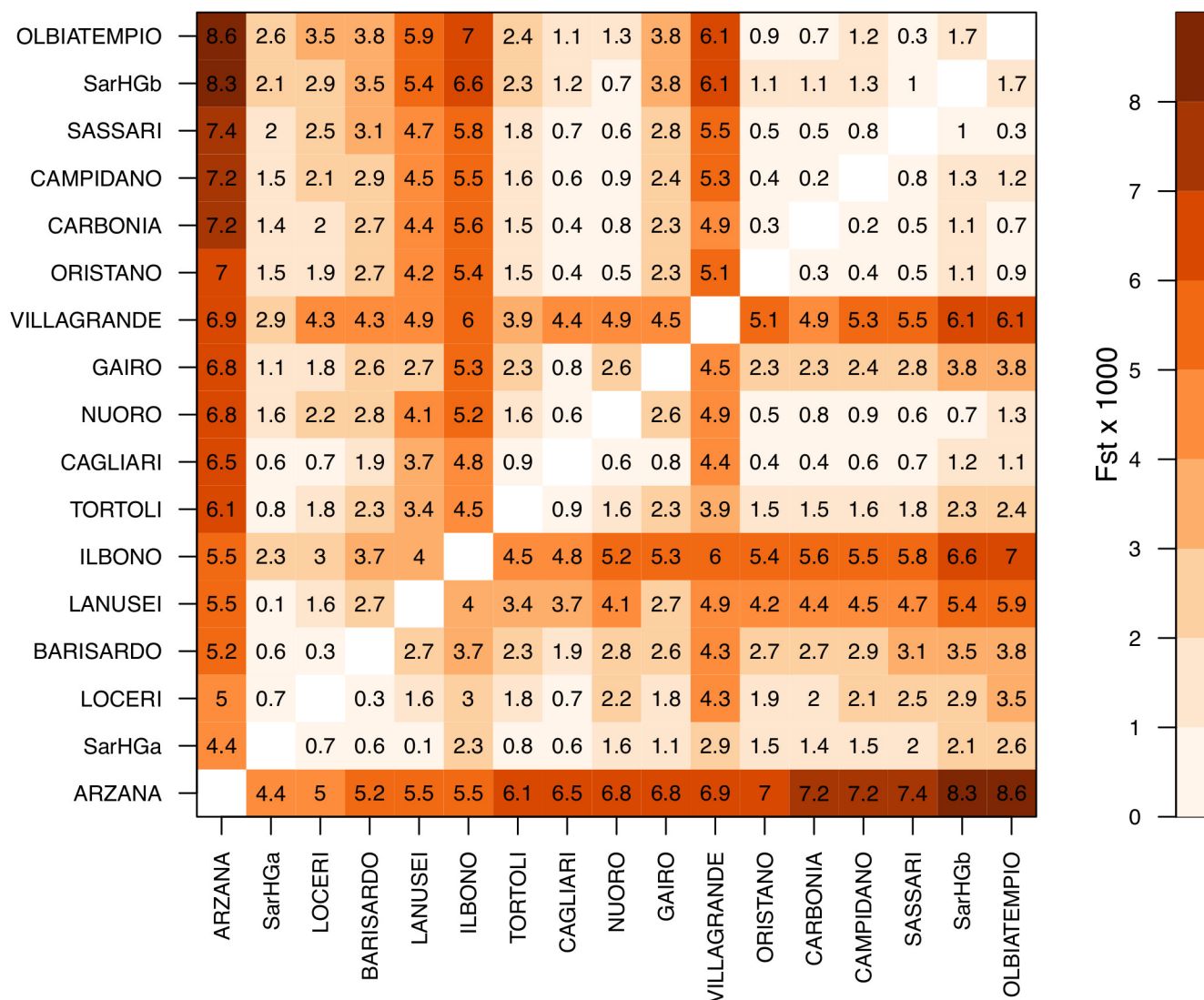


In the format provided by the authors and unedited.

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-Asadi⁷, **Magdalena Zoledziowska**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. Abecasis⁶, **David Schlessinger**⁸, **Francesco Cucca** ^{5,9} and **John Novembre**^{4*}

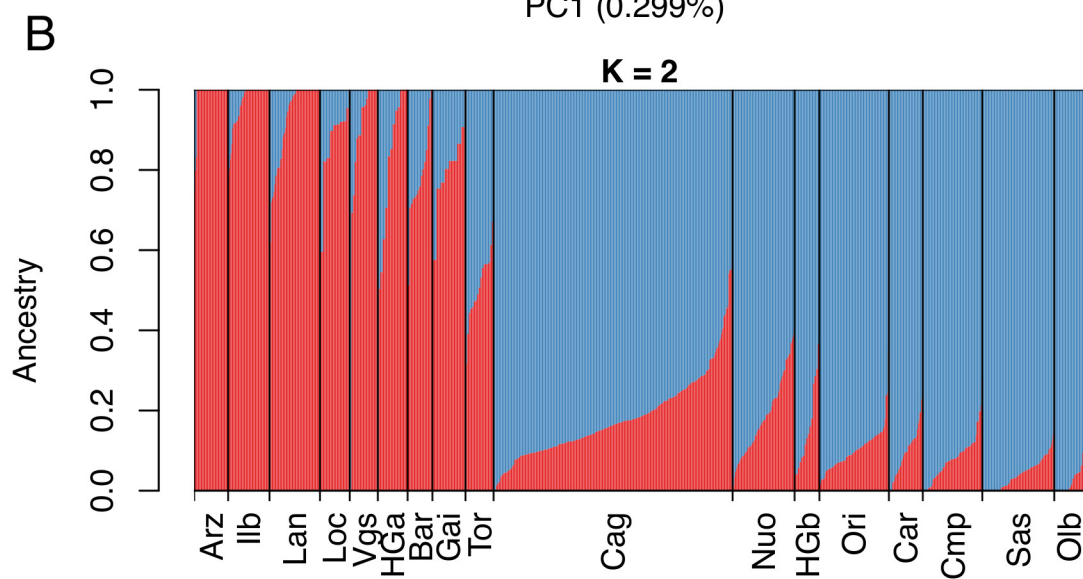
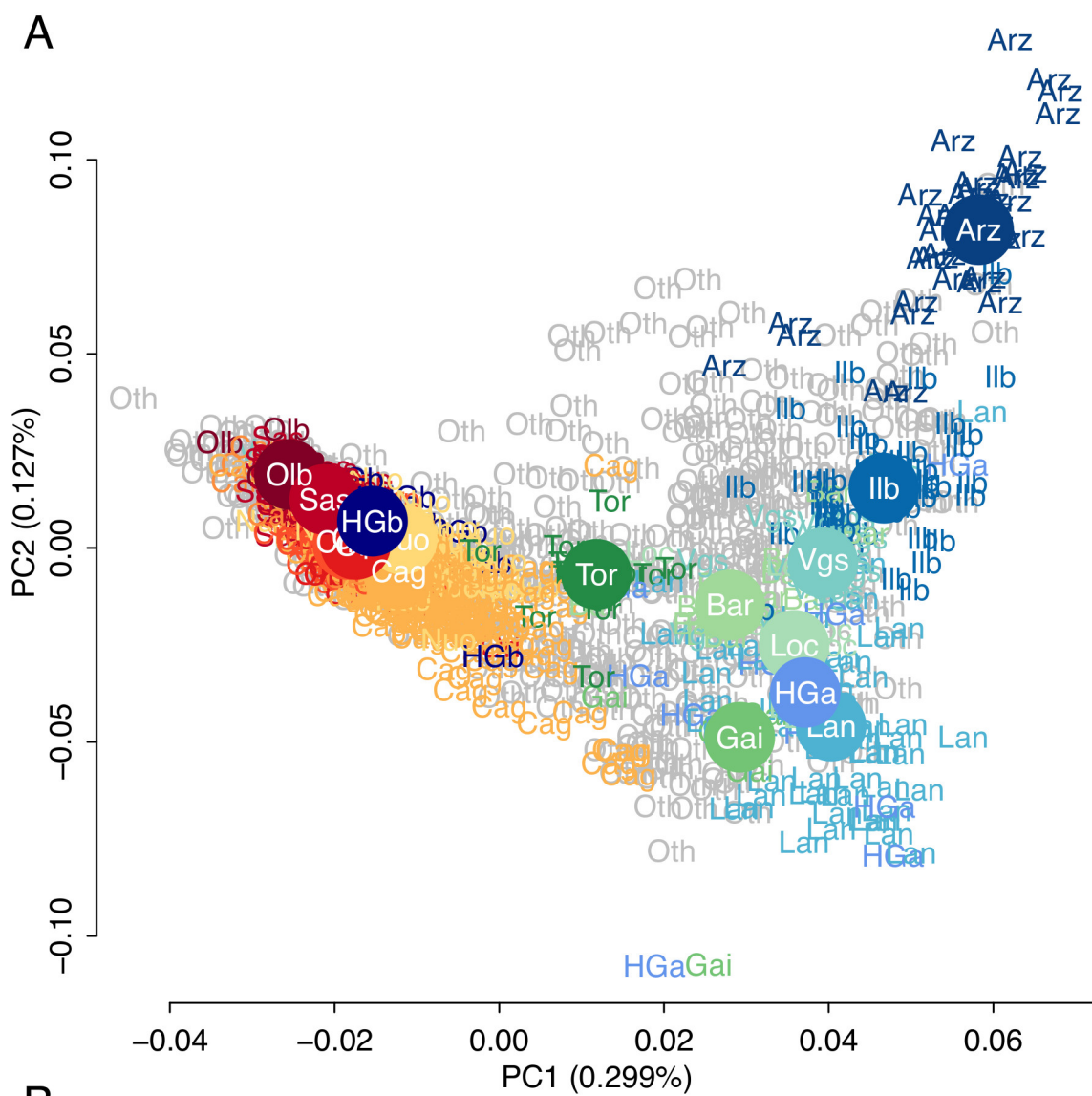
¹Center for Genetic Epidemiology, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA. ²Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Behavior, University of California, Los Angeles, Los Angeles, CA, USA. ³Department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



Supplementary Figure 1

Fst matrix within Sardinia, including the HGDP Sardinians.

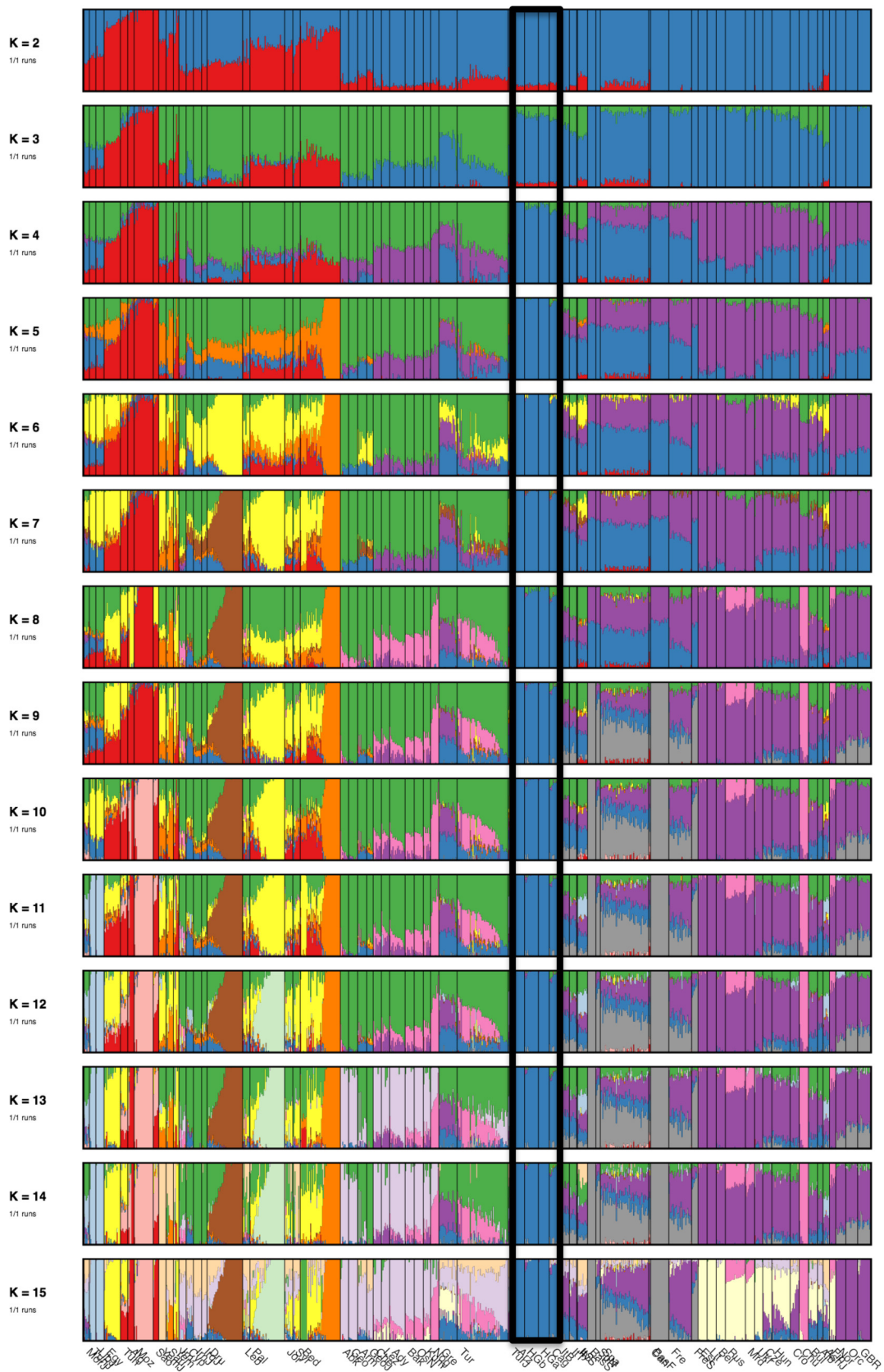
The HGDP Sardinians are recorded as being collected from the Gennargentu region (A. Piazza, personal communication). Consistent with this record, the HGDP Sardinians show close affinity to individuals from Ogliastra, although roughly half of the samples are more similar to the broader sample outside of Ogliastra. Thus, we labeled the two subgroups of HGDP as “SarHGa” and “SarHGb”.



Supplementary Figure 2

Relationship between HGDP Sardinians and newly sequenced Sardinians.

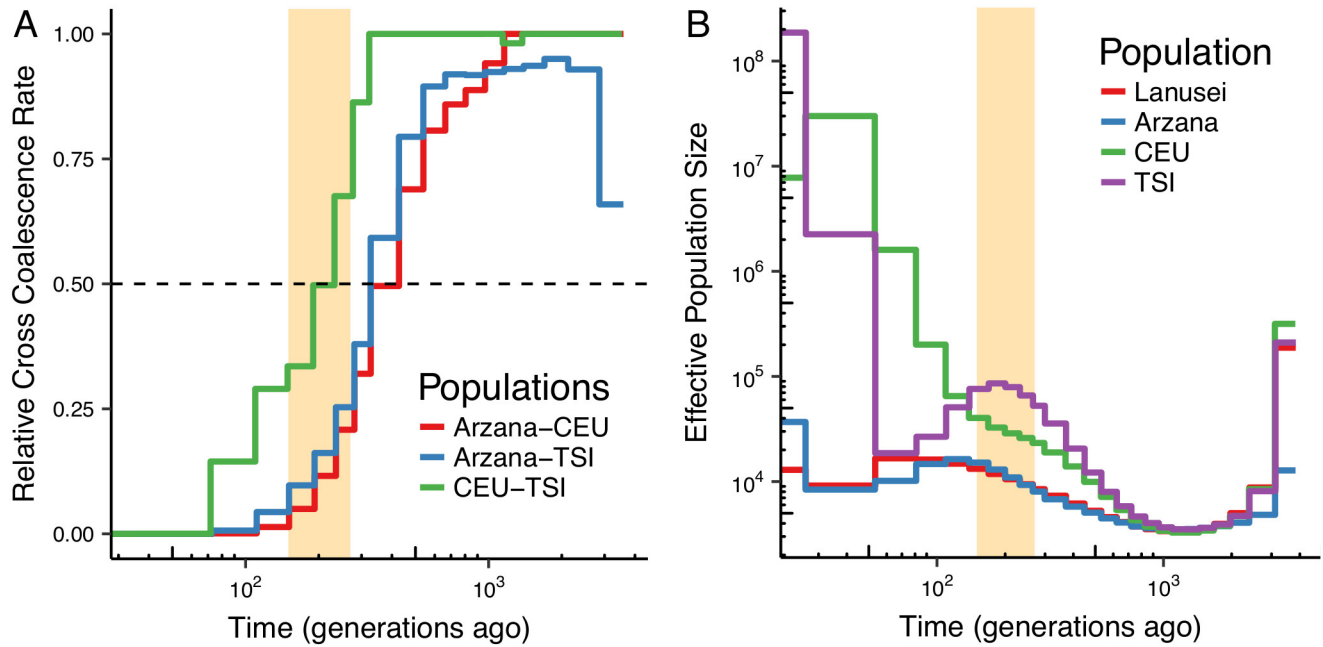
a. PCA results of merged Sardinian whole-genome sequences and the HGDP Sardinians. **b.** Admixture results of the merged dataset. The HGDP Sardinians are recorded as being collected from the Gennargentu region (A. Piazza, personal communication). Consistent with this record, the HGDP Sardinians show close affinity to individuals from Ogliastra, although roughly half of the samples are more similar to the broader sample outside of Ogliastra. Thus, we labeled the two subgroups of HGDP as “HG_a” and “HG_b”.



Supplementary Figure 3

Admixture results using the merged dataset of Sardinia and Human Origins Array data, for $K = 2$ to $K = 15$.

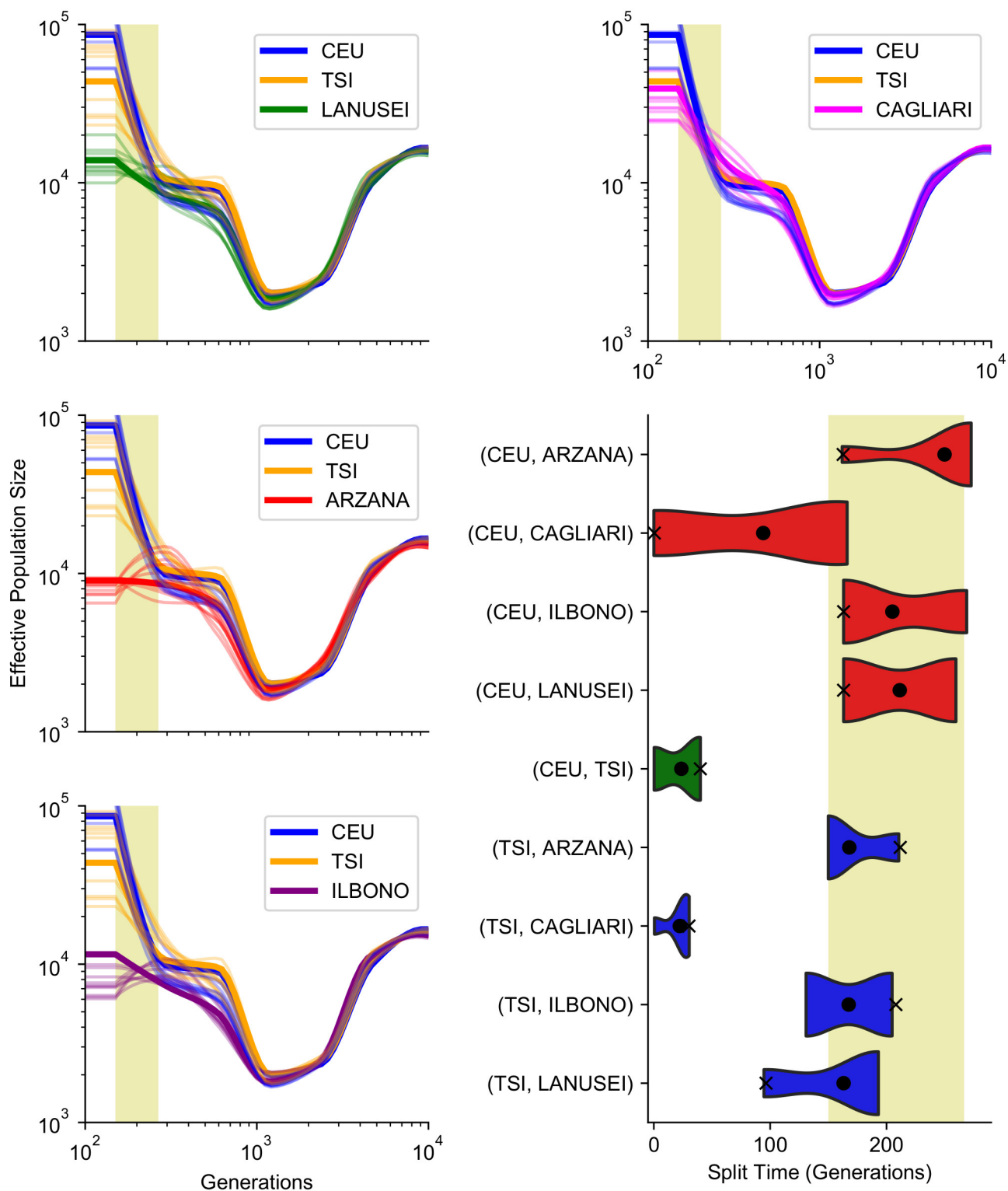
For $K = 4$ to $K = 15$, Sardinians (HGDP Sardinians, Arzana, and Cagliari, grouped by the black box) all cluster together with nearly 100% of a unique component of ancestry that is not found at high levels outside of Sardinia, consistent with their relative isolation and drift. Visualization across all K values was generated using Pong (v. 1.4.5; *Bioinformatics* **32**, 2817–2823, 2016).



Supplementary Figure 4

Coalescent-based inference of demographic history using MSMC.

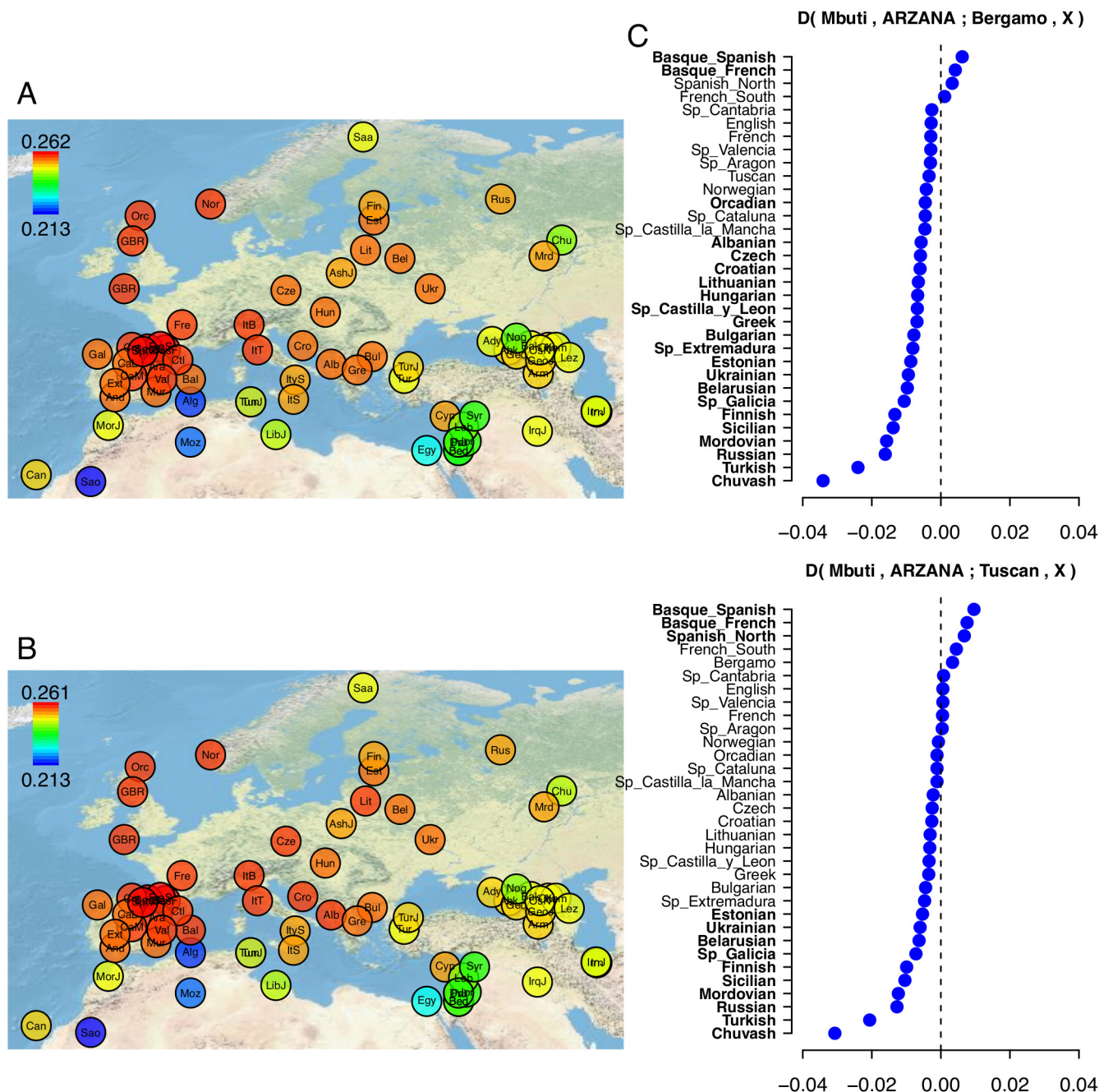
a, Inferred relative cross-coalescence rate between pairs of populations through time, based on four haplotypes each from Arzana, CEU, and TSI. A cross-coalescence rate of 0.5 is arbitrarily defined as the divergence time, which differs in definition from divergence time estimated by SMC++. Thus, one should be cautious with direct numerical comparisons of divergence time between the two methods. **b**, Population size history inference based on eight haplotypes of high-coverage individuals from each of Lanusei (LAN), Arzana (ARZ), 1000 Genomes CEU and TSI. The shaded box denotes approximately the Neolithic period, around 4,500 to 8,000 years ago, converted to units of generations assuming 30 years per generation.



Supplementary Figure 5

Inference of demographic history using SMC++ for each Sardinian subpopulation.

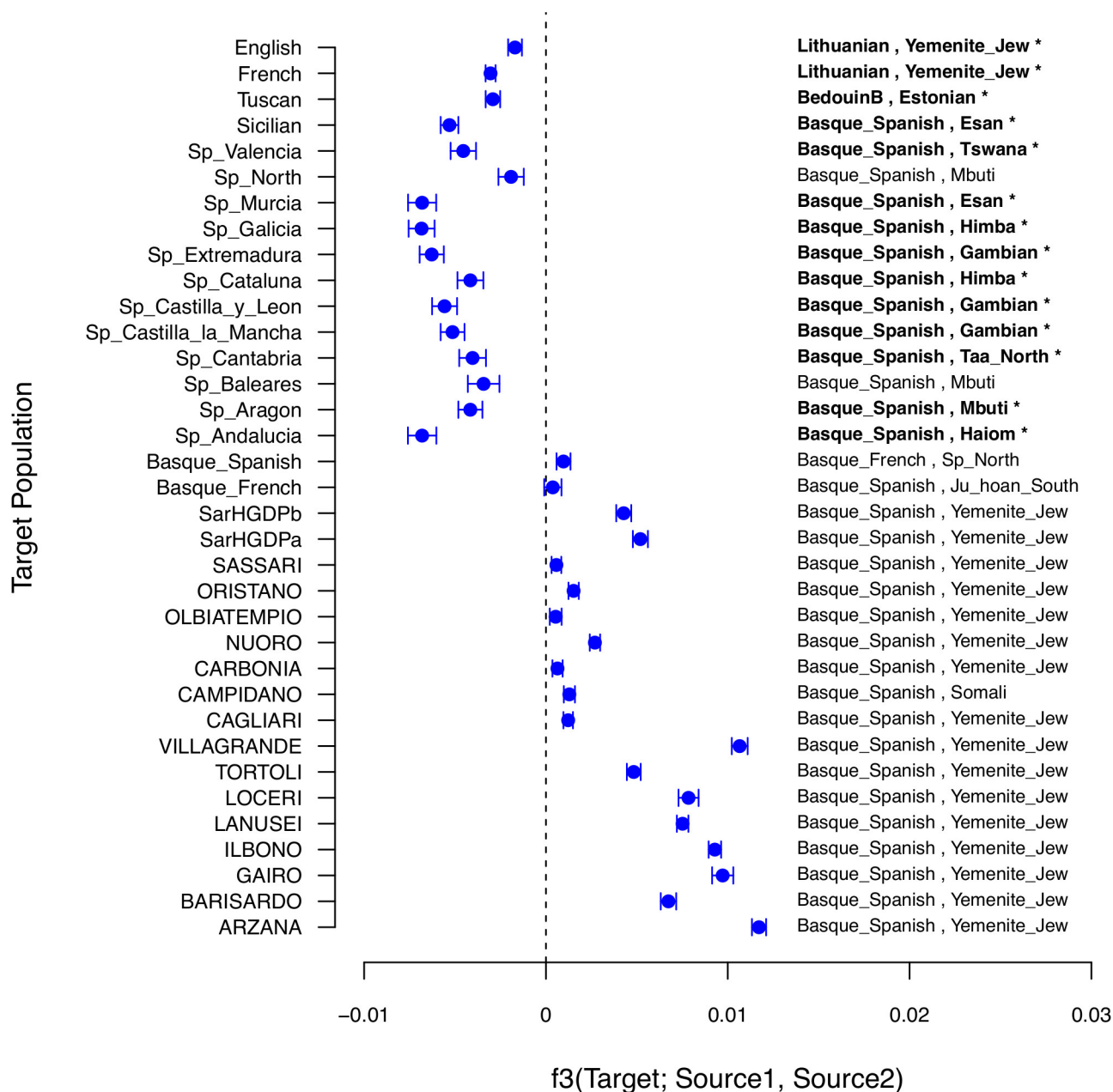
a,b, Population size history. For each population, we used 2 high-coverage individuals and 40 low-coverage individuals for analysis. The parameters used for SMC++ were $t_1 = 150$ and knots = 10. We found that all populations from Ogliastro (ARZANA, LANUSEI, and ILBONO) displayed lower effective population sizes in the recent past than mainland European populations (CEU, TSI) (**a**) or the population from Cagliari (**b**). Uncertainty in the estimates is reflected through ten bootstrap samples, shown in the same but lighter color to the trajectory estimated using the entire dataset. The shaded box denotes approximately the Neolithic period, around 4,500 to 8,000 years ago, converted to units of generations assuming 30 years per generation. **c**, Population divergence time estimates. Each black cross denotes the point estimates output by SMC++; uncertainty and mean point estimates from ten bootstrap samples are shown by the violin plot and the black dot, respectively.



Supplementary Figure 6

Relationship between Sardinia and mainland populations.

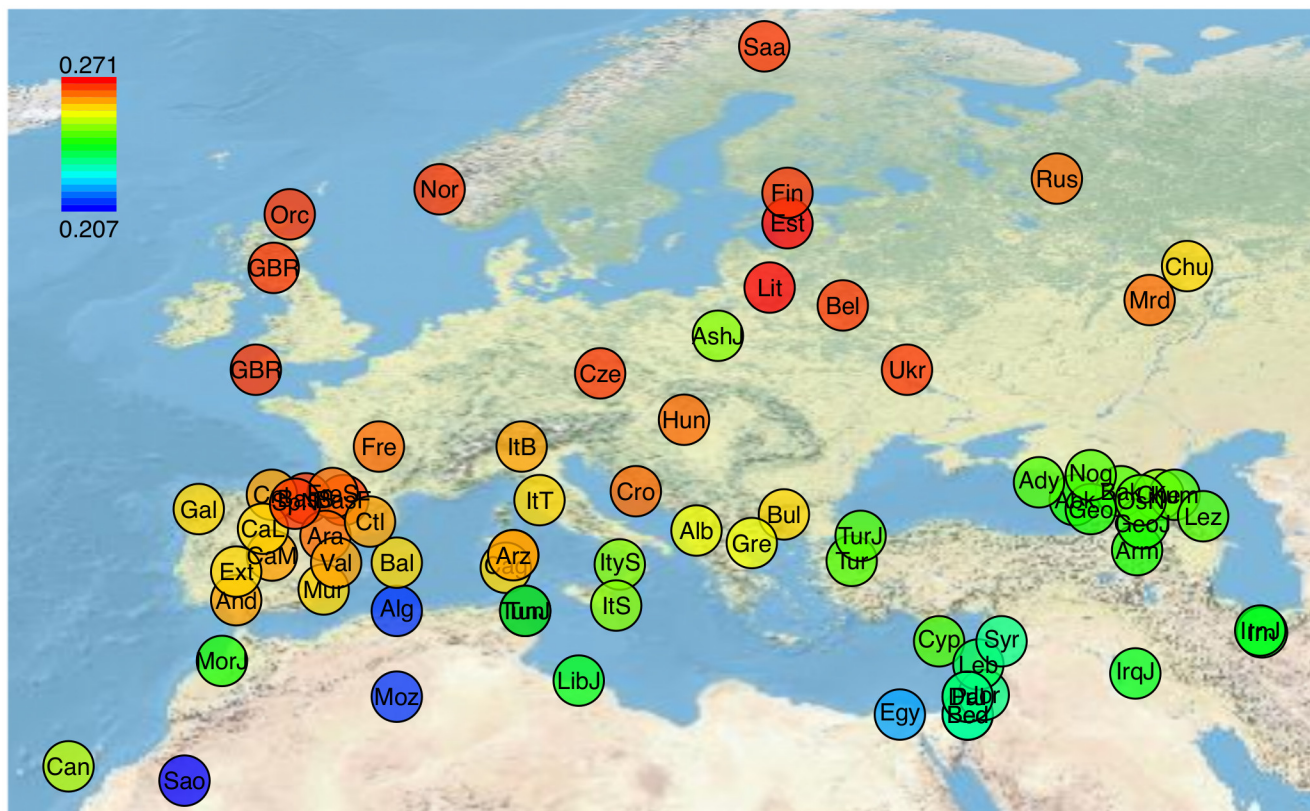
a. Outgroup f_3 results of the form $f_3(\text{Mbuti}; \text{Arzana}, X)$, where X is a mainland population from Human Origins Array data. Outside of Sardinia, Sardinians show the highest amount of shared drift with the Basque. **b.** Similar display of results for $f_3(\text{Mbuti}; \text{Cagliari}, X)$. **c.** To formally test for excess sharing between Sardinians and Basque, we computed D statistics of the form $D(\text{Mbuti}, \text{Arzana}; \text{Tuscan or Bergamo}, X)$, where X is a mainland European population. Relative to Arzana–Tuscan or Arzana–Bergamo sharing, excess of sharing between Arzana and X would result in significant positive values of this D statistic; dearth of sharing between Arzana and X would result in significant negative values. Mainland populations with significant results ($|Z| > 4$) in this analysis are bolded on the y axis.



Supplementary Figure 7

Analysis of admixture signal using f_3 statistics.

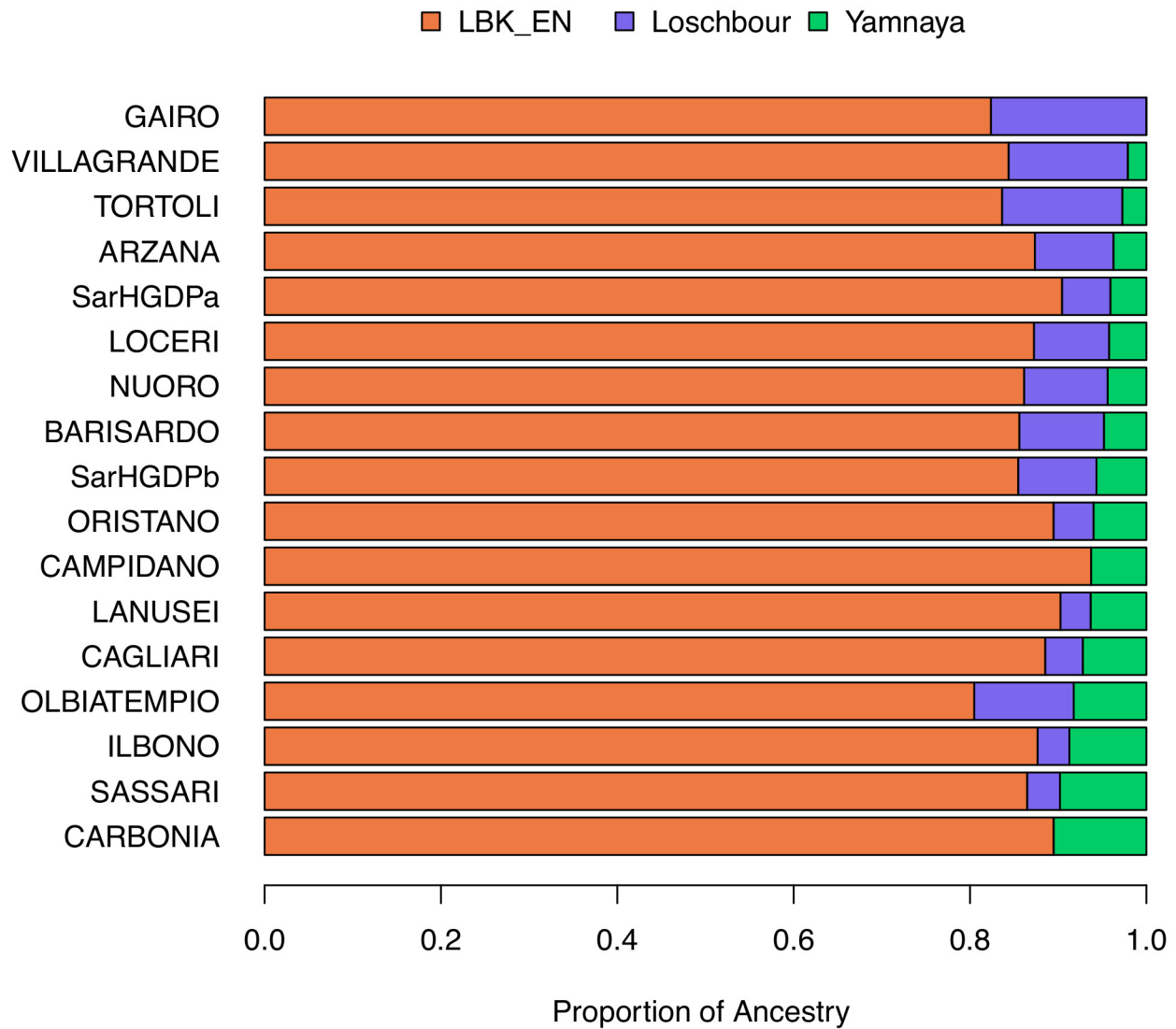
For each (target) Sardinian populations with sample size greater than eight and selected European populations from Human Origins Array data, we computed f_3 of the form $f_3(\text{target}; \text{source 1}, \text{source 2})$, where source 1 and source 2 are all possible pairs of populations in Human Origins Array data. Then, for each target population, we display the pair of source populations that produced the lowest (not necessarily negative) f_3 value. Significantly negative f_3 statistics, bolded and marked with an asterisk, provide evidence for admixture, which we observed for the mainland European populations of English, French, Tuscan, Sicilian, and Spanish from most regions of Spain (Z range from -4.5 to -11.1). Error bars represent the standard error of the estimated f_3 based on the block jackknife procedure. None of the Sardinian or Basque populations showed evidence of admixture by this test.



Supplementary Figure 8

Relationship of ancient pre-Neolithic hunter-gatherers across Europe.

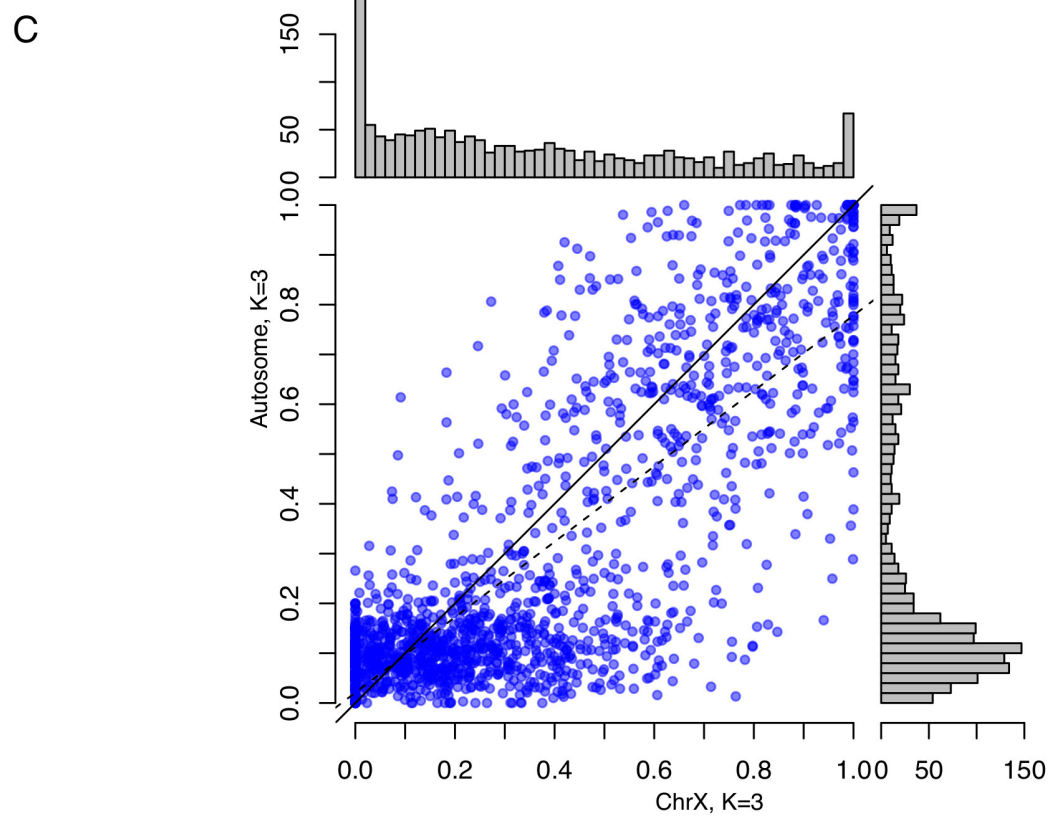
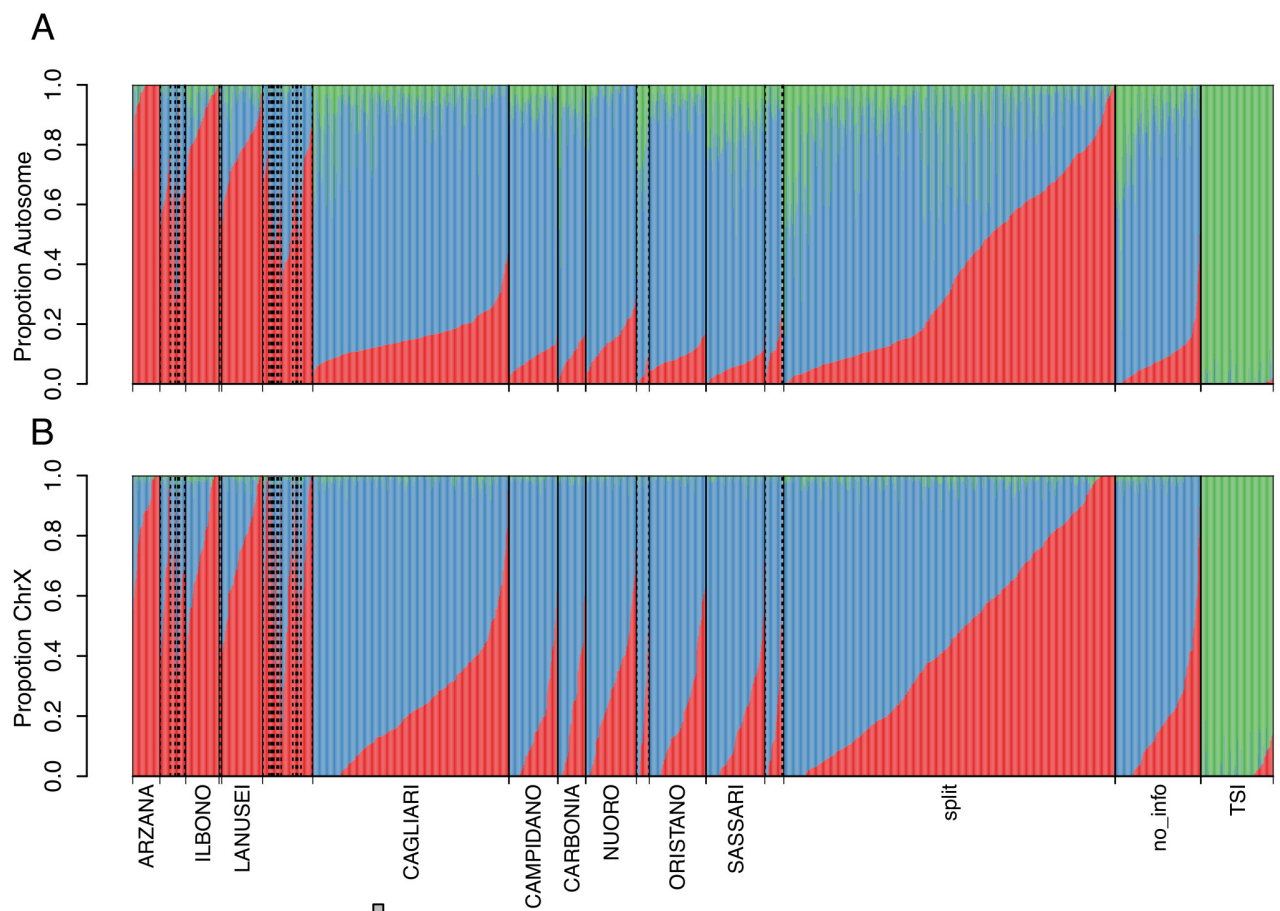
The map visualizes outgroup f_3 statistics of the form $f_3(\text{Mbuti}; \text{Loschbour}, X)$, where X is a population across the merged dataset of Sardinia and Human Origins Array data. Population abbreviations are the same as in Fig. 4.



Supplementary Figure 9

Mixture proportions of the three-component ancestries among Sardinian populations.

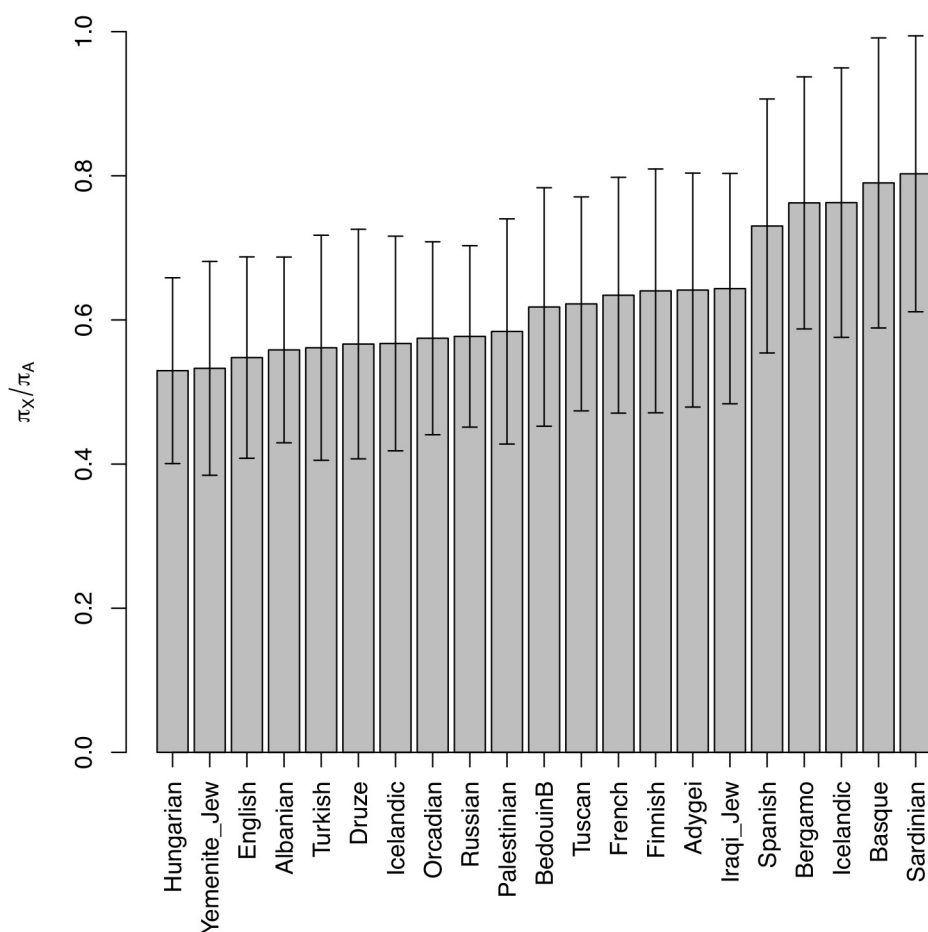
Using a method first presented in Haak et al. (*Nature* **522**, 207–211, 2015), we computed unbiased estimates of mixture proportions without a parameterized model of relationships between the test populations and the outgroup populations based on f_4 statistics. The three-component ancestries were represented by early Neolithic individuals from the LBK culture (LBK_EN), pre-Neolithic hunter-gatherers (Loschbour), and Bronze Age steppe pastoralists (Yamnaya). See Supplementary Table 5 for standard error estimates computed using a block jackknife.



Supplementary Figure 10

Admixture analysis contrasting chromosome X and the autosome based on 1,577 unrelated Sardinians and TSI individuals at $K = 3$.

a, Results using the autosomal data. **b**, Results using chromosome X data. **c**, Correlation of the 'red' ancestral component from **a** for each individual between chromosome X (x axis) and the autosome (y axis). The solid black line is the $y = x$ line; the dashed black line is the regression trend line. Overall, there is enrichment of the red ancestry component (found most prominently among Sardinians from Ogliastra) on the X chromosome as compared to the autosomes. We observed qualitatively similar results whether we used only the females in the dataset, removed the Arzana individuals, or compared chromosome X estimates only to those on chromosome 7, which is the most closely matched autosome in terms of sequenced base pairs and gene density in the human reference genome.



Supplementary Figure 11

Ratio of heterozygosity estimates between chromosome X and the autosome.

For each of the 21 deeply sequenced female Europeans found in SGDP, we computed the ratio of heterozygosity on the X chromosome versus the autosome. The analysis is restricted to the 3,606 and 787 10-kb windows believed to be neutrally evolving (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)) on the autosome and X chromosome, respectively. Error bars represent standard errors estimated from block jackknife procedures in 10-Mb blocks.

ID	HGD ID	Parental Origin		Grandparental Origin (Paternal)		Grandparental Origin (Maternal)		Origin Call	Seq?
		F	M	GF	GM	GF	GM		
171	1062	Tortoli	Tortoli	Tortoli	Loceri	Tortoli	Tortoli	Tortoli	Yes
832	1073	Lanusei	Ilbono	Lanusei	Lanusei	Lanusei	Ilbono	Lanusei	Yes
833	1074	Loceri	Seui	Loceri	Loceri	Loceri	Loceri	Loceri	Yes
2306	1076	Lanusei	Lanusei	-	Lanusei	Lanusei	Lanusei	Lanusei	No
2483	1072	Lanusei	Lanusei	Lanusei	Lanusei	Lanusei	Lanusei	Lanusei	No
4676	1064	Ilbono	Ilbono	Ilbono	Ilbono	Ilbono	Loceri	Ilbono	Yes
4841	1066	Lanusei	Gairo	Lanusei	Lanusei	Gairo	Gairo	NA	Yes
4842	1065	Ulassai	Ulassai	Perdas- defogu	Ulassai	Ulassai	Ulassai	Ulassai	Yes

Supplementary Table 1: parental and grandparental origin of duplicate samples between Sardinia whole genome sequencing data and HGD Sardinians. Of the 27 HGD Sardinians on the Human Origin Array dataset, 8 were included in our sample. For seven of these we had grandparental information that imply the samples derive from the villages of Tortoli, Lanusei, Loceri, Ilbono, and Ulassai in the Ogliastra province. These samples were likely shared with the HGD by early researchers collecting for the Lanusei/SardiNIA project. F, M, GF, GM represents father, mother, grandfather, and grandmother, respectively. Origin call list the assigned sample origin based on parental and grandparental birthplace information (**Methods**). Seq? shows whether the individual was included in the whole-genome sequenced dataset reported here.

Cluster A (N = 12)		Cluster B (N = 15)	
HGDP01062	HGDP01073	HGDP00666	HGDP00674
HGDP01063	HGDP01074	HGDP00667	HGDP01068
HGDP01064	HGDP01075	HGDP00668	HGDP01069
HGDP01065	HGDP01076	HGDP00669	HGDP01070
HGDP01066		HGDP00670	HGDP01077
HGDP01067		HGDP00671	HGDP01078
HGDP01071		HGDP00672	HGDP01079
HGDP01072		HGDP00673	

Supplementary Table 2: list of HGDP Sardinians divided into two clusters based on joint PCA analysis with the sequenced Sardinians (**Supplementary Figure 2**). We named the two sub-clusters Cluster A (SarHGDPa) and Cluster B (SarHGDPb). Cluster A HGDP Sardinians are more closely resembling those from the Gennargentu region in our dataset.

Test Population	Source 1	Source 2	Admixture Date (gen)	Fitted amplitude ($\times 10^{-5}$)	P-value	α_1 (s.e.)	α_2 (s.e.)
CAGLIARI	Wambo	Spanish (Castilla y Leon)	62.57 +/- 6.61	2.70 +/- 0.310	1.28×10^{-13}	0.0039 (0.0036)	0.0094 (0.0044)
CAGLIARI	Esan	Spanish (Castilla y Leon)	59.06 +/- 5.98	2.56 +/- 0.289	3.75×10^{-14}	0.0040 (0.0037)	0.0097 (0.0046)
CAGLIARI	Luhya	Bergamo	63.38 +/- 7.27	2.47 +/- 0.278	1.20×10^{-13}	0.013 (0.0025)	0.019 (0.0034)
CAGLIARI	Esan	Bergamo	55.25 +/- 5.21	2.46 +/- 0.250	2.25×10^{-18}	0.029 (0.0026)	0.029 (0.0034)
CAGLIARI	Wambo	Orcadian	59.12 +/- 8.09	2.46 +/- 0.327	1.13×10^{-8}	0.013 (0.0024)	0.019 (0.0032)
CAMPIDANO	Luhya	Tuscan	63.56 +/- 12.80	2.59 +/- 0.506	0.0285	0.0086 (0.0030)	0.018 (0.0039)
ORISTANO	Wambo	Estonian	101.05 +/- 20.08	3.51 +/- 0.638	0.0195	0.049 (0.0029)	0.052 (0.0040)
ORISTANO	BantuSA	Hungarian	88.73 +/- 12.89	3.33 +/- 0.534	1.80×10^{-5}	0.021 (0.0024)	0.027 (0.0030)
ORISTANO	Mbuti	Hungarian	88.23 +/- 14.99	3.17 +/- 0.597	0.0045	0.018 (0.0020)	0.022 (0.0024)
ORISTANO	BantuSA	Estonian	87.92 +/- 15.46	3.11 +/- 0.562	0.001275	0.048 (0.0028)	0.050 (0.0038)
ORISTANO	Gana	Lithuanian	84.84 +/- 14.18	3.08 +/- 0.629	0.039	0.037 (0.0026)	0.031 (0.0032)
SASSARI	Khwe	Lithuanian	82.21 +/- 15.92	2.53 +/- 0.430	2.25×10^{-8}	0.039 (0.0027)	0.033 (0.0035)
SASSARI	Tshwa	Lithuanian	65.72 +/- 12.37	2.48 +/- 0.389	0.0039	0.036 (0.0025)	0.030 (0.0032)
SASSARI	BantuSA	Lithuanian	67.81 +/- 9.62	2.42 +/- 0.330	0.00042	0.041 (0.0028)	0.035 (0.0037)
SASSARI	Mbuti	Lithuanian	69.74 +/- 13.83	2.42 +/- 0.367	0.045	0.035 (0.0024)	0.029 (0.0031)
SASSARI	Hoan	Lithuanian	72.20 +/- 14.72	2.40 +/- 0.336	0.042	0.034 (0.0024)	0.028 (0.0030)

Supplementary Table 3: Admixture proportions estimated by f4 ratio test. α_i given are the estimated Subsaharan admixture proportion with respect to different outgroups. For α_1 outgroups used were Finnish and Chimp; For α_2 outgroups used were Han and Chimp.

population	D(Mbuti, Sardinia, Stuttgart, Loschbour)	Z score
SarHGDPa	-0.0143	-3.337
CAMPIDANO	-0.013	-3.255
VILLAGRANDE	-0.0122	-2.894
SarHGDPb	-0.0122	-2.85
ARZANA	-0.0117	-2.788
CAGLIARI	-0.0114	-2.871
LOCERI	-0.0114	-2.531
CARBONIA	-0.0113	-2.833
ORISTANO	-0.0112	-2.788
SASSARI	-0.011	-2.755
BARISARDO	-0.011	-2.607
TORTOLI	-0.0107	-2.5
LANUSEI	-0.0102	-2.486
ILBONO	-0.0102	-2.452
NUORO	-0.0097	-2.38
OLBIATEMPIO	-0.0086	-2.07
GAIRO	-0.0078	-1.842

Supplementary Table 4: D-statistics demonstrating closer affinity of Neolithic European farmer (Stuttgart) than pre-Neolithic hunter-gatherer (Loschbour) to each Sardinian population. Note that the Basal European component in Stuttgart would tend to push the D-statistics towards positive values.

Population	resnorm	LBK_EN (s.e.)	Loschbour (s.e.)	Yamnaya (s.e.)
GAIRO	0.0000288	0.824 (0.053)	0.176 (0.053)	0 (0)
VILLAGRANDE	0.0000271	0.844 (0.054)	0.135 (0.10)	0.0210 (0.069)
TORTOLI	0.000033	0.836 (0.047)	0.136 (0.092)	0.0272 (0.067)
ARZANA	0.0000245	0.874 (0.045)	0.0891 (0.089)	0.0373 (0.063)
SarHGDPa	0.0000302	0.904 (0.049)	0.0549 (0.091)	0.0407 (0.066)
LOCERI	0.0000193	0.873 (0.052)	0.0852 (0.098)	0.0422 (0.070)
NUORO	0.0000318	0.861 (0.042)	0.0950 (0.083)	0.0439 (0.059)
BARISARDO	0.0000217	0.856 (0.051)	0.0960 (0.093)	0.0480 (0.064)
SarHGDPb	0.0000369	0.855 (0.048)	0.0887 (0.093)	0.0566 (0.066)
ORISTANO	0.0000229	0.895 (0.043)	0.0453 (0.078)	0.0598 (0.054)
CAMPIDANO	0.000023	0.937 (0.032)	0 (0.006)	0.0626 (0.031)
LANUSEI	0.0000289	0.903 (0.045)	0.0343 (0.084)	0.0630 (0.060)
CAGLIARI	0.0000242	0.885 (0.044)	0.0426 (0.080)	0.0720 (0.055)
OLBIATEMPIO	0.0000222	0.805 (0.049)	0.113 (0.085)	0.0824 (0.059)
ILBONO	0.0000349	0.877 (0.048)	0.0360 (0.087)	0.0872 (0.060)
SASSARI	0.0000232	0.865 (0.044)	0.037 (0.080)	0.0980 (0.054)
CARBONIA	0.0000287	0.895 (0.035)	0 (0.025)	0.105 (0.031)

Supplementary Table 5: Estimated ancestry proportion and standard error from each of the ancient populations. LBK_EN represents the Neolithic farmer ancestry; Loschbour represents the pre-Neolithic hunter-gatherer ancestry; Yamnaya represents the post-Neolithic Steppe pastoralist ancestry.

Ancestry bin range	Stuttgart		Loschbour		Yamnaya	
	slope (s.e.)	Z	slope (s.e.)	Z	slope (s.e.)	Z
0%-100%	0.00263		0.00274		-0.000441	
	(0.000438)	6.00	(0.000418)	6.56	(0.000257)	-1.71
10%-90%	0.00197		0.00227		0.000272	
	(0.000544)	3.62	(0.000523)	4.34	(0.000328)	0.83

Supplementary Table 6: Correlation between outgroup f3 statistics and the proportion of Gennargentu ancestry. Outgroup f3 statistics of the form f3(Mbuti; Ancient, Sardinian) was calculated in steps of 5% of Gennargentu ancestry, and the slope of the correlation is reported here for each of the three ancient populations tested. Standard errors (s.e.) and Z-scores are estimated by blocked jackknife resampling removing 10 Mb at a time. As the effect in Figure 6B appears to be more extreme due to individuals with extreme Gennargentu ancestry (0-10% or 90-100%), we also re-estimated the correlation removing those bins.