	0.0195	0.049	0.0029
Sassari	Khwe	Lithuanian	82.21	15.92	2.53	0.430	2.25×10^{-8}	0.039	0.0027
	Ogliastra, HGDP Sardinians								
Arzana, Bari Sardo, Gairo, Ilbono, Lanusei, Loceri, Tortolì, Villagrande, SarHGDPa, SarHGDPb	No successful fit								

P value has been corrected for multiple hypothesis testing (number of pairs of source populations and analysis-wide number of test populations). Admixture proportions of the sub-Saharan ancestry are estimated with the f_4-ratio test, using Finnish and Chimp as the outgroups. ALDER, Admixture-induced Linkage Disequilibrium for Evolutionary Relationships; HGDP, Human Genome Diversity Project.

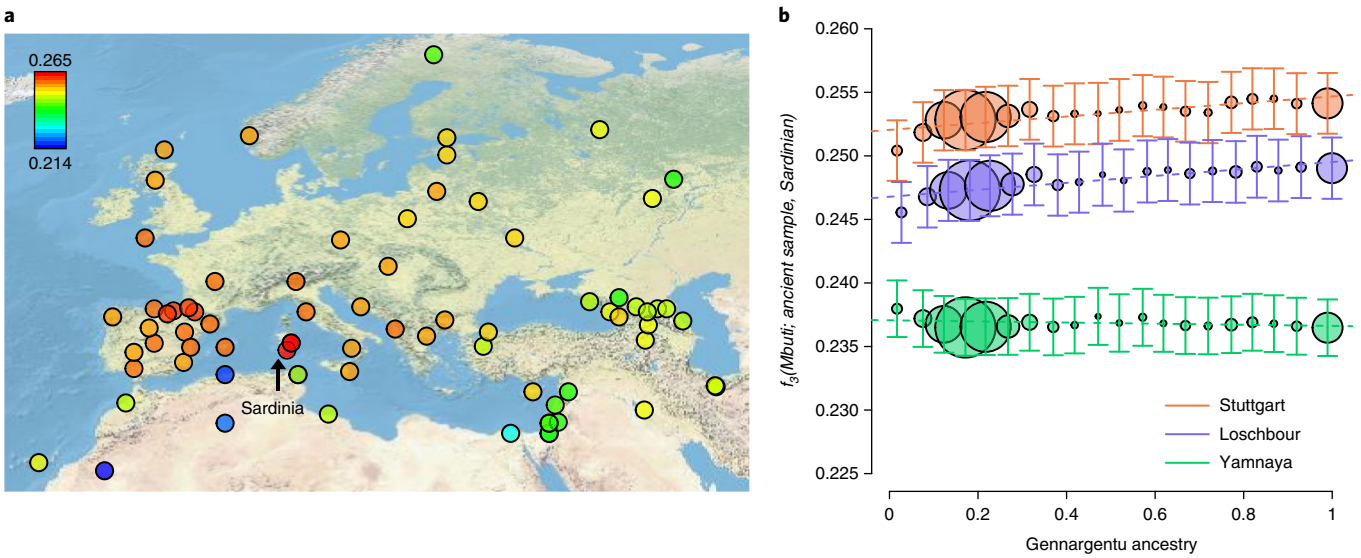

Fig. 6 | Similarity of ancient samples to populations across Europe and within Sardinia. a, Outgroup f_3 statistics of the form f_3*(Mbuti; Stuttgart, X)*, where X is a population across the merged dataset of Sardinian and HOA data. Higher f_3 values suggest a larger shared drift between a pair of populations. The arrow indicates Sardinian populations. **b**, Outgroup f_3 statistics of the form f_3*(Mbuti; Ancient, Sardinian)* across Sardinian samples binned in steps of 5% of Gennargentu ancestry estimated in Fig. 2b. The increase of outgroup f_3 statistics as a function of ancestry is positive for Stuttgart and Loschbour (0.00263 and 0.00274, respectively) and slightly negative for Yamnaya (-4.4×10^{-4}). Ancient samples used include a reference Neolithic farmer individual (Stuttgart, orange), a reference pre-Neolithic hunter-gatherer individual (Loschbour, blue), and a reference steppe population (Yamnaya, green) from the merged dataset (see Haak et al.[18] in Methods). Error bars represent the s.e.m. of the estimated f_3 values from the blocked jackknife procedure. The sizes of the circles are proportional to the number of samples per bin (max $N = 281$ per bin).

that the best proxy for African admixture is sub-Saharan African populations rather than Mediterranean North African populations, and we inferred the date of admixture as ~1,800–3,000 years ago (assuming 30 years per generation). The lack of a strong signal of North African autosomal admixture may be due to inadequate coverage of modern North African diversity in our reference sample, such that the sub-Saharan component of admixture we detected may be an indirect reflection of recent North African admixture (particularly if the North African source was admixed with sub-Saharan Africans; for example, see Pickrell et al.[65]). Alternatively, it may be due to a poor representation of ancestral North Africans. Present-day North African ancestry reflects large-scale recent gene flow during the Arab expansion (~1,400 years ago[58]). The sub-Saharan African admixture observed in the non-Ogliastra samples could be mediated through an influx of migrants from North Africa before the Arab expansion, for example, during the eras of trade relations and occupations from the Phoenicians, Carthaginians, and Romans (~700 B.C. to ~200 B.C., see Dyson and Rowland[10]).

While we can confirm that Sardinians principally have Neolithic ancestry on the autosomes, the high frequency of two Y-chromosome haplogroups[33,36–39] (I2a1a1 at ~39% and R1b1a2 at ~18%) that are not typically affiliated with Neolithic ancestry is one challenge to this model. Whether these haplogroups rose in frequency due to extensive genetic drift and/or reflect sex-biased demographic processes has been an open question. Our analysis of X chromosome versus autosome diversity suggests a smaller effective size for males, which can arise due to multiple processes, including polygyny, patrilineal inheritance rules, or transmission of reproductive success[66]. We also find that the genetic ancestry enriched in Sardinia is more prevalent on the X chromosome than the autosome, suggesting that male lineages may more rapidly trace back to the mainland. Considering that the R1b1a2 haplogroup may be associated with post-Neolithic steppe ancestry expansions in Europe[19], and the recent timeframe when the R1b1a2 lineages expanded in Sardinia[33], the patterns raise the possibility of recent male-biased steppe ancestry migration to Sardinia, as has been reported among mainland Europeans at large[45] (though see Lazaridis and Reich[67] and Goldberg et al.[68]). Such a recent influx is difficult to square with the overall divergence of Sardinian populations observed here. Thus, our results make clear that future studies aimed to understand sex-biased processes in the history of Sardinia, and European populations in general, will be illuminating, especially as systems of mating and dispersal may have shifted alongside modes of subsistence[69].

For the purposes of understanding complex trait evolution in Sardinian history, the results suggest that while Sardinia has clearly had influence from pre-Neolithic sources and contact with steppe ancestry populations, the demographic history is one of substantial isolation and abundant Neolithic ancestry relative to the mainland. For traits with a strong sex-linked component, our results encourage accounting for the sex-biased processes detected here. The relatively constant size of Sardinian populations predicts that there has been less of an influx of rare variants[70], as well as an increase of homozygosity, relative to other expanding populations. These two factors may increase the impact of dominance components of variation[71] and reduce the allelic heterogeneity of complex traits[72–74]. Armed with a better understanding of Sardinian prehistory and demographic events, we anticipate a more nuanced understanding of complex trait variation and disease incidences in Sardinia. The affinity to Neolithic farmer populations (and to a lesser extent, pre-Neolithic hunter-gatherer populations) also means that Sardinia is a potential reservoir for variants that may have been lost in mainland Europeans.

URLs. SardiNIA project, https://sardinia.nia.nih.gov/. MSMC Tools, https://github.com/stschiff/msmc-tools. Impute2, 1000 genomes phasing reference panel, https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#reference. Complete Genomics 69 Genomes Data, http://www.completegenomics.com/public-data/69-genomes/. Uniquely mapped reads, https://oc.gnz.mpg.de/owncloud/index.php/s/RNQAkHcNiXZz2fd. Simons Genome Diversity Project, https://www.simonsfoundation.org/life-sciences/simons-genome-diversity-project-dataset/. Neutral regions of the autosome and X chromosome, http://hammerlab.biosci.arizona.edu/Neutralome/Neutralome.bed. GotCloud, https://genome.sph.umich.edu/wiki/GotCloud. Beagle 3.3.1, http://faculty.washington.edu/browning/beagle/b3.html. SAMtools, http://samtools.sourceforge.net/. Bcftools, https://samtools.github.io/bcftools/bcftools.html. HapMap3, https://www.sanger.ac.uk/resources/downloads/human/hapmap3.html. PLINK, http://zzz.bwh.harvard.edu/plink/simulate.shtml. mapdata: Extra Map Databases, https://CRAN.R-project.org/package=mapdata. ggplot2, https://CRAN.R-project.org/package=ggplot2. AdmixTools, https://github.com/DReichLab/AdmixTools. ALDER version 1.03, http://cb.csail.mit.edu/cb/alder/. OpenStreetMap, https://www.openstreetmap.org/#map=6/54.910/-3.432. SMC++, https://github.com/popgenmethods/smcpp. SHAPEIT2, http://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41588-018-0215-8.

Received: 6 December 2016; Accepted: 30 July 2018;
Published online: 17 September 2018

References

1. Lettre, G. & Hirschhorn, J. N. Small island, big genetic discoveries. *Nat. Genet.* **47**, 1224–1225 (2015).
2. Sidore, C. et al. Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nat. Genet.* **47**, 1272–1281 (2015).
3. Naitza, S. et al. A genome-wide association scan on the levels of markers of inflammation in Sardinians reveals associations that underpin its complex regulation. *PLoS Genet.* **8**, e1002480 (2012).
4. Zoledziewska, M. et al. Height-reducing variants and selection for short stature in Sardinia. *Nat. Genet.* **47**, 1352–1356 (2015).
5. Steri, M. et al. Overexpression of the cytokine BAFF and autoimmunity risk. *N. Engl. J. Med.* **376**, 1615–1626 (2017).
6. Cucca, F. et al. The distribution of DR4 haplotypes in Sardinia suggests a primary association of type I diabetes with DRB1 and DQB1 loci. *Hum. Immunol.* **43**, 301–308 (1995).
7. Marrosu, M. G. et al. The co-inheritance of type 1 diabetes and multiple sclerosis in Sardinia cannot be explained by genotype variation in the HLA region alone. *Hum. Mol. Genet.* **13**, 2919–2924 (2004).
8. Pugliatti, M. et al. The epidemiology of multiple sclerosis in Europe. *Eur. J. Neurol.* **13**, 700–722 (2006).
9. Cao, A. & Galanello, R. Beta-thalassemia. *Genet. Med.* **12**, 61–76 (2010).
10. Dyson, S. L., & Rowland, R. J. *Archaeology and History in Sardinia from the Stone Age to the Middle Ages: Shepherds, Sailors, & Conquerors* (University of Pennsylvania Museum of Archaeology and Anthropology: Philadelphia, PA, USA, 2007).
11. Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. *The History and Geography of Human Genes* (Princeton University Press: Princeton, NJ, USA, 1994).
12. Eaves, I. A. et al. The genetically isolated populations of Finland and Sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes. *Nat. Genet.* **25**, 320–323 (2000).
13. Calò, C. M., Melis, A., Vona, G. & Piras, I. S. Sardinian population (Italy): a genetic review. *Int. J. Mod. Anthropol.* **1**, 39–64 (2008).
14. Cavalli-Sforza, L. L. & Piazza, A. Human genomic diversity in Europe: a summary of recent research and prospects for the future. *Eur. J. Hum. Genet.* **1**, 3–18 (1993).
15. Barbujani, G. & Sokal, R. R. Zones of sharp genetic change in Europe are also linguistic boundaries. *Proc. Natl Acad. Sci. USA* **87**, 1816–1819 (1990).
16. Zavattari, P. et al. Major factors influencing linkage disequilibrium by analysis of different chromosome regions in distinct populations: demography, chromosome recombination frequency and selection. *Hum. Mol. Genet.* **9**, 2947–2957 (2000).
17. Elhaik, E. et al. Geographic population structure analysis of worldwide human populations infers their biogeographical origins. *Nat. Commun.* **5**, 3513 (2014).
18. Cann, H. M. Human genome diversity. *C. R. Acad. Sci. III, Sci. Vie* **321**, 443–446 (1998).
19. Haak, W. et al. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207–211 (2015).
20. Keller, A. et al. New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nat. Commun.* **3**, 698 (2012).
21. Lazaridis, I. et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409–413 (2014).
22. Li, J. Z. et al. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100–1104 (2008).
23. Skoglund, P. et al. Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. *Science* **336**, 466–469 (2012).
24. Allentoft, M. E. et al. Population genomics of Bronze Age Eurasia. *Nature* **522**, 167–172 (2015).
25. Hofmanová, Z. et al. Early farmers from across Europe directly descended from Neolithic Aegeans. *Proc. Natl Acad. Sci. USA* **113**, 6886–6891 (2016).
26. Mathieson, I. et al. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**, 499–503 (2015).
27. Sikora, M. et al. Population genomic analysis of ancient and modern genomes yields new insights into the genetic ancestry of the Tyrolean Iceman and the genetic structure of Europe. *PLoS Genet.* **10**, e1004353 (2014).

28. Ghirotto, S. et al. Inferring genealogical processes from patterns of Bronze-Age and modern DNA variation in Sardinia. *Mol. Biol. Evol.* **27**, 875–886 (2010).
29. Fraumene, C., Petretto, E., Angius, A. & Pirastu, M. Striking differentiation of sub-populations within a genetically homogeneous isolate (Ogliastra) in Sardinia as revealed by mtDNA analysis. *Hum. Genet.* **114**, 1–10 (2003).
30. Morelli, L. et al. Frequency distribution of mitochondrial DNA haplogroups in Corsica and Sardinia. *Hum. Biol.* **72**, 585–595 (2000).
31. Pala, M. et al. Mitochondrial haplogroup U5b3: a distant echo of the epipaleolithic in Italy and the legacy of the early Sardinians. *Am. J. Hum. Genet.* **84**, 814–821 (2009).
32. Olivieri, A. et al. Mitogenome diversity in Sardinians: a genetic window onto an island's past. *Mol. Biol. Evol.* **34**, 1230–1239 (2017).
33. Francalacci, P. et al. Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. *Science* **341**, 565–569 (2013).
34. Caramelli, D. et al. Genetic variation in prehistoric Sardinia. *Hum. Genet.* **122**, 327–336 (2007).
35. Vona, G. The peopling of Sardinia (Italy): history and effects. *Int. J. Anthropol.* **12**, 71–87 (1997).
36. Contu, D. et al. Y-chromosome based evidence for pre-neolithic origin of the genetically homogeneous but diverse Sardinian population: inference for association scans. *PLoS One* **3**, e1430 (2008).
37. Morelli, L. et al. A comparison of Y-chromosome variation in Sardinia and Anatolia is more consistent with cultural rather than demic diffusion of agriculture. *PLoS One* **5**, e10419 (2010).
38. Semino, O. et al. The genetic legacy of Paleolithic *Homo sapiens sapiens* in extant Europeans: a Y chromosome perspective. *Science* **290**, 1155–1159 (2000).
39. Rootsi, S. et al. Phylogeography of Y-chromosome haplogroup I reveals distinct domains of prehistoric gene flow in europe. *Am. J. Hum. Genet.* **75**, 128–137 (2004).
40. Chikhi, L., Nichols, R. A., Barbujani, G. & Beaumont, M. A. Y genetic data support the Neolithic demic diffusion model. *Proc. Natl Acad. Sci. USA* **99**, 11008–11013 (2002).
41. Passarino, G. et al. Y chromosome binary markers to study the high prevalence of males in Sardinian centenarians and the genetic structure of the Sardinian population. *Hum. Hered.* **52**, 136–139 (2001).
42. Olalde, I. et al. The Beaker phenomenon and the genomic transformation of northwest Europe. *Nature* **555**, 190–196 (2018).
43. Kivisild, T. The study of human Y chromosome variation through ancient DNA. *Hum. Genet.* **136**, 529–546 (2017).
44. Skoglund, P. et al. Genomic insights into the peopling of the Southwest Pacific. *Nature* **538**, 510–513 (2016).
45. Goldberg, A., Gunther, T., Rosenberg, N. A. & Jakobsson, M. Ancient X chromosomes reveal contrasting sex bias in Neolithic and Bronze Age Eurasian migrations. *Proc. Natl Acad. Sci. USA* **114**, 2657–2662 (2017).
46. Günther, T. et al. Ancient genomes link early farmers from Atapuerca in Spain to modern-day Basques. *Proc. Natl Acad. Sci. USA* **112**, 11917–11922 (2015).
47. Hellenthal, G. et al. A genetic atlas of human admixture history. *Science* **343**, 747–751 (2014).
48. Loh, P. R. et al. Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* **193**, 1233–1254 (2013).
49. Moorjani, P. et al. The history of African gene flow into Southern Europeans, Levantines, and Jews. *PLoS Genet.* **7**, e1001373 (2011).
50. Barbujani, G., Bertorelle, G., Capitani, G. & Scozzari, R. Geographical structuring in the mtDNA of Italians. *Proc. Natl Acad. Sci. USA* **92**, 9171–9175 (1995).
51. Pistis, G. et al. High differentiation among eight villages in a secluded area of Sardinia revealed by genome-wide high density SNPs analysis. *PLoS One* **4**, e4654 (2009).
52. Sanna, S. et al. Variants within the immunoregulatory *CBLB* gene are associated with multiple sclerosis. *Nat. Genet.* **42**, 495–497 (2010).
53. Zoledziewska, M. et al. Variation within the *CLEC16A* gene shows consistent disease association with both multiple sclerosis and type 1 diabetes in Sardinia. *Genes Immun.* **10**, 15–17 (2009).
54. Pilia, G. et al. Heritability of cardiovascular and personality traits in 6,148 Sardinians. *PLoS Genet.* **2**, e132 (2006).
55. Petkova, D., Novembre, J. & Stephens, M. Visualizing spatial population structure with estimated effective migration surfaces. *Nat. Genet.* **48**, 94–100 (2016).
56. Novembre, J. & Stephens, M. Interpreting principal component analyses of spatial population genetic variation. *Nat. Genet.* **40**, 646–649 (2008).
57. Botigué, L. R. et al. Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. *Proc. Natl Acad. Sci. USA* **110**, 11791–11796 (2013).
58. Henn, B. M. et al. Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genet.* **8**, e1002397 (2012).
59. Paschou, P. et al. Maritime route of colonization of Europe. *Proc. Natl Acad. Sci. USA* **111**, 9211–9216 (2014).
60. Terhorst, J., Kamm, J. A. & Song, Y. S. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat. Genet.* **49**, 303–309 (2017).
61. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919–925 (2014).
62. Raghavan, M. et al. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* **505**, 87–91 (2014).
63. Patterson, N. et al. Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
64. Blasco Ferrer, E. *Paleosardo: Le Radici Linguistiche Della Sardegna Neolitica* (De Gruyter, Berlin and New York, 2010).
65. Pickrell, J. K. et al. Ancient west Eurasian ancestry in southern and eastern Africa. *Proc. Natl Acad. Sci. USA* **111**, 2632–2637 (2014).
66. Heyer, E., Chaix, R., Pavard, S. & Austerlitz, F. Sex-specific demographic behaviours that shape human genomic variation. *Mol. Ecol.* **21**, 597–612 (2012).
67. Lazaridis, I. & Reich, D. Failure to replicate a genetic signal for sex bias in the steppe migration into central Europe. *Proc. Natl Acad. Sci. USA* **114**, E3873–E3874 (2017).
68. Goldberg, A., Günther, T., Rosenberg, N. A. & Jakobsson, M. Reply to Lazaridis and Reich: robust model-based inference of male-biased admixture during Bronze Age migration from the Pontic-Caspian Steppe. *Proc. Natl Acad. Sci. USA* **114**, E3875–E3877 (2017).
69. Wilkins, J. F. & Marlowe, F. W. Sex-biased migration in humans: what should we expect from genetic data? *Bioessays* **28**, 290–300 (2006).
70. Keinan, A. & Clark, A. G. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* **336**, 740–743 (2012).
71. Joshi, P. K. et al. Directional dominance on stature and cognition in diverse human populations. *Nature* **523**, 459–462 (2015).
72. Lohmueller, K. E. The impact of population demography and selection on the genetic architecture of complex traits. *PLoS Genet.* **10**, e1004379 (2014).
73. Simons, Y. B., Turchin, M. C., Pritchard, J. K. & Sella, G. The deleterious mutation load is insensitive to recent population history. *Nat. Genet.* **46**, 220–224 (2014).
74. Uricchio, L. H., Zaitlen, N. A., Ye, C. J., Witte, J. S. & Hernandez, R. D. Selection and explosive growth alter genetic architecture and hamper the detection of causal rare variants. *Genome Res.* **26**, 863–873 (2016).

Acknowledgements

The authors would like to thank Iosif Lazaridis, Pontus Skoglund, Nick Patterson, Sohini Ramanchandran, Alan Rogers, and Robert Brown for discussion and technical assistance, as well as members of the Novembre and Lohmueller labs for constructive comments regarding this research. This study was funded in part by the National Institutes of Health (NIH), including support via National Human Genome Research Institute grants HG005581, HG005552, HG006513, HG007022 to G.R.A., and HG007089 to J.N.; via National Heart, Lung, and Blood Institute grant HL117626 to G.R.A.; via National Institute of General Medical Sciences grant GM108805 to J.N., F32GM106656 to C.W.K.C., and T32GM007197 to J.H.M. and A.B.; via National Institute of Neurological Disorders and Stroke grant T32NS048004 to C.W.K.C.; by the Intramural Research Program of the NIH, National Institute on Aging, with contracts N01-AG-1-2109 and HHSN271201100005C to the Italian National Research Council (Consiglio Nazionale delle Ricerche); and by National Science Foundation fellowship DGE-1746045 to J.H.M. and H.A. This research was also supported by Sardinian Autonomous Region (L.R. no. 7/2009) grant cRP3-154, PB05 InterOmics MIUR Flagship Project, and grant FaReBio2011 (Farmaci e Reti Biotecnologiche di Qualità) to F.C.

Author contributions

F.C., G.R.A., D.S., and J.N. conceived of the study. C.W.K.C., C.S., D.S., F.C., G.R.A., and J.N. designed the study. C.W.K.C., J.H.M., C.S., H.A., and A.B. performed the analyses. C.W.K.C., J.H.M., C.S., A.B., K.E.L., G.R.A., D.S., F.C., and J.N. interpreted the data. C.S., M.Z., M.P., F.B., A.M., G.P., M.S., A.A., G.R.A., D.S., and F.C. contributed to data collection and the initial preparation for genetic analysis. C.W.K.C. and J.N. wrote the paper with input from all coauthors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0215-8.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to C.W.K.C. or J.N.

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

Methods

Cohort description. We included in this study individuals from the SardiNIA/Progenia longitudinal study of aging[2,54] based in the Ogliastra region and from the case-control studies of multiple sclerosis[52] and type 1 diabetes[53] across the general population of Sardinia. For the case-control study cohort, we required individuals to have at least three Sardinian grandparents. All participants gave informed consent, with protocols approved by the institutional review boards of the University of Cagliari, the National Institute on Aging, and the University of Michigan.

Whole-genome sequenced Sardinian dataset. The dataset includes 3,514 individuals sequenced at low coverage (average coverage 4.2×) and 131 individuals sequenced at high coverage (average coverage 36.7×); 2,090 individuals belong to the SardiNIA cohort, while the remaining 1,424 are derived from the case-control study. A subset of 2,120 low-coverage individuals was previously described[2]. The additional 1,394 individuals and the 131 high-coverage individuals were aligned, recalibrated, and quality-checked using the same criteria[2] to guarantee sample uniformity. Variant calling was performed using GotCloud as described by Sidore et al.[2] (see URLs), which also include a step of genotyping refinement with Beagle[75] (see URLs) to increase genotype accuracy in the low-coverage individuals. For the X chromosome, we first performed standard variant calling to generate genotype likelihoods (GLs) for each individual genotype. We then set the heterozygous call GL to a value of 500 among males and ran genotype refinement as described for autosomal markers. The most likely homozygous genotypes for males generated by Beagle were then converted to haploid genotypes.

To process the high-coverage data, we created a pileup of raw sequence reads using SAMtools version 0.2 ('samtools mpileup'; see URLs), filtering out bases with a base quality score < 20 and reads with a map quality score < 20. We then used the bcftools version 1.2 variant caller ('bcftools call') in conjunction with custom scripts (see URLs) to call SNPs.

Filtering individual samples by poor sequencing quality and relatedness. For each individual, we examined the proportion of imputed genotypes with the highest posterior genotype probability < 0.9 and removed eight outlier individuals with an excessive proportion (> 0.008), likely due to an overall low coverage of these samples.

To prune the dataset for related individuals, we first extracted a subset of 153.7 K SNPs with maximum pairwise r^2 of 0.2 (pruned from 1.21 M SNPs overlapping between the Sardinian WGS data and HapMap 3; see URLs) and then computed the genome-wide proportion of pairwise identity by descent (pihat) using PLINK version 1.08 (see URLs). The distribution of pihat showed distinct modes corresponding to different degrees of relatedness, as well as extensive low-level sharing (pihat < 0.1) between individuals, consistent with long-term isolation. We removed one individual from each pair of individuals with pihat > 0.07 to retain 1,577 approximately unrelated individuals, including 615 individuals from the SardiNIA sample and 962 individuals from the case-control cohort.

Defining sample origin based on self-reported grandparental ancestry. We assigned a four-part ancestral origin to each study participants based on the self-reported geographical birth locations of each of their parents and grandparents. We first categorized each location by three levels of resolutions: (1) macro-regions (for example, Sardinia, Southern Italy, France, Tunisia); (2) provinces within Sardinia (for example, Cagliari, Sassari, Ogliastra); and (3) town-level (for example, Arzana, Lanusei, Tortolì). For each parental lineage, we preferentially used grandparental origin, if available, or parental information to represent the ancestry from both grandparents if grandparental information was missing. Given the more detailed information provided by the participants of the SardiNIA project, we defined SardiNIA samples down to town resolution, but only defined the case-control samples down to province resolution unless noted otherwise. In the initial PCA and admixture analyses stratified by these labels, we found that genetic ancestry did not significantly differ for individuals having four or three parts of their ancestry coming from a particular location, nor did it differ by whether the four-part origin came from self-reported grandparental or self-reported parental origin (data not shown). However, we observed more heterogeneity among individuals with a two-part origin. Thus, unless noted otherwise, for any analysis where discrete geographical labeling is used we restricted to individuals having at least three out of the four-part origin from the same geographical location.

Merging with other datasets. To merge the Sardinia sequenced data with the Human Origin dataset[21], we repeated the variant calling pipeline in the Sardinian dataset specifically at the 600,841 variable sites released with the Human Origin dataset. Comparisons of this call set with the array genotypes on the same individuals suggest that the resulting genotype calls are of high quality (genotype discordant rate = 0.43 and 0.24% at heterozygous sites and all call sites, respectively). We then merged this call set with the Human Origins dataset across the autosome (594,924 SNPs), of which 95,853 are monomorphic in Sardinia. Some basic information and summary statistics of the reference panels and populations used in the merge can be found in Supplementary Table 7. Unless denoted specifically, the Human Origins merge is with the Lazaridis et al. data[21], which also provided aDNA samples for the Neolithic farmer (LBK380 or Stuttgart) and the pre-Neolithic hunter-gatherer Loschbour. For estimating mixture proportions in a three-way model of European admixture, we merged our dataset with the version of HOA data published by Haak et al.[19], which contains additional ancient samples (particularly additional early Neolithic farmers, LBK_EN, and post-Neolithic steppe pastoralists, Yamnaya) but fewer SNPs (354,212).

PCA. Sardinia-specific PCA was conducted using all unrelated individuals genotyped at SNPs found in HapMap 3 (Altshuler et al.[76]). For regional PCA, since a significant imbalance of sample sizes across populations may distort PCA, a random subset of ten unrelated Sardinians from Arzana and Cagliari were chosen to represent Sardinia and merged with the Human Origin dataset. Only populations from North Africa, the Middle East, Caucasus, and Europe from the HOA data were included. PCA analysis was performed using EIGENSTRAT version 5.0 after removing one SNP of each pair of SNPs with $r^2 \geq 0.8$ (in windows of 50 SNPs and steps of 5 SNPs) as well as SNPs in regions known to exhibit extended long-range LD (see Price et al.[77]).

Admixture analysis. Similar to the PCA analyses, Sardinia-specific admixture analysis was conducted using all unrelated individuals genotyped at SNPs found in HapMap 3. The regional analysis was conducted with subsampling of ten Sardinians, each with self-reported Cagliari and Arzana ancestries, merged with individuals from relevant populations from the HOA. Analysis was performed using ADMIXTURE version 1.22, following the recommended practice in the manual for LD filtering (removing one SNP of each pair of SNPs with $r^2 \geq 0.1$ in windows of 50 SNPs and steps of 5 SNPs). Ten independent unsupervised runs for $K = 2$–15 were performed, and for each value of K the run with maximum likelihood as estimated by the program was retained.

EEMS. The EEMS analysis was conducted using the same set of SNPs as the PCA. As EEMS requires fine-scale geographically indexed samples, we only used individuals whose four grandparents were all born in the same location at the town-level. This resulted in 181 individuals across the island for analysis in the Sardinia-only analysis. For the Mediterranean region analysis, the merged dataset with HOA was used. Because of the scale of the Mediterranean region, we only used the two Sardinian populations of Cagliari and Sassari, these being the two Sardinian populations with the largest sample sizes that are geographically sufficiently distant not to be merged by EEMS. The populations from HOA data used in this analysis are: Spanish (Castilla y León, Castilla-La Mancha, Extremadura, Cantabria, Cataluña, Valencia, Murcia, Andalucia, Baleares, Aragon, Galicia); Spanish_North; French_South; French; Bergamo; Italian_South; Tuscan; Sicilian; Mozabite; Algerian; Tunisian; and Spanish_Basque. We used the default settings for the EEMS hyperparameters. For each run, we ran a burn-in of 1 million iterations followed by an additional 1 million iterations with posterior samples taken every 1,000 iterations. We assessed the convergence of the Markov chain Monte Carlo by the posterior probability trace plot. We further assessed model fit by comparing the expected distance fitted by EEMS to the raw observed distances. Repeating the analysis with different combinations of grid sizes and random seeds produced qualitatively similar results. Results were displayed geographically using the 'mapdata' package and 'ggplot2' in R.

Admixture f_3 analyses and D-statistics. For admixture f_3 analyses, aimed at testing for evidence of admixture in a target population, we computed the f_3 statistics using all pairs of populations from Europe (including Turkey/Greece, Italian Peninsula, and Iberian Peninsula), Caucasus, Middle East, North Africa, and sub-Saharan Africa (Supplementary Table 7). For the outgroup f_3 analyses, aimed at estimating the amount of shared drift between a pair of populations, we computed the f_3 statistics between a Sardinian (Arzana or Cagliari) and another mainland population, while using the Mbuti individuals as the outgroup. Both f_3 and D-statistics were calculated using AdmixTools version 3.0 (see Patterson[63]). Statistical significance was assessed using the default blocked jackknife implementation in AdmixTools. Results were displayed geographically using the 'nps' set of maps available through the OpenStreetMap package in R (see URLs).

ALDER. We used the full set of HOA SNPs, except for SNPs lacking recombination map information (based on sex-averaged deCODE map[78]), or found in regions of long-range LD (see Price et al.[77]). For each Sardinian test population, we tested all pairwise combinations of mainland populations as in the f_3 admixture analysis using ALDER version 1.03 (see URLs).

In contrast to the approach taken by Loh et al.[48], we opted to be conservative with interpreting ALDER results where the two-reference LD decay fit is inconsistent with the one-reference LD decay fit from the program (that is, a 'successful fit' with warnings). In general, in successful fits where the two-reference decay curve and the one-reference decay curve agree with each other, the amplitude of the fit tends to be negatively correlated with the f_3 statistics of the same source/target population triplets, even if the f_3 statistics is positive (that is suggesting no evidence of admixture). However, when the decay curves under the two scenarios do not agree with each other, the correlation with the f_3 statistics is also poorer and/or becomes positive. These results suggest that complications

from the shared past demography between source and target populations could have influenced the LD decay curve fitting[48]. Thus, we reported only successful fits for up to 5 pairs of populations with the highest estimated amplitude among those with significant evidence of admixture ($P < 0.05$ after multiple testing correction by ALDER and Bonferroni correction for testing 15 Sardinian subpopulations), if available (Supplementary Table 3). The pairs of source populations with the highest estimated amplitude of LD decay are the populations closest to the true ancestral populations[48] among those available in our analyses. We thus estimated the admixture proportion using these pairs of source populations using the f_4-ratio estimator[63] (Supplementary Table 3).

Estimating the mixture proportions of ancient populations. Following Haak et al.[19], we estimated the mixture proportion with respect to the early European farmers (LBK_EN), western hunter-gatherers (Loschbour), and Yamnaya steppe pastoralists (Yamnaya) using the 'lsqlin' command in Matlab, based on a matrix of relationships between the test sample, the three ancient reference samples, and a set of 15 worldwide outgroups (Ami, BiAka, Bougainville, Chukchi, Eskimo, Han, Ju_hoan_North, Karitiana, Kharia, Mbuti, Onge, Papuan, She, Ulchi, and Yoruba)[19,21]. We assessed the uncertainty of these estimates with a blocked jackknife with 10 Mb blocks.

SMC++. Our main SMC++ analysis is based on 4 high-coverage unrelated individuals and 90 low-coverage unrelated individuals. The CEU and TSI high-coverage individuals were sequenced by Complete Genomics (see URLs) and variants were called using the same pipeline that was applied to the 131 high-coverage Sardinian samples described earlier and merged with 90 randomly selected individuals from the same population in 1000 Genomes. For Sardinia, we selected 2 high-coverage individuals each from Lanusei and Arzana (all with 4 grandparents from each village), and 50 and 40 low-coverage individuals each from Lanusei and Arzana, to match 90 samples in 1000 Genomes. Only biallelic sites were used. Moreover, we kept only regions where reads could be uniquely mapped (see URLs). For our supplementary analysis to explore finer-scale differentiation among Sardinian populations, we selected 2 high-coverage individuals each from CEU, TSI, Arzana, Lanusei, Ilbono, and Cagliari, merged with 40 low-coverage individuals from each population. The high-coverage Sardinian individuals all have four grandparents from the same population, with the exception of one Cagliari individual who has three grandparents from Cagliari and one grandparent with no information.

We used SMC++ version 1.9.3 to estimate population size trajectories and divergence time between populations. We used default parameters except that we set T1, the most recent time point for population size history inference, to 150 generations and the number of spline knots used to anchor the size history to 10. We had simulated a plausible European population growth model[79] and found that this combination of T1 and number of spline knots produced the best fitted population size trajectory from the simulated demography (data not shown). We evaluated the variability in the estimated size trajectory and the divergence times by resampling ten replicates of the genome in blocks of 10 Mb. Notably, in bootstrap samples we sometimes observed bimodally distributed estimates of divergence times. Therefore, we report the point estimates for divergence times from the whole dataset and the average of the ten bootstrap replicates, as reflected in Fig. 5a and Supplementary Fig. 5c. Importantly, the order of divergence time estimates among the pairs we examined remained unchanged. An alternative combination that could also recapitulate the simulated demography is T1 = 100 and spline knots = 12; however, we found that this combination of parameters produces less stable population size trajectories for the recent past based on the bootstrapping results. Following the practice in Terhorst et al.[60], we used a mutation rate of 1.25×10^{-8} per basepair per generation to scale time.

MSMC. We additionally phased the high-coverage variant call set from SMC++ analysis with SHAPEIT2 (version 2.r790; see URLs) using the 1000 Genomes Phase 3 reference panel (see URLs). Input files for MSMC were then generated using custom scripts from the MSMC github repository (see URLs). We used MSMC version 0.1.0 to estimate effective population size and cross-coalescent rates (CCRs). For effective population size inference, we used four individuals (eight phased haplotypes) from each of CEU, TSI, Arzana, and Lanusei; for CCR inference, we used pairs of two individuals from each population. We defined the estimated divergence time between a pair of populations as the first time point at which the CCR is at or above 0.5. The mutation rate used to scale time was 1.25×10^{-8} per basepair per generation.

Allele sharing. We computed the allele sharing ratios between pairs of populations as the probability that two randomly drawn carriers of the allele of a given minor allele frequency (MAF) are from different populations, normalized by the panmictic expectation[80,81]. Specifically, we defined x_i and x_j as the relative fraction of sample sizes and p_i and p_j as the frequency of the minor allele in populations i and j. Then, the probability that two randomly drawn carriers are from different populations is the probability of sampling two carriers from different populations over the total probability of sampling two carriers. In terms of the variables defined here, this is: $2x_ix_jp_ip_j / (x_i^2p_i^2 + x_j^2p_j^2 + 2x_ix_jp_ip_j)$. This quantity was normalized by the panmictic expectation, which is $2x_ix_j$.

X chromosome versus autosome analysis. Both autosomal and X-chromosome data were first filtered to retain only SNPs with MAFs > 0.02 in Europeans (1000 Genomes Europeans + 1,577 unrelated Sardinians). This left 6,740,788 SNPs on the autosome and 221,434 SNPs on the X chromosome. SNPs were then pruned by LD using 868 unrelated Sardinian females by removing one SNP of each pair of SNPs with $r^2 \geq 0.1$ (in windows of 50 SNPs and steps of 5 SNPs), leaving 433,704 autosomal SNPs and 18,918 X-chromosome SNPs. We ran ADMIXTURE with $K = 3$, using all of the unrelated Sardinians and TSI from 1000 Genomes. In general, TSI individuals form the first ancestry component, while Sardinians are distributed in two different components as shown in Fig. 1b. Then, for the autosome and X chromosome, we compared the distribution of the component showing the largest F_{ST} from the TSI-dominated component to evaluate the excess of the Sardinian-specific ancestry on the X chromosome. Significance was assessed by permuting individual ancestries 1 million times, as well as by bootstrapping individuals. Analysis was done on both males and females using ADMIXTURE version 1.3 (Shringarpure et al.[82]), and confirmed by rerunning the analysis in the subset of 868 unrelated Sardinian females and in 839 non-Arzana Sardinian females. We also compared chrX/chr7 only. SNPs were filtered similarly as stated earlier; in total we compared 26,164 SNPs on chr7 to 18,918 SNPs on the X chromosome.

To compare the heterozygosity of the X chromosome versus the autosome, we utilized the WGS data released by the Simons Genome Diversity Project (SGDP)[83] (see URLs; accessed in March 2015). We restricted our analysis to only the 21 female Europeans in the SGDP and only used data from the presumed neutral regions of the autosome and the X chromosome[84] consisting of 3,606 and 787 10 kb windows, respectively. Heterozygosity was computed as the number of non-missing heterozygous sites of an individual normalized by the total genomic span in the neutral region. We then compared the heterozygosity ratio across populations.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Allele frequency summary data analyzed in the study have been deposited with the European Genome-phenome Archive under accession number EGAS00001002212. The disaggregated individual-level sequence data for 2,105 samples (adult volunteers of the SardiNIA cohort longitudinal study) analyzed in this study are from Sidore et al.[2] and are available from the database of Genotypes and Phenotypes under project identifier phs000313.v4.p2. The remaining individual-level sequence data are from a case-control study of autoimmunity from across Sardinia, consent and local institutional review board approval having been obtained. These data are only available for sharing and collaborating on by request from the project leader, Francesco Cucca, Consiglio Nazionale delle Ricerche, Italy.

References

75. Browning, B. L. & Browning, S. R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **84**, 210–223 (2009).
76. Altshuler, D. M. et al. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
77. Price, A. L. et al. Long-range LD can confound genome scans in admixed populations. *Am. J. Hum. Genet.* **83**, 135–139 (2008).
78. Kong, A. et al. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**, 1099–1103 (2010).
79. Tennessen, J. A. et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
80. Gravel, S. et al. Demographic history and rare allele sharing among human populations. *Proc. Natl Acad. Sci. USA* **108**, 11983–11988 (2011).
81. Nelson, M. R. et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **337**, 100–104 (2012).
82. Shringarpure, S. S., Bustamante, C. D., Lange, K. & Alexander, D. H. Efficient analysis of large datasets and sex bias with ADMIXTURE. *BMC Bioinformatics* **17**, 218 (2016).
83. Mallick, S. et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
84. Woerner, A. E., Veeramah, K. R., Watkins, J. C., Hammer, M. F. & Novembre, J. The role of phylogenetically conserved elements in shaping patterns of human genomic diversity. *Mol. Biol. Evol.* **35**, 2284–2295 (2018).

nature research

Corresponding author(s): Charleston W.K. Chiang, John Novembre

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	The data generation pipeline and associated softwares used were detailed in a previous publication (Sidore et al. Nature Genetics 2015). Any additional data used in this publication were generated in the same pipeline to ensure data uniformity. For variant calling of the high coverage individuals, we used samtools (v0.2) and bcftools (v1.2).
Data analysis	Softwares used for data analysis include PLINK (v1.08), EIGENSTRAT (v.5), ADMIXTURE (v1.22 and v1.3), EEMS (v.0.0.0.9), Admixtools (v3.0), ALDER (v1.03), SMC++ (v1.9.3), and MSMC (v0.1.0).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Allele frequency summary data analyzed in the study will be deposited to EGA under accession number EGAS00001002212. The disaggregated individual-level sequence data for 1887 samples - from adult volunteers of the SardiNIA cohort longitudinal study - analysed in this study are from Sidore et al (2015) and are available from dbGAP under project identifier phs000313.v3.p2. The remaining individual-level sequence data are from a case-control study of autoimmunity from across Sardinia, and per the obtained consent and local IRB, these data are only available for collaboration by request from the project leader (Francesco Cucca, Consiglio Nazionale delle Ricerche, Italy).

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We used all available data and the maximally unrelated subsets as detailed in the Methods for population genetic analysis
Data exclusions	Individuals appearing as the outliers in quality of imputed genotypes upon visual inspection were excluded (8 out of 3,514). Genetically identified related individuals were also excluded.
Replication	General findings are consistent with results from limited Sardinian samples from external reference datasets such as HGDP or SGDP.
Randomization	For subset analyses, a randomly selected subsets are used. In EEMS analysis, we specifically selected individuals with all four grandparents born in the same location to reduce impact of recent motility in human populations (see Methods).
Blinding	The investigators were not blinded. There is no testing of an intervention vs. placebo group, and thus no blinding is necessary.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	We included in this study all individuals from the SardiNIA/Progenia longitudinal study of aging based in the Ogliastra region and from the case-control studies of Multiple Sclerosis and Type 1 Diabetes across the general population of Sardinia. For the SardiNIA study, over 6000 individuals older than 13 years of age were recruited from four villages in Lanusei, Sardinia. For the case-control study cohort, we required individuals to have at least three Sardinian grandparents.
Recruitment	Details of the sample recruitments are described in prior publications (Sidore et al. Nature Genetics 2015, Sanna et al. Nat.

Recruitment	Genet. 2010, Zoledziewska et al. Genes Immun. 2013, and cited in the current report). As this is a population genetic analysis, any bias in recruitment, if present, is unlikely to impact the estimates of allele frequencies genome-wide.