Supplementary Note 1: Archaeological background

1.1. North Africa

Youssef Bokbot and Abdeslam Mikdad

1.1.1. The Cave of Ifri n'Amr o'Moussa

The Ifri n'Amr o'Moussa (IAM) cave is located close to Beht River, about 17 km northeast of the city of Khemisset, Morocco (Figure S1.1). The cave was identified as an archaeological site by Youssef Bokbot in 2005, but the surrounding area had been known since 1936, through the work of Armand Ruhlman on the prehistoric fortified sites of Oued Beht. The cave is located within the Zemmour plateau, a region with significant agricultural and mining potential that have been exploited since the Neolithic era.

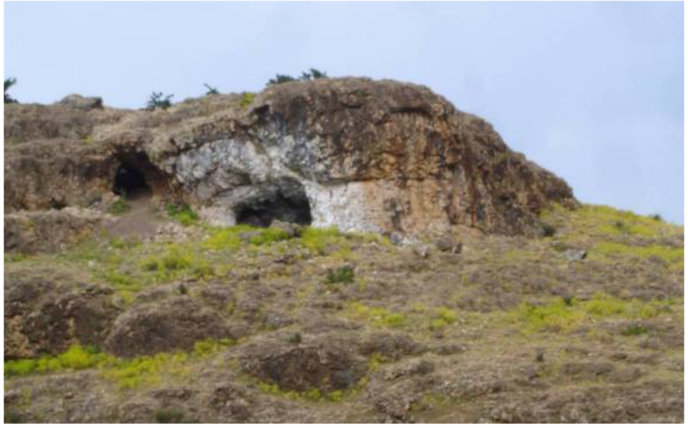


Figure S1.1 - The cave of Ifri n'Amr o'Moussa

The archaeological excavations in the cave of Ifri n'Amr o'Moussa started in 2006, leading to the discovery of habitats, burials and archaeological artifacts, belonging to two prehistoric civilizations:

- Chalcolithic or Copper Age (2,400 to 1,800 years BCE), represented by the Bell-Beaker culture, which spread throughout western and central Europe.

- Ancient Neolithic (5,400 to 4,800 years BCE), probably corresponding to the Cardial civilization that flourished throughout the Mediterranean.

The first excavation season at Ifri n'Amr o'Moussa in 2006 led to the discovery of copper objects and bone industry in stratigraphy, as well as several shards of bell-shaped ceramic. This finding positioned Ifri n'Amr o'Moussa as one of the few Bell-Beaker sites excavated in Morocco, and the major site of this civilization in the western basin of the Mediterranean[1]. The bone industry is very abundant in the first layer, consisting mostly on eyed needles and punches. The presence of eyed needles confirms the classification of this industry within the Chalcolithic culture. It is worth mentioning that some of the objects excavated in Ifri n'Amr o'Moussa are unique in their genres throughout all North Africa. For example, a necklace made from a boar's tusk, intentionally embracing the silhouette of a snake, as well as the mandible of a hedgehog transformed into a pendant (Figure S1.2).

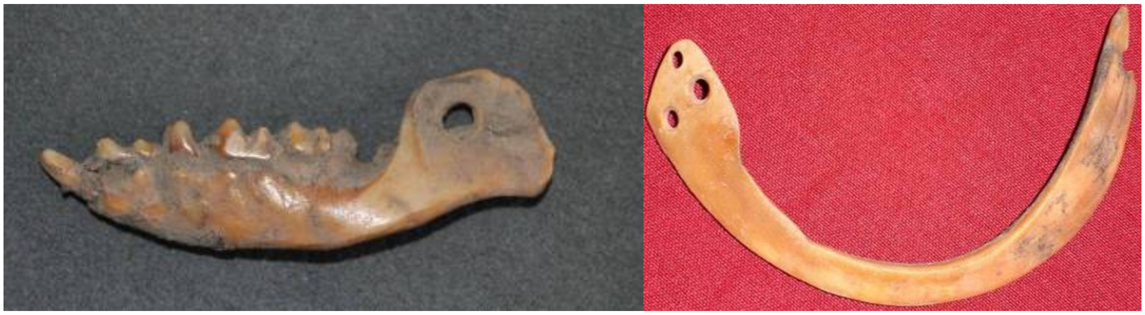


Figure S1.2 - Pendant made of the mandible of a hedgehog and necklace made from a boar's tusk, both excavated at Ifri n'Amr o'Moussa.

Two cooper artifacts were also excavated at Ifri n'Amr o'Moussa, a point and a punch. The metallic point from the Ifri n'Amr o'Moussa cave has an clear resemblance to the Western Europe "palmella" points, particularly frequent in the Iberian Peninsula and the French Atlantic and Mediterranean coasts. These metallic points are found in contexts dated to the Chalcolithic or the Late Neolithic, depending on the region. The copper punch is also related to contexts dating from the same period, between 4,750 - 4,050 BCE, and associated to the Bell-Beaker culture.

The Ifri n'Amr o'Moussa cave also delivered an Early Neolithic/Epipaleolithic sequence. Although not fully studied yet, this older sequence includes Neolithic layers marked by ash and charcoal deposits. These layers contain ceramics similar to Cardial pottery, decorated with simple rocking-stamped motifs, made with smooth and denticulated shells. Cereals remains obtained in the same layers containing the Cardial-like ceramic gave dates that documents the arrival of domesticated cereals and pottery printed with marine shells in the foothills of the Middle Atlas towards the third quarter of the 8th millennium BCE. The archaeozoological study of the site of the Neolithic/Epipaleolithic level of Ifri n'Amr o'Moussa[2] has revealed a great diversity of wildlife species. Based on that, it seems that these Neolithic populations performed hunting activities and were relatively mobile. However, it has been also observed the presence of a small number of domestic animals, indicating the possession of some head of cattle. This result implies that, in addition to a complex hunting activity, these were semi-sedentary populations could have maintained some domestic animals as well.

The exceptional character of the cave of Ifri n'Amr o'Moussa, is further reinforced by the discovery of seven human skeletons buried in sepulchral structures[3] in the Early Neolithic/Epipaleolithic layer. During the 2006 and 2007 season, an adult human skeleton (IAM.1) and two young children (IAM.2 and IAM.3) were excavated[4]. In 2010, four additional individuals were discovered (IAM.4, IAM.5, IAM.6 and IAM.7)[3]. All skeletons were in the same level, indicating the cave was used as a necropolis.

-IAM.1: This individual has an estimated age of 40 years and was identified as a female. The subject rested in supine left lateral flexion, with the head to the east.

- IAM.2 and IAM.3: The two children excavated in burials 2 and 3 have estimated ages of six and eighteen months, respectively. The first subject rested in dorsal decubitus, along a north-south axis with the head to the north. The second subject, visibly older than the first, was placed in the left lateral decubitus position, on an east-west axis with the head to the east.

- IAM.4: Grave 4 belongs to an adult subject, of 40 years of age, female and small. The individual is folded in a dorsal position, with the head to the east.

- IAM.5: Grave 5 contains an infant of about 5 years of age. This individual rested in left lateral decubitus, with the head to the east.

- IAM.6: Grave 6 contains a large adolescent who was about 16 years old. The individual rested in dorsal decubitus, with the head to the east and the lower limbs folded.

- IAM.7: Grave 7 contains a child who is very young and rests in supine position with his head to the east in a circular pit.

All the burials in Ifri n'Amr o'Moussa site are devoid of any artifacts, except for an original funeral ritual, which consists of placing a millstone on the skull (Figure S1.3). The size of the millstone is proportional to the age of the individual. An adult is entitled to a large stone, a child to a medium stone, and an infant to a very small one adapted to its calvaria. All the pits were shallow and carefully surrounded by stones. The archeological level where the samples were excavated is attributed to the early Neolithic *lato sensu*. It is also characterized by an important presence of small game, such as beef and swine[5]. These burials could indicate that Ifri n'Amr o'Moussa cave was in fact a Neolithic necropolis. This is entirely possible since Neolithic necropolises in caves are well documented in the region. These include El Kiffen[6], El-Mnasra[7], El Harhoura II[8] caves.

These burials were dated from 4,850 to 5,250 cal BCE (Figure S1.4), the same period when the Carial civilization radiated in the Mediterranean area. However, the ritual of depositing grindstones on the skulls of individuals have not been found in any other archaeological site of the Maghreb, nor in the Iberian Peninsula. The only cases of comparison, as we know, come from the island of Cyprus, notably the tombs 207 and 641 of Khirokitia, which date from the Pre Pottery Neolithic B technology[9]. Samples selected for this study are detailed in Table S1.1.

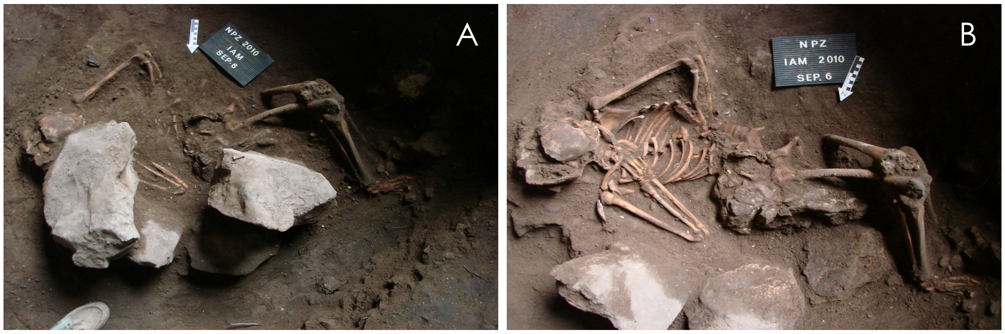


Figure S1.3 - Human remains excavated at Ifri n'Amr o'Moussa in grave number 6. Picture A shows the millstone place above the skull. Picture B shows how the millstone crushed the skull of the individual.

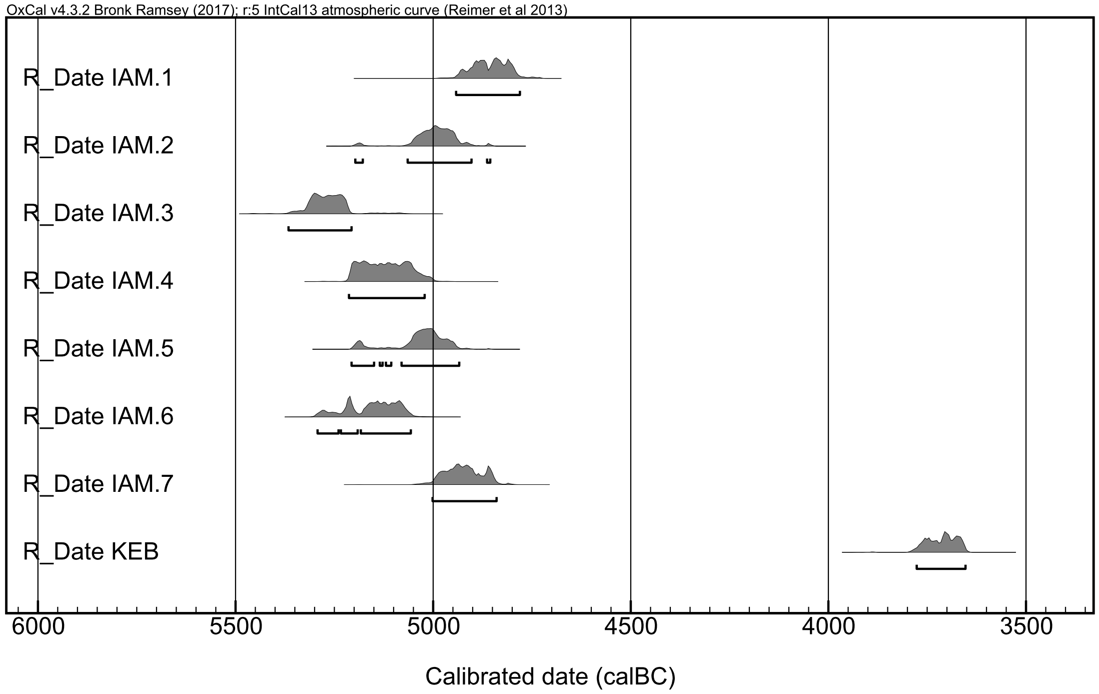


Figure S1.4 - Calibration of the radiocarbon dates obtained from Ifri n'Amr o'Moussa (IAM.#) and Kehf el Baroud (KEB) using IntCal13[10] and the software Oxcal v4.3.2[11].

1.1.2. The Cave of Kehf el Baroud

The Kehf el Baroud cave (Figure 1.5) is located about fifty kilometers east of the city of Casablanca. The cave consists of two rooms that communicate through a small gallery[12]. This archaeological site was discovered during the 50's by the Moroccan Caves Club and was initially excavated by De Wailly in the 60's. Although the initial project was abandoned, excavations continued in the site in 1993 and 1994, led by Bokbot and Mikdad.

In the cave of Kehf el Baroud, three main archeological layers have been identified:

1. A gray layer at the top: This first level is ashy, rich in snails and it contains many ceramic shards, cut flints, a copper point and a schist plate. It corresponds to a heavy habitation phase.

2. An intermediate white layer: This level contains less ceramic shards and it corresponds to a low human occupation of the cave. Human remains have been recovered from this layer.
3. A yellow layer at the bottom: This level is sterile.

Two samples has been radiocarbon dated from Kehf el Baroud site. The sample from the gray layer gave a date of 2,800 ± 110 BCE, while the white layer was dated around 3,210 ± 110 BCE[13].

Although both old and recent alterations have affected the Kehf el Baroud cave, it is clear that the archaeological layers are intact. The gray layer is the most important phase of the occupation of the site. Several artifacts have been retrieved from this level: a few campaniform shards (Figure S1.6), a cooper ax (Figure S1.7), bone industries, sieves and punches and lithic industries[13]. Based on the radiocarbon dates and the archaeological material, it is clear that the gray level corresponds to a Bell-Beaker horizon.

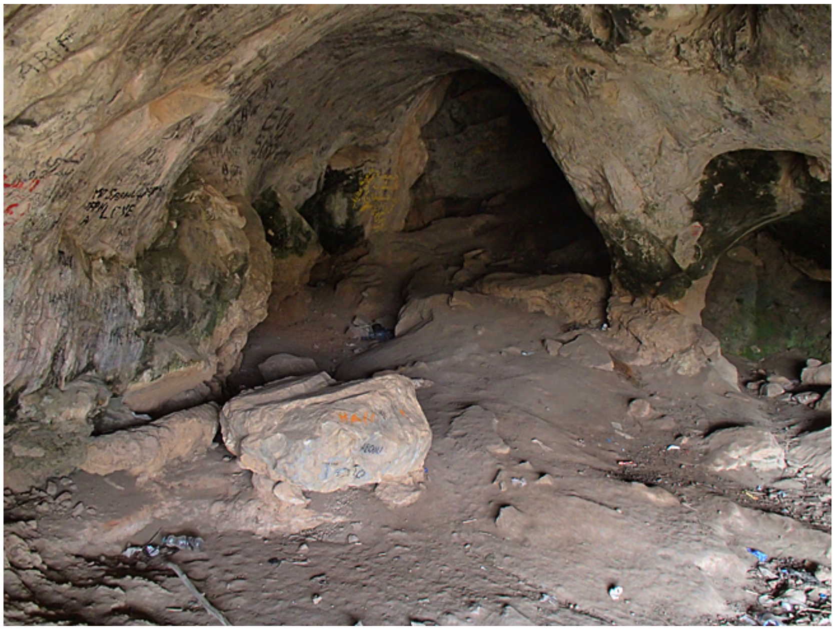


Figure S1.5: The cave of Kehf el Baroud

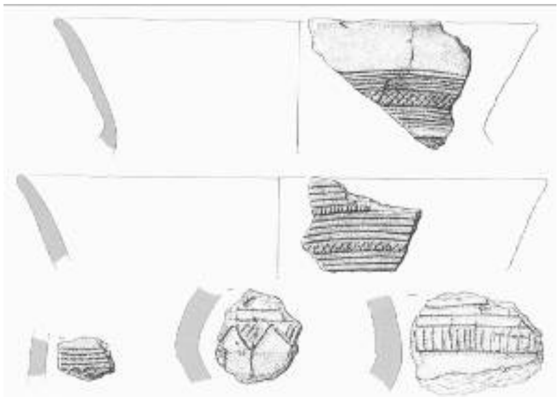


Figure S1.6: Bell-Beaker ceramic excavated at Kehf el Baroud

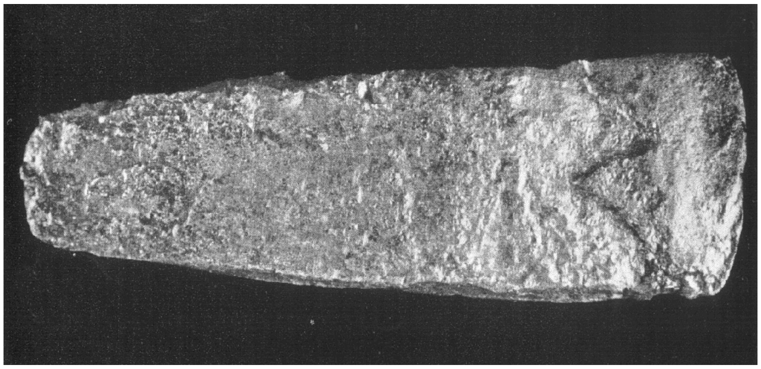


Figure S1.7: Flat copper ax excavated at Kehf el Baroud

The human remains analyzed in this study were obtained from the white layer, whose pottery remains are quite similar to Late Neolithic and Early Chalcolithic pottery from other locations in Morocco, such as the two neighboring necropolis of Rouazi-Skhirat and El Kiffen, and from Spain and Portugal, primarily in Los Millares and Vila Nova de Sao Pedro. The dating of the white layer performed by De Wailly[13], giving an age of about 3,200 BCE, has been disputed by most researchers. However, this date has been confirmed previously by thermoluminescence[12] and by calibrated radiocarbon dating in this study (Table S1.1, Figure 1.4).

1.2. Southern Spain

Dimas Martín-Socas, María Dolores Camalich Massieu, Francisco Javier Rodríguez-Santos, Jonathan Santana Cabrera and Aioze Trujillo Mederos

1.2.1. El Toro site

El Toro (36° 57’ 26” N; 4° 32’ 10” W) is a cave site located in Antequera (Málaga), concretely in the Sierra del Torcal, at an elevation of 1,190 meters above sea level. Sierra del Torcal is a wide karstic mountain range that separates Mediterranean Andalusia from the Sub-Betic System of southern Iberian Peninsula. The morphogenesis of this region is characterized by calcareous rocks and diaclastic systems that have conditioned the directions of karst flow (Figure S1.8). Numerous pits have been identified, some of which have provided evidence of occupation during the Neolithic period[14].

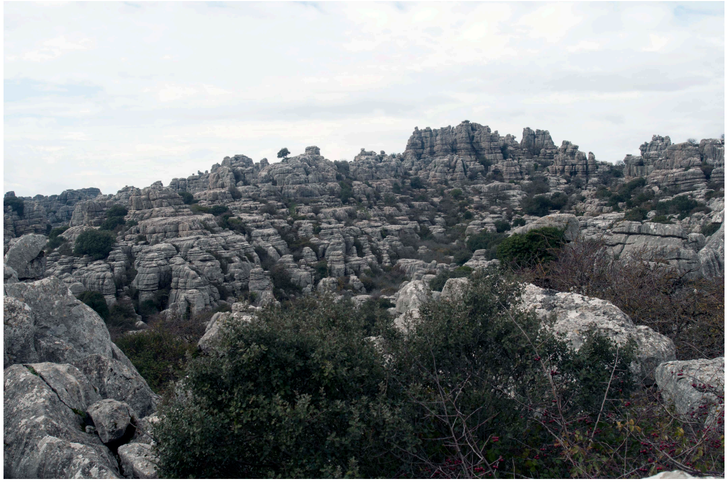


Figure S1.8: View of Sierra del Torcal (Málaga, Spain)

Today, the cave of El Toro shows an internal structure characterized by large fallen blocks, which collapsed prior to the Neolithic occupation. During the first quarter of the 4th millennium, the structure of the cave changed, possibly as the result of a tectonic movement or a collapse of the karst system. These changes included the closure of the original entrance, the configuration of a new entrance, the formation of 17-meter pit and the modification of the sedimentary fill in the southern sector.

Two of the authors of the present study (D.M-S. and M.D.C.) carried out five excavation campaigns in the cave (1977, 1980, 1981, 1985 and 1988), in the immediate area of the entrance (Figure S1.9). A stratified 2.40 m depth sequence was identified, divided into a series of archaeological units following the Harris Matrix[15]. These units were grouped into four chrono-cultural phases and one sterile phase[14,16] (Figure S1.10):

- Phase I: Superficial layer, where evidence of the more recent occupation of the cave has been identified (Roman, medieval and modern times).

- Phase II: Layer dating to the end of the third millennium BP[14], which is characterized by a decrease in occupation with less evidence of material remains and domestic activities.

- Phase III: Layer corresponding to the recent Neolithic, dated 3,370 - 3,220 BCE (4,250 - 3,950 cal BCE)[17].

- Sterile deposit, probably associated to a period of abandonment of the site.

- Phase IV: Layer corresponding to the Early Neolithic phase, dated between 6,200 - 5,980 BP (5,280 - 4,780 cal BC)[17,18] (Table S1.1, Figure S1.11).

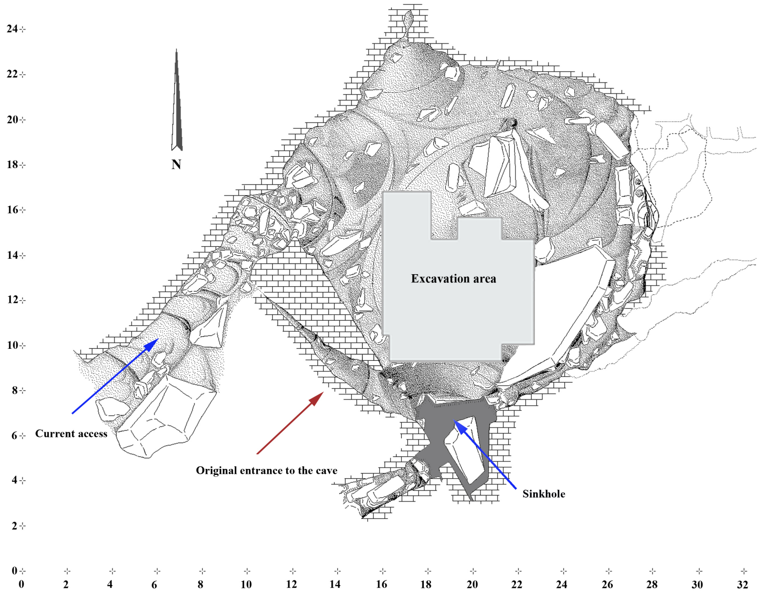


Figure S1.9: Structure of El Toro cave

Regarding phase VI, the archaeological complex includes lithic remains, bones and ceramics. Associated fauna remains are mainly those of ovicaprids, while seeds excavated correspond mostly to cereals. Functional analysis highlights a predominance of meat processing activities, along with evidence of bone, wood, leather and clay work. Also, it has been evidenced the production of ceramic *in situ*[19]. Pottery remains from phase VI in El Toro cave, are characterized by the use of highly diverse decoration techniques. Ceramics from El Toro were decorated including incise, printed, plastic, "boquique" and "almagra" techniques (Figure S1.12). The material culture observed in El Toro cave allows us to classify the remains from phase VI within the Andalusian Early Neolithic culture.

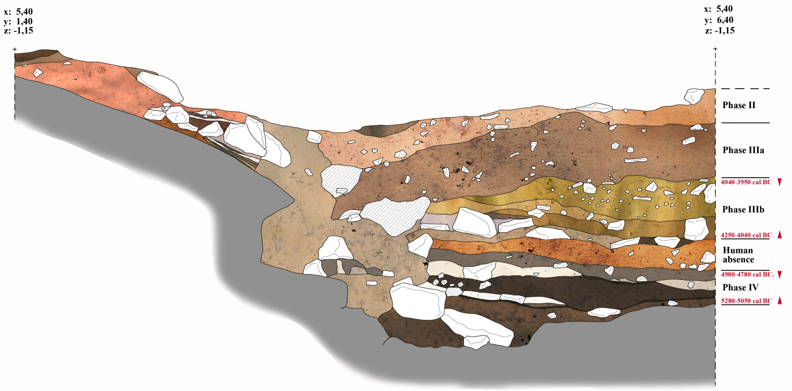


Figure S1.10: Stratigraphic sequence of El Toro cave

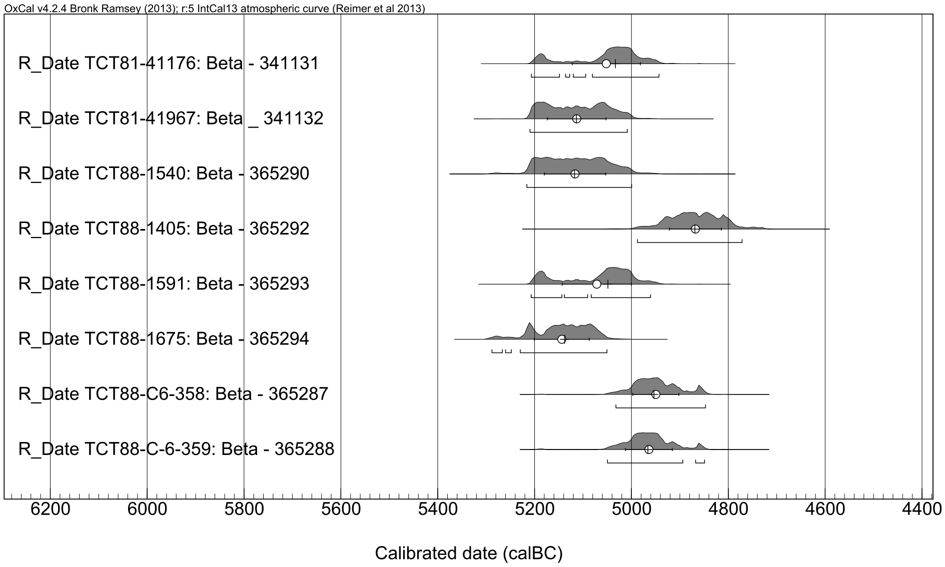


Figure S1.11: Calibration of the radiocarbon dates obtained from Early Neolithic remains excavated in El Toro cave, using IntCal13[10] and the software Oxcal v4.2.4[11].

Human remains were excavated in the Early Neolithic phase. The remains appeared scattered and without anatomical connection. They were distributed in two well-differentiated areas: contexts A and B. Context A is associated with the area of domestic occupation and is the source of most of the human remains. Context B is clearly a ritual area that is located in a small hollow, and it is spatially independent of the domestic space. The ritual area is characterized by the deposit of a human skullcap, with evidence of manipulation (Beta-365288: 6060 ± 30 BP), and the jaw of an adult individual (Beta-365287: 6050 ± 30 BP), accompanied by four ceramic vessels.

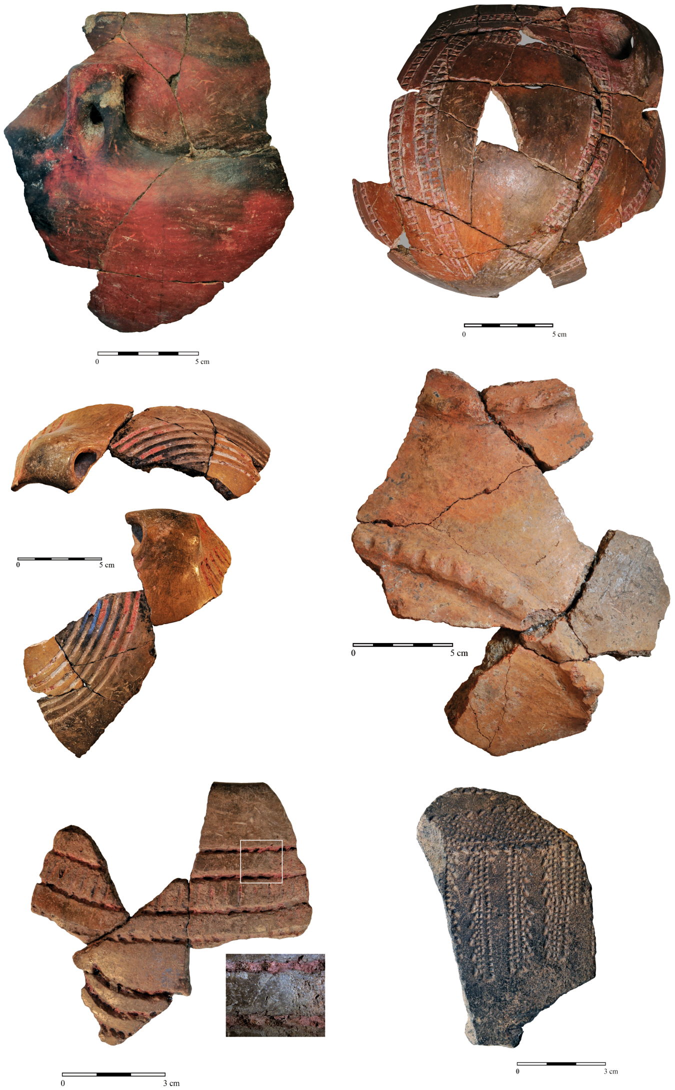


Figure S1.12: Several examples of pottery decoration techniques observed in remains excavated at El Toro site

Site ID	Sample ID	Library ID	Sample type	Museum ID	Radiocarbon dating ID	Conventional date	2 sigma calibration
IAM	**IAM.1**	AEH078	phalanx	IAM06S1.44	Beta - 443596	**5970 ± 30 BP**	**4937 - 4786 cal BCE**
	IAM.2	AEH079	petrous bone	IAM07S2	Beta - 443597	**6080 ± 30 BP**	**5055 - 4930 cal BCE**
	IAM.3	AEH080	petrous bone	IAM07S3	Beta - 443598	**6290 ± 40 BP**	**5325 - 5210 cal BCE**
	IAM.4	AEH081	petrous bone	IAM10S4.5	Beta - 443599	**6160 ± 30 BP**	**5215 - 5005 cal BCE**
		AEH166					
	IAM.5	AEH082	petrous bone	IAM10S5	Beta - 443600	**6100 ± 30 BP**	**5199 - 5176 cal BCE** **5066 - 4942 cal BCE**
		AEH075	phalanx	IAM10S5.12			
		AEH164					
	IAM.6	AEH083	petrous bone	IAM10S6.151	Beta - 443601	**6210 ± 30 BP**	**5290 - 5265 cal BCE** **5230 - 5195 cal BCE** **5180 - 5060 cal BCE**
		AEH167					
	IAM.7	AEH085	petrous bone	IAM10S7	Beta - 443602	**6030 ± 30 BP**	**5000 - 4840 cal BCE**
		AEH084	phalanx	IAM10S7.68			
KEB	**KEB.1**	AEH076	phalanx	KEB93.AV N.101	Beta - 443603	**4940 ± 30 BP**	**3780 - 3650 cal BCE**
		AEH165	phalanx				
	KEB.2	AEH077	phalanx	KEB94.AV.BL N.105	-	4940 ± 30 BP	3780 - 3650 cal BCE
	KEB.3	AEH086	phalanx	KEB94.AV.BL N.158			
	KEB.4	AEH087	phalanx	KEB N.154			
		AEH168	phalanx				
	KEB.5	AEH160	teeth	KEB93.94 d1			
	KEB.6	AEH161	teeth	KEB93.94 d2			
	KEB.7	AEH162	teeth	KEB93.94 d3			
	KEB.8	AEH163	teeth	KEB93.94 d4			
TOR	**TOR.1**	AEH017	teeth	46158-2	-	*6090 ± 110 BP*	*5280 - 4780 cal BCE*
	TOR.2	AEH018	teeth	1540a			
	TOR.3	AEH020	teeth	46645			
	TOR.4	AEH021	teeth	22123			
	TOR.5	AEH074	skull fragment	40858-2	**Beta – 365288**	**6060 ± 30 BP**	**5040 - 4850 cal BCE**
		AEH091					
	TOR.6	AEH092	teeth	46262-1;TCT-85	-	*6090 ± 110 BP*	*5280 - 4780 cal BCE*
		AEH169					
	TOR.7	AEH093	teeth	1706;TCT-88			
		AEH170					
	TOR.8	AEH094	teeth	20896;TCT-80			
		AEH171					
	TOR.9	AEH095	teeth	46645;TCT-85			
	TOR.10	AEH096	teeth	46456-1;TCT-85			
	TOR.11	AEH097	teeth	328; TCT-88	**Beta - 365287**	**6050 ± 30 BP**	**5030 - 4850 cal BCE**
	TOR.12	AEH098	teeth	1405-2;TCT-88	-	*6090 ± 110 BP*	*5280 - 4780 cal BCE*
BOT	**BOT.1**	AEH099	teeth	27/275	-	-	-

Table S1.1 - Archaeological samples included in this study. This table includes information on sample ID, library ID, sample type, museum ID, radiocarbon dating ID, conventional and 2-sigma calibrated radiocarbon dates. Datings performed directly on human remains are indicated with bold fonts. Datings performed with other material co-excavated with the samples are indicated with italics.

Analysis of the human bones of context A shows a specific treatment of the funerary remains during the Early Neolithic period. The anthropological record consists of small bones, some cranial and jaw fragments, as well as loose teeth[19]. Micro-morphological studies show that the presence of small bones and teeth is the result of anthropic processes contemporaneous with the Early Neolithic human occupation of the cave of El

Toro. The most plausible explanation for the presence of a specific part of the skeleton, coupled with the absence of long bones and skulls, is that it responds to specific funerary practices. In this case, it would consist on (a) a non-definitive primary burial of the deceased person for their skeletonization and (b) the transfer of the remains to their final resting place, as observed in other European contexts[20]. This practice would result in the selection of certain bones, so that fragmented and smaller remains would become part of the waste materials mixed with those from domestic activities[19]. El Toro cave shows evidence of different activities associated with everyday domestic life, such as food processing, skin treatment or ceramic manufacturing. It has been also demonstrated that El Toro cave was used for keeping the livestock during occupancy times, as well as successive combustion episodes[16]. These results explain the extensive dispersion of all human bones in the room space and demonstrate that their treatment does not differ from other waste related to domestic activities.

1.2.1. Los Botijos site

The Cave of the Botijos is a well-known Neolithic archaeological site in the Benalmádena region (Málaga). As in El Toro Cave, diverse pottery decoration techniques have been observed in Los Botijos, including printed, incised, boquique and almagra[21]. Although the human remain included in this study has not been dated, the archaeological materials recovered from the cave allowed us to assign the remains to the Andalusian Early Neolithic culture.

Supplementary Note 2: DNA extraction, library preparation and enrichment

Rosa Fregel, Morten Rasmussen, Andre Elias Rodrigues-Soares, Joshua Kapp and Alexandra Sockell

2.1. DNA extraction

To avoid contamination, both DNA extraction and library preparation steps were performed in a dedicated clean laboratory. Well-preserved teeth and petrous bones were sampled from Ifri n'Amr ou Moussa (IAM), Kehf el Baroud (KEB), El Toro (TOR) and Los Botijos (BOT) (Table S1.1), as the cementum layer in teeth roots and the inner part of petrous bones are considered the best sources for ancient DNA (aDNA)[22]. Additionally, we used phalanx samples when petrous bones and teeth were unavailable, expecting the phalanges' high bone density to also allow a better conservation of the endogenous DNA. Bone samples were extracted from all seven individuals from IAM. A total of eight different well-conserved phalanx/teeth samples were extracted from KEB. Twelve different individuals were sampled from TOR. All TOR samples were teeth, except for TOR.5 that was a piece of the skull excavated on the site (see Supplemental Note 1 and Table S1.1 for details). As KEB and TOR are mass burials we cannot rule out *a priory* the possibility that the same individual could be sampled twice. Finally, one teeth sample was included from BOT.

Bone samples were decontaminated by removing the superficial layer using a Dremel tool and a metallic bit. After decontamination, a piece of bone was pulverized using a Retsch™ MM 400 Mixer Mill. By choosing the mixer mill instead of direct drilling of the bone, we expected to have a higher endogenous DNA recovery[23]. Teeth samples were decontaminated by wiping the surface with tissue paper soaked in 10% bleach solution, rinsed with 70% ethanol and UV-irradiated for 15 min on each side. To obtain the root cementum, teeth were cut in the middle line and the dentine from root was removed by drilling using a Dremel tool. Finally, a piece of the root cementum was cut with a Dremel and pulverized with a mixer mill. All metallic material was thoroughly cleaned with a bleach solution after use, rinsed with 70% ethanol and UV-irradiated for one hour to avoid cross-contamination between samples.

DNA was extracted following the protocol proposed by Dabney et al.[24]. Briefly, up to 120 mg of bone/teeth powder was incubated overnight at 37°C in the extraction buffer (0.45 M EDTA pH 8, 0.25 mg/ml proteinase K). Then, DNA was purified by means of a silica-based protocol, using binding buffer (5 M Guanidine hydrochloride, 40% isopropanol, 0.05% Tween 20), and MinElute spin columns (QIAGEN). To reduce the number of centrifugation steps, we used the spin columns with an extension reservoir (BioRad) previously treated with bleach and UV light to avoid contamination. Finally, aDNA was eluted in 50 μl of elution buffer (1 mM EDTA pH 8, 10 mM Tris-HCl pH 8, 0.05% Tween 20).

In order to improve endogenous DNA recovery, several bone and teeth samples were treated with sodium hypochlorite following Korlevic et al.[25]. The extraction process was the same as before[24], but with an initial 15 min incubation step with a 0.5% sodium hypochlorite solution.

2.2. Library preparation

Double-stranded libraries were prepared following Meyer and Kircher[26]. Briefly, 25 μl of DNA were submitted to blunt-end repair using T4 polynucleotide kinase (0.5 U/μl) (Thermo Fisher) and T4 DNA polymerase (0.1 U/μl) (Thermo Fisher). Samples were then purified using the Nucleotide Removal Kit (QIAGEN), but using 10 volumes of the PNI buffer. Then, p5/p7 adapters were ligated to the DNA using T4 DNA ligase (0.125 U/μl) (Thermo Fisher) and the adapter filled-in using the large fragment of *Bst* polymerase (0.3 U/μl) (NEB). Finally, DNA libraries were amplified and indexed by PCR, using AmpliTaq Gold Hot Start polymerase (Thermo Fisher), primer IS4[26] and 1 μl of an index primer with a seven nucleotide barcode (Illumina). The amplified libraries were then purified using SPRI beads (Beckman Coulter) and quantified on a Qubit fluorometer (Thermo Fisher). The average library DNA concentrations was 21.76 ± 1.99 ng/μl (Table S2.1). The minimum value was observed for sample AEH082 with 3.02 ng/μl.

2.3. Shotgun sequencing

Shotgun libraries were ran on an Illumina NextSeq 500 platform (paired-end reads, 2 x 75 bp). Shotgun data was analyzed following the pipeline described in Supplementary Note

3. The amount of endogenous DNA was calculated by dividing the number of reads after filtering by the total number of trimmed reads. In order to normalize results, an equal number of raw reads (4 million) was used for comparison between samples. Endogenous DNA in the shotgun libraries (no sodium-hypochlorite treatment) accounted for 1.68 ± 1.29% of the total (median = 0.49%, IQR = 0.072%–2.064%). Duplicate rate on shotgun libraries was 1.20 ± 0.09% (Table S2.1).

Site	Sample ID	Library ID	DNA concentration (ng/µl)	Insert size	Endogenous DNA (%)	Duplicates rate (%)
IAM	IAM.1	AEH078	27.6	40.8	0.008	1.62
	IAM.2	AEH079	27.8	35.9	0.104	1.33
	IAM.3	AEH080	22.2	41.5	0.30	1.04
	IAM.4	AEH081	27.8	40.4	0.79	1.12
		AEH166*	18.0	47.3	0.28	2.31
	IAM.5	AEH082	3.02	41.2	3.66	1.87
		AEH075	16.7	59.9	2.78	1.06
		AEH164*	19.0	47.5	2.09	1.51
	IAM.6	AEH083	28.8	40.5	1.05	1.03
		AEH167*	20.8	45.5	0.07	5.93
	IAM.7	AEH085	25.2	39.9	2.13	1.05
		AEH084	19.5	39.8	3.29	1.33
KEB	KEB.1	AEH076	20.6	56.1	0.49	1.15
		AEH165*	20.4	47.9	1.96	2.61
	KEB.2	AEH077	17.2	50.5	0.016	1.40
	KEB.3	AEH086	29.4	42.7	0.037	1.06
	KEB.4	AEH087	28.4	46.1	1.93	1.13
		AEH168*	12.6	50.2	3.66	7.23
	KEB.5	AEH160*	17.1	47.3	0.11	4.44
	KEB.6	AEH161*	19.8	49.8	12.11	5.48
	KEB.7	AEH162*	17.4	47.3	0.49	5.32
	KEB.8	AEH163*	25.0	47.6	5.92	7.24
TOR	TOR.1	AEH017	12.6	45.6	0.11	0.85
	TOR.2	AEH018	21.6	46.8	0.10	0.86
	TOR.3	AEH020	19.8	47.1	0.02	1.09
	TOR.4	AEH021	24.8	47.9	0.03	1.07
	TOR.5	AEH074	20.2	54.7	4.30	1.49
		AEH091	28.2	47.9	2.00	1.57
	TOR.6	AEH092	28.6	47.4	17.17	1.22
		AEH169*	14.4	41.3	33.03	6.98
	TOR.7	AEH093	29.6	46.5	0.78	1.23
		AEH170*	14.1	41.3	1.13	2.80
	TOR.8	AEH094	26.6	46.7	1.21	1.16
		AEH171*	11.1	41.3	5.62	3.49
	TOR.9	AEH095	26.6	45.2	0.05	0.97
	TOR.10	AEH096	26.2	47.5	0.01	1.03
	TOR.11	AEH097	23.4	46.7	0.36	1.19
	TOR.12	AEH098	26.2	48.4	0.10	1.12
BOT	BOT.1	AEH099	30.2	48.6	2.45	1.48

Table S2.1 - Detailed information about all shotgun libraries. Samples treated with bleach are indicated with an asterisk.

For eight samples, we tested the use of the pre-extraction sodium-hypochlorite treatment to improving the recovery of endogenous DNA[25]. The improvement on both KEB and TOR samples was low, with enrichment rates between 1.44X and 4.66X. Libraries treated with sodium hypochlorite also exhibited lower complexity, with duplicate rates increasing from 2.27X to 6.40X.

Archaeological site	Sample ID	Enrichment (bleach)	
		endogenous DNA	duplicates
IAM	IAM.4	0.35	2.06
	IAM.5	0.75	1.42
	IAM.6	0.07	5.75
KEB	KEB 1	4.00	2.27
	KEB 4	1.90	6.40
TOR	TOR.6	1.92	5.73
	TOR.7	1.44	2.28
	TOR.8	4.66	3.01

Table S2.2 - Comparison of enrichment on endogenous DNA and fraction of duplicate reads for libraries with and without sodium-hypochlorite treatment

For IAM samples the effect of the sodium-hypochlorite treatment was clearly detrimental (Table S2.2), while exhibiting the same rate of complexity loss. It is possible that the bad performance of sodium-hypochlorite treatment in IAM can be related to high degradation of the sample. However, both IAM samples were obtained from the petrous bone and, although the bone powder was homogenized and obtained from the same region[27], we cannot rule out differences on the starting bone material.

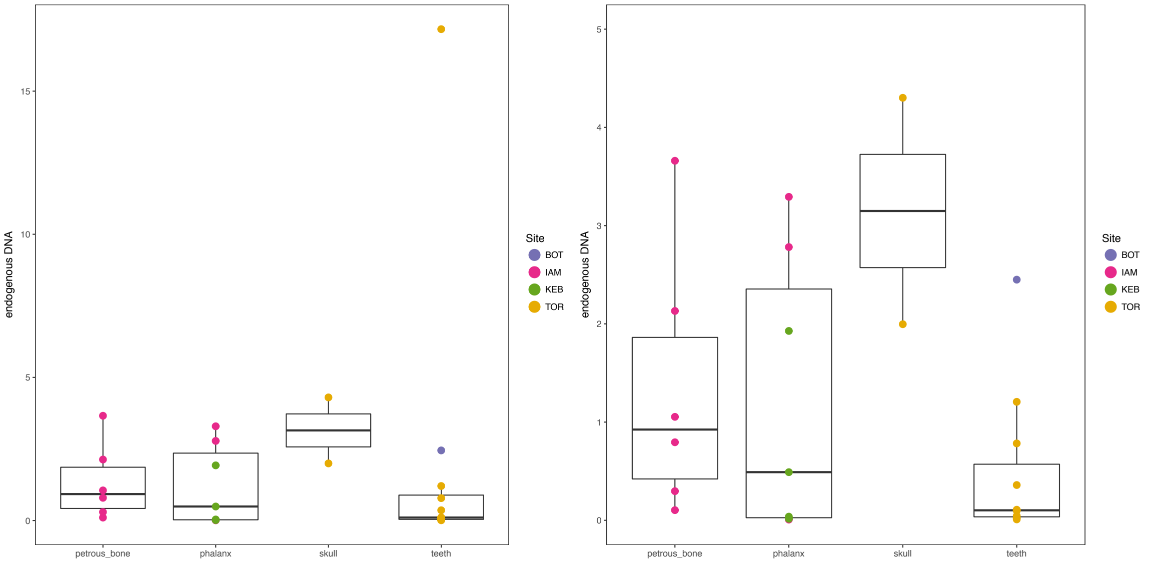


Figure S2.1 - Comparison of endogenous DNA rates based on the type of source material (A) Shows all libraries. (B) Detail of the plot excluding the outlier from TOR.

Excepting the remarkably good conserved piece of skull from TOR, the source material that yielded the highest endogenous content was the petrous bone (Figure S2.1). Phalanx samples also delivered relatively good endogenous DNA yields, on average better than teeth. For two samples from IAM, it was possible to get the petrous bone and a phalanx extracted. In those cases, both materials produced similar endogenous DNA rates. For example, for IAM.5 the endogenous DNA percentage was 3.66% and 2.78% respectively for the petrous bone and the phalanx. In the case of IAM.7, those values accounted for 2.13% and 3.29%, with a higher rate for the phalanx sample (Table S2.1). These results indicate that phalanges are also good sources for aDNA, as it has been already observed on forensic analyses[28].

2.4. WISC capture

Samples with low endogenous DNA content were enriched for human endogenous DNA using whole-genome in-solution capture (WISC)[29]. Briefly, human genomic DNA was used to prepare libraries with adapters containing T7 RNA polymerase promoters. These T7 libraries were submitted to *in vitro* transcription with biotinylated UTP to generate RNA baits covering the whole human genome. Then, the biotinylated RNA baits were used for capturing endogenous DNA from aDNA libraries, where the human DNA is pulled with magnetic streptavidin-coated beads and the unbound DNA is washed away. Human cot-1 and salmon sperm DNA was used to block repetitive and interspecific DNA from binding the baits, and adapter blockers to minimize adapter hybridization. The captured DNA was then purified, treated with RNase to remove baits, and finally amplified by PCR using adapter primers. The minimum number of PCR cycles needed for amplification and the final DNA concentration was determined by qPCR using the KAPA Library Quantification Kit (Kapa Biosystems).

WISC capture was applied to 26 shotgun libraries. Samples treated with bleach were excluded from the WISC capture due to their low complexity. Sample TOR.9 was excluded from WISC capture and further analyses due to absence of damage patterns (see Supplementary Note 3). Finally, sample TOR.6 was also excluded from capture due to relatively high endogenous DNA content (17.2%). Overall, WISC performance was good, with an average value of enrichment on unique reads was 14.6 ± 7.9X, with minimum and maximum values of 1.9X and 84.7X, respectively. As expected, WISC capture also produced a decrease on the libraries complexity. The average duplicate rate increase after capture was 19.3 ± 6.6X, with values ranging between 3.1X and 59.7X.

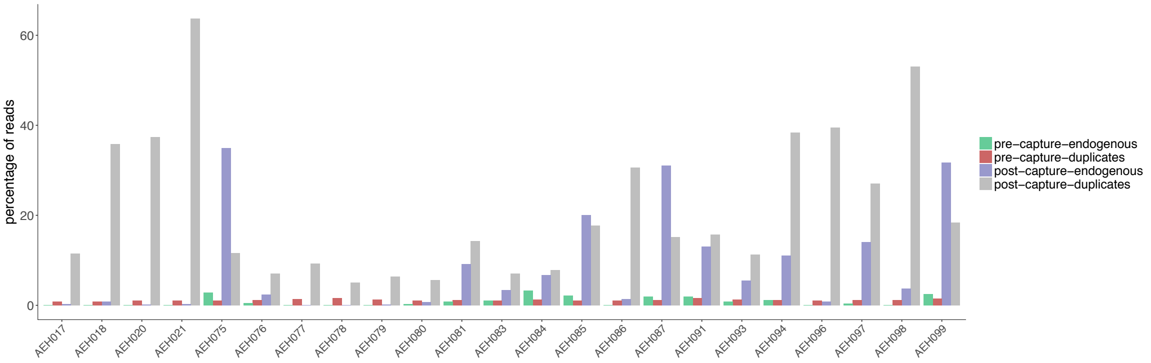


Figure S2.2 - Comparison of % endogenous content in pre-capture and post-capture sequencing data of libraries captured using WISC

Archaeological site	Sample ID	Library ID	Enrichment	
			endogenous DNA	duplicates
IAM	IAM.1	AEH078	2.1	3.1
	IAM.2	AEH079	1.9	4.8
	IAM.3	AEH080	2.3	5.4
	IAM.4	AEH081	11.5	12.8
	IAM.5	AEH075	12.6	10.9
	IAM.6	AEH083	3.3	6.9
	IAM.7	AEH085	9.4	16.9
		AEH084	2.0	5.9
KEB	KEB.1	AEH076	4.9	6.1
	KEB.2	AEH077	6.0	6.6
	KEB.3	AEH086	37.2	28.9
	KEB.4	AEH087	16.1	13.4
TOR	TOR.1	AEH017	2.8	13.5
	TOR.2	AEH018	8.2	41.9
	TOR.3	AEH020	6.4	34.4
	TOR.4	AEH021	11.7	59.7
	TOR.5	AEH091	6.5	10.0
	TOR.7	AEH093	7.1	9.2
	TOR.8	AEH094	9.2	33.2
	TOR.10	AEH096	84.7	38.2
	TOR.11	AEH097	39.0	22.8
	TOR.12	AEH098	36.7	47.6
BOT	BOT.1	AEH099	12.93	12.37

Table S2.3 - Comparison of pre-capture and post-capture sequencing data of samples captured using WISC

As shown in Figure S2.3, the performance of WISC capture was better in TOR than in KEB and IAM sites. The average enrichment on unique reads for TOR was 21.23 ± 16.28X. Overall, the site that had worse WISC results was IAM, with a unique reads enrichment of 5.64 ± 3.31X. Finally, KEB produced intermediate results with an average enrichment on unique reads of 16.08 ± 14.98X.

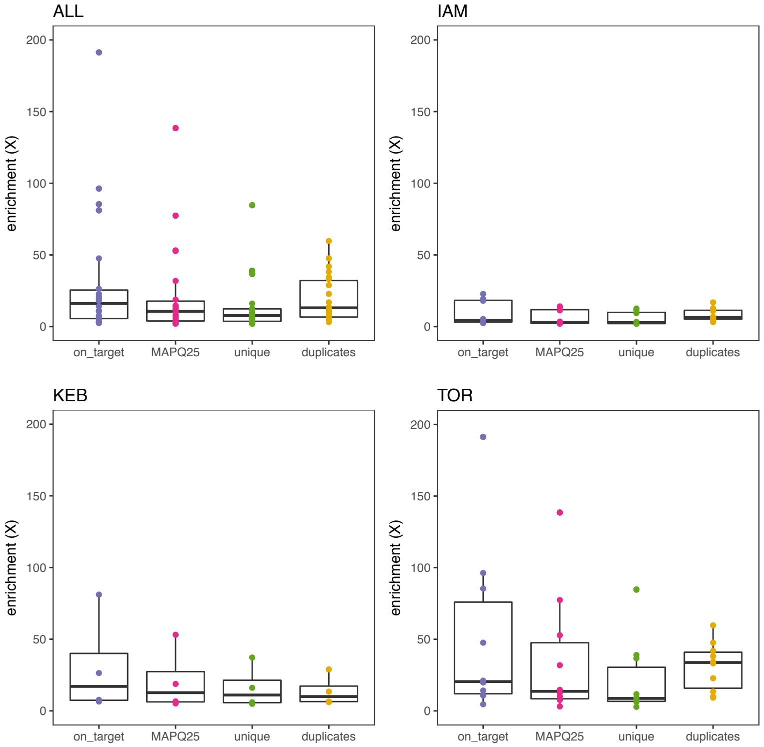


Figure S2.3 - Comparison of WISC performance on the all samples and for each of the three archaeological sites

2.5. MEGA capture

Selected samples were also captured using baits designed for enriching aDNA on SNPs contained in the Illumina MEGA array. A capture panel matching the sites on the MEGA array was submitted for SeqCap EZ design (Nimblegen, Roche). Due to the degraded nature of aDNA, the design was limited to probes covering SNP position, and allowing up to 20 close matches, defined by Roche as a sequence that differs by no more than 5 insertions, deletions or SNPs from the probe. The final capture panel targets ~90% of the 1.7 million MEGA sites. For MEGA capture, libraries were pooled together to obtain ~500 ng of DNA. Only libraries with similar endogenous DNA content and from the same site were pooled together. Samples were captured in multiplex following the manufacturer recommendations.

MEGA capture consistently showed an increase of coverage on the targeted SNPs, with an average increase of intersected MEGA array SNPs of 94.5 ± 28.2X (See table S2.4 for details). Enrichment values ranged form 24.1X in sample TOR.6 to 163X in sample TOR.7.

Site ID	Sample ID	Library ID	shotgun			MEGA capture			Improvement on SNP coverage
			number of SNPs intersected	Coverage (%)	Depth	Number of SNPs intersected	Coverage (%)	Depth	
IAM	**IAM.3**	AEH080	358	0.02%	1.00	48,902	2.85%	1.03	136.6
	IAM.4	AEH081	996	0.06%	1.00	111,019	6.73%	1.07	111.5
	IAM.5	AEH075	4,255	0.24%	1.00	222,165	14.07%	1.12	52.2
	IAM.6	AEH083	1,230	0.07%	1.00	145,827	9.01%	1.10	118.6
	IAM.7	AEH084	3,935	0.22%	1.00	98,382	5.89%	1.06	25.0
		AEH085	2,564	0.15%	1.00	222,434	14.45%	1.15	86.8
KEB	**KEB.1**	AEH076	699	0.04%	1.00	91,445	5.38%	1.04	130.8
	KEB.4	AEH087	2,713	0.15%	1.00	379,858	26.79%	1.25	140.0
TOR	**TOR.5**	AEH074	6,758	0.38%	1.00	208,056	12.87%	1.10	30.8
		AEH091	2,861	0.16%	1.01	166,282	10.14%	1.08	58.1
	TOR.6	AEH092	25,502	1.46%	1.01	614,933	50.77%	1.46	24.1
	TOR.7	AEH093	1,071	0.06%	1.00	175,193	10.85%	1.10	163.6
	TOR.8	AEH094	1,656	0.09%	1.00	248,628	15.95%	1.14	150.1

Table S2.4 - Comparison of pre-capture and post-capture sequencing data of samples captured using MEGA capture

Supplementary Note 3: ***Mapping, filtering and authentication criteria***

Rosa Fregel and María C. Ávila-Arcos

Samples with high endogenous DNA content (>10%) and post-capture libraries were sequenced to saturation on an Illumina NextSeq 500 platform (paired-end reads, 2 x 75 bp and 2 x 42 bp). Quality features of the sequencing data were evaluated at different stages using FASTQC version 0.11.5 (http://www.bioinformatics.babraham.ac.uk). Reads were trimmed and adapters removed using AdapterRemoval[30] version 1.5.4, with a minimum trim length of 30 bp and a minimum base quality of 20. Paired-end reads were merged with the default minimum overlap length (11 bp). Trimmed merged reads were then mapped to the human reference genome (GRCh37) using BWA[31] version 0.7.12. For libraries sequenced using the 2 x 42 bp paired-end method, unmerged reads up to 141 bp were also kept for analysis, replicating the same insert size as 2 x 75 bp paired-end merged reads. For libraries sequenced using the 2 x 42 bp paired-end method, we also used the clipOverlap function form bamUtil v.1.0.14 to trim overlaps smaller than 11 bp on paired-end reads (http://genome.sph.umich.edu/wiki/BamUtil). The mapping was performed using "bwa -aln" with the seed option (-l) disabled. Filtering of the mapped reads involved: a) removing reads with a mapping quality lower than 30, b) removing duplicated reads and c) excluding reads with alternative mapping coordinates (i.e. controlling for the information in the XA, XT and X0 tags). All filtering was performed using SAMtools[32] version 0.1.18. Indel Realigner from GATK[33] version 2.5.2 was used for improving the quality of alignments around indels. Bam files from different runs were merged per sample using SAMtools merge[32]. At this point, several samples were removed from the analysis due to low read counts: IAM.1, IAM.2, KEB.2, KEB.5 and TOR.9.

MapDamage[34] version 2.0.2 was used to determine the presence of misincorporations and DNA fragmentation patterns expected for ancient DNA. In order to minimize the effect of post-mortem damage in donwnstream analysis, mapDamage was also used to rescale the quality of bases likely affected by cytosine deamination. Contamination rates based on mtDNA were calculated using ContamMix[35] version 1.0-10. For this analysis, we used the mtDNA bam files with 4 bp trimmed at both ends, to avoid damage interfering with contamination estimations.

Post-mortem damage patterns were as expected. In all cases, average insert sizes were around 40 - 50 bp (Table S2.1, Figure S3.1). Deamination patterns at the 3' end were ~40% for IAM samples, ~27% for KEB samples and ~23% for TOR and BOT (Table S3.1, Figure S3.1). Regarding contamination rates (Table S3.1), the values observed from capture data were higher to those from shotgun data. On average, contamination rates were 4.5% for IAM, 2.8% for KEB and 6.8% for TOR. Although sample IAM.3 was included in the mtDNA analysis, it is important to take into consideration that mtDNA coverage for this sample was not high enough for providing a reliable contamination estimate, and IAM.3 results should be taken with caution.

Archaeological site	Sample ID	3' damage	5' damage	Contamination average	Contamination upper bound	Contamination lower bound
IAM	**IAM.3**	42.9%	43.0%	6.42	21.49	1.90
	IAM.4	37.0%	36.7%	4.66	6.64	3.26
	IAM.5	32.1%	32.5%	3.68	5.26	2.62
	IAM.6	44.8%	44.2%	5.28	8.48	3.02
	IAM.7	41.0%	40.2%	2.35	3.39	1.67
KEB	**KEB.1**	24.4%	24.6%	3.03	6.31	1.23
	KEB.3	34.8%	33.5%	3.16	5.73	1.72
	KEB.4	20.5%	20.3%	2.49	3.27	1.89
	KEB.6	24.4%	24.5%	3.77	6.72	1.98
	KEB.7	31.2%	31.0%	1.44	4.31	0.31
	KEB.8	27.2%	27.3%	3.03	6.23	1.27
TOR	**TOR.1**	33.8%	33.6%	1.42	2.07	0.98
	TOR.2	26.9% / -	26.3% / -	14.50 / -	17.82 / -	11.71 / -
	TOR.3	30.02% / -	28.7% / -	7.99 / -	14.16 / -	4.82 / -
	TOR.4	24.6% / -	24.5% / -	17.52 / -	24.64 / -	12.22 / -
	TOR.5	22.4% / 60.4%	22.8% / 60.6%	19.95 / 4.47	31.58 / 8.76	11.64 / 1.96
	TOR.6	19.6%	20.8%	4.15	5.89	2.94
	TOR.7	14.7%	14.8%	7.06	8.36	5.88
	TOR.8	21.0%	21.4%	3.95	4.80	3.26
	TOR.10	13.6% / -	13.2% / -	10.64 / -	18.07 / -	5.60 / -
	TOR.11	28.6%	27.7%	3.88	5.00	2.98
	TOR.12	24.0%	23.1%	1.50	2.30	0.93
BOT	**BOT.1**	20.5%	19.9%	3.91	4.90	3.11

Table S3.1 - Ancient DNA authentication parameters

Several samples from TOR exhibited high contamination rates. This is probably due to extensive archaeological and anthropological work done on the samples, prior to the design of this project. The higher value of contamination was observed for the skull sample, which we suspect was previously handled without gloves. The presence of one source of contamination was perfectly clear from both mtDNA and Y-chromosome results, where two different haplogroups were observed in both cases. For the mtDNA, reads from both H and J haplogroups were obtained, whereas SNPs for G-M201 and R-M269 were present in the Y-chromosome analysis. To see if it was possible to reduce the presence of contamination, a second library was prepared from a different region of the skull sample and with a deeper surface removal by drilling. However, the same value of contamination was observed, indicating that the contaminant DNA penetrated deep in the sample, maybe through sweat during manipulation. With the aim of including those samples in further analysis, we applied an additional damage-filtering step on the "contaminated" merged bam files. For that, we used the MR score from mapDamage[34], which correlates with the expected number of C → T changes on the molecule due to post-mortem damage. This filtering was applied to all samples with an upper bound contamination estimation higher than 10%: TOR.2, TOR.3, TOR.4, TOR.5 and TOR.10. After filtering for MR ≥ 0.8, sample TOR.5 presented a contamination rate of 4.47%, and the sample was clearly classified as J2b1a and G-M201 for the mtDNA and the Y chromosome, respectively (See Supplementary Note 4 and 5). As we are filtering for reads with clear damage, cytosine deamination at the end of the reads increased from ~22% to ~60% (Table S3.1). For the other samples (TOR.2, TOR.3, TOR.4 and TOR.10), the number of merged reads was lower than for TOR.5, and after filtering for MR ≥ 0.8, the mtDNA coverage was too low for analysis. For that reason, samples TOR.2, TOR.3, TOR.4 and TOR.10 were removed from subsequent analyses. Although sample TOR.5 is included in the mtDNA and chromosome Y analyses, results from this sample should be taken with caution.

Reads obtaining from sequencing the extraction and library prep blanks were also included in the aforementioned analyses. Although all the lab procedures were carried out with the highest standards for avoiding contamination from modern molecules, amplified DNA were observed in both extraction and library prep blanks. Average DNA concentration was 19.20 ng/µl for the extraction blanks and 9.63 ng/µl for the library prep blanks. Electrophoretic analysis of the aDNA libraries and the blanks indicate that the overall insert size of aDNA libraries was higher than that on extraction blanks. In the case of the library preparation blanks, electrophoresis showed that most of the amplified DNA was related to adapter dimmers. A total of ~4 million reads were sequenced for all the extraction and library prep blanks. Although the endogenous human DNA rates of blanks overlap those of the aDNA libraries (Figure S3.3), duplicate rates are in average 70X higher, indicating that the PCR reactions for these samples started from very few molecules. MapDamage analysis of the blanks corroborated the absence of post-mortem damage, ruling out the possibility of cross-contamination between samples.

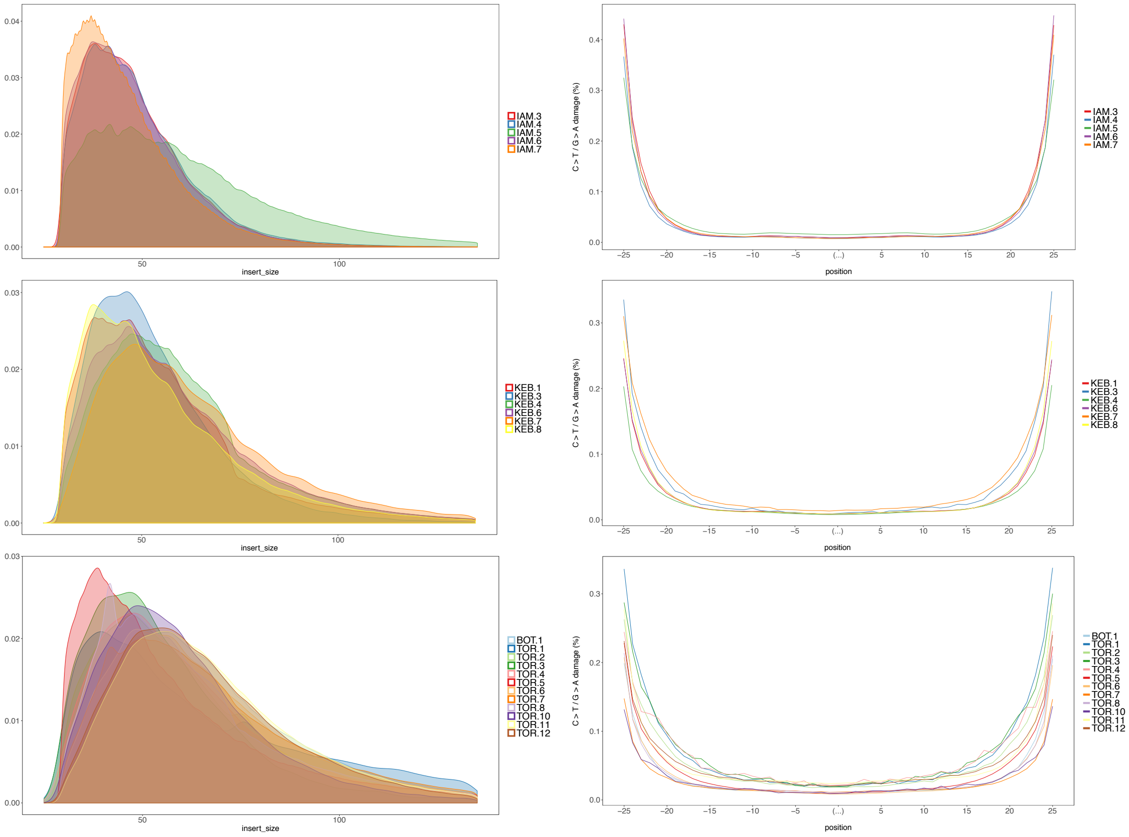


Figure S3.1 - Insert length distribution and damage pattern plots

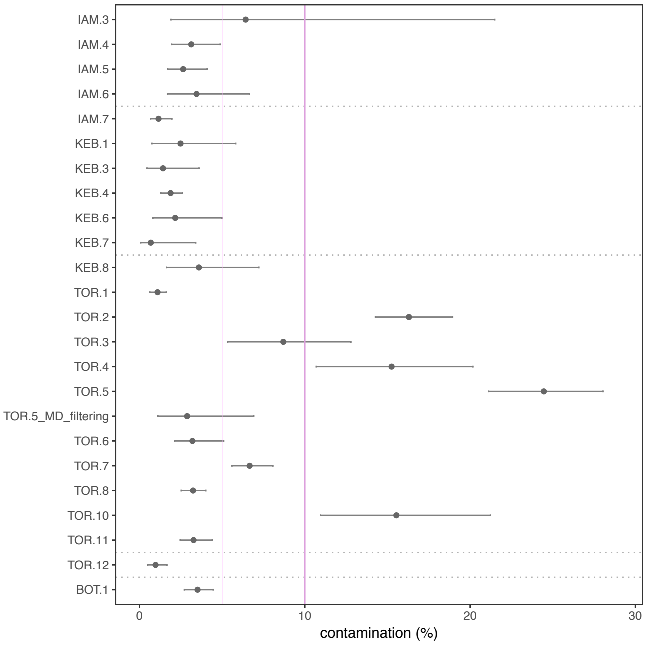


Figure S3.2 - Contamination rates for all samples, including TOR.5 after post-mortem damage filtering.

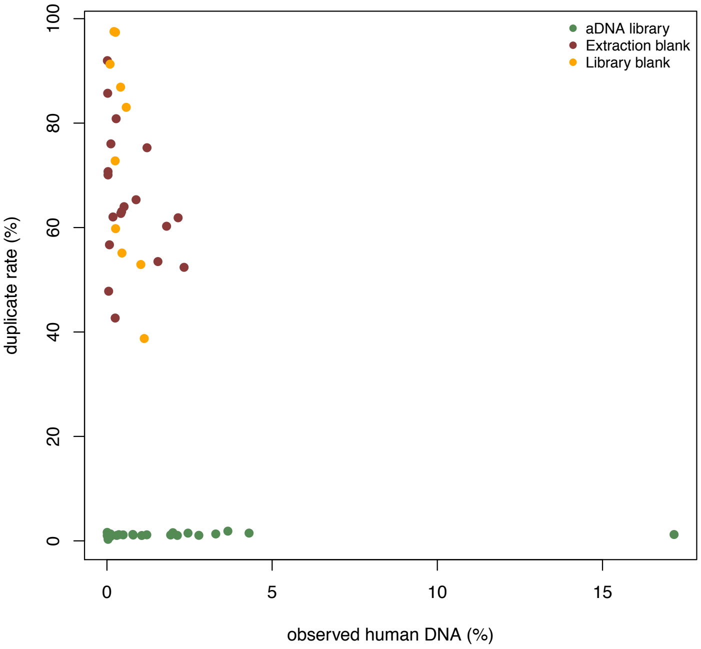


Figure S3.3 - Relationship between observed human DNA and library complexity in aDNA samples and contamination blanks

Supplementary Note 4: Mitochondrial DNA analysis

Rosa Fregel

For analyzing the mtDNA genome, reads were re-mapped directly to the revised Cambridge Reference Sequence (rCRS)[36] using the same filtering criteria as before. Only samples with an average mtDNA depth of at least 10X were considered. For generating the consensus sequence, we used SAMtools and BCFtools[32] version 0.1.19. A list of variants was then obtained using SAMtools mpileup, with a minimum depth of 5. Haplogroups were determined with HaploGrep[37] version 2.0, using PhyloTree build 17 version (http://www.phylotree.org)[38]. MtDNA haplotypes were manually curated by visual inspection using Tablet[39] version 1.16.09.06.

Previoulsy published sequences belonging to haplogroups of interest were obtained from NCBI (https://www.ncbi.nlm.nih.gov). MtDNA genomes were then aligned to the rCRS with BioEdit Sequence Alignment program[40], transformed into haplotypes using HaploSearch[41] and classified into haplogroups using HaploGrep[37]. Finally, we used Network version 5 (http://www.fluxus-engineering.com) to build reduced-median networks[42] to determine the phylogeographical assignation of the mtDNA lineages. Indels around nucleotides 309, 522, 573 and 16193, and hotspot mutations (e.g. 16519) were excluded from phylogenetic analysis.

Archaeological site	Sample ID	Reads number	Depth	Coverage	Haplogroup
IAM	IAM.3	1,737	5.1	96.87%	M1b1
	IAM.4	17,423	51.9	100.0%	U6a1b
	IAM.5	20,343	82.1	99.98%	U6a1b
	IAM.6	7,445	20.8	99.95%	U6a7
	IAM.7	41,890	129.4	99.97%	U6a3
KEB	KEB.1	5,380	17.5	99.88%	X2b
	KEB.3	6,058	20.2	99.95%	K1a1b1
	KEB.4	30,884	135.4	100.0%	K1a1b1
	KEB.6	4,879	14.0	99.93%	K1a4a1
	KEB.7	4,467	12.3	99.83%	T2b3
	KEB.8	8,359	22.7	99.92%	X2b
TOR	TOR.1	31,500	169.6	100.0%	T2c1d
	TOR.5	6,148	17.6	99.92%	J2b1a
	TOR.6	17,004	53.1	99.99%	T2b3
	TOR.7	30,678	126.4	100.0%	T2b3
	TOR.8	45,708	234.4	100.0%	K1a1
	TOR.11	28,654	124.5	100.0%	K1a2a
	TOR.12	32,628	173.7	100.0%	J2b1a
BOT	BOT.1	29,300	123.6	100.0%	K1a4a1

Table S4.1 - MtDNA results

We recovered 18 complete mtDNA genomes from IAM (n=5), KEB (n=6), TOR (n=7) and BOT (n=1). The average depth was 80.8 ± 31.8X, with a minimum depth of 5.1X (Table S4.1). Although sample IAM.3 had coverage lower than 10X, we include it in the analysis. Haplogroup classification is shown in Tables S7 and S8, and in Figure S7.

All samples from IAM (~5,000 BCE) belong to haplogroup U6 (Table S4.1, Figure S4.1), except for IAM.3 that is tentatively classified as haplogrupo M1. Haplogroup U6 is considered a North African autochthonous lineage, with a coalescence age of 42,000 - 52,000 BP. U6 is related to the back migration to Africa from Eurasia in Paleolithic times[43,44]. Haplogroup U6 is relatively frequent in modern day North African populations (e.g. 7.5% in Morocco according to Pennarun et al.[45]), indicating a maternal continuity in the region since Paleolithic times. Although the result from IAM.3 is not conclusive due to low coverage, it is congruent with North African mtDNA history. Haplogroup M1 is also related to the Paleolithic black migration to Africa and considered a North African autochthonous lineage, although its coalescence age (~26,000 BP) is younger than that of U6[45]. Haplogroup M1 is also present today in North Africa (e.g. 2.1% in Morocco[45]), and it is also in agreement with population continuity in the region. The presence in IAM of two prominent North African autochthonous lineages such as U6 and M1 supports maternal continuity in the area and implies an Eurasian origin for IAM people.

Surprisingly, the mtDNA composition of KEB (~3,000 BCE) is completely different from IAM (Table S4.1, Figure S4.1), and presents lineages previously observed in Early and Middle Neolithic samples in Europe and the Near East, such us X, K and T2[46,47]. KEB.1 and KEB.8 have the same mtDNA lineage belonging to X2b haplogroup. X2 is a mtDNA lineage that is believed to have arisen ~20,900 years ago[48], but its place of origin is uncertain. The highest frequency of X2 is observed today in the Caucasus (4.3%), although it is also relatively frequent in the Near East (2.9%) and the Mediterranean Europe (2.5%)[49]. Within the Near East, Haplogroup X2 is especially frequent in Druze, a religious group considered to have retained ancient Neolithic genetic ancestry by cultural isolation and endogamous practices[50]. X2 is present also in modern populations of the Maghreb region[49], with an average frequency of 1.21%. KEB.3 and KEB.4 lineages belong both to K1a1b1 and KEB.6 belongs to K1a4a1. K is considered a typical European Neolithic lineage, and it has been thoroughly observed in Neolithic populations (e.g. Mathieson et al.[51]), with a frequency of ~10%[52]. Haplogroup K originated ~38,000 years ago, and independently diversified in both the Near East and Europe[53]. Haplogroup K is common in modern populations of the Near East (~10%) and Europe (~7%), and it has also been observed in North Africa (4.8%)[54]. Finally, KEB.7 belongs to T2b3 and it is identical to two genomes recovered from TOR. T2 is another of the mtDNA haplogroups related to Neolithic populations in Europe (see Mathieson et al.[51]), with an average frequency of 12%[52]. Haplogroup T2 is supposed to have originated ~21,000 years ago in the Near East[55]. This haplogroup represents ~8% of the European mtDNA lineages and ~5% of the Near Easterns[55] and North Africans[54]. The striking difference in mtDNA composition between IAM and KEB is suggestive of a population replacement, or at least, to an admixture event previous to the Late Neolithic period in Morocco, bringing to the area mtDNA lineages related to Neolithic populations in the Near East and Europe.

MtDNA lineages observed in the southern Iberian sites belong to K1a, T2b, T2c and J2b haplogroups (Table S4.1, Figure S4.1), coinciding with the expected for a Neolithic European population. BOT.1 mtDNA lineage belongs to K1a4a1, the same haplogroup as KEB.6. Samples TOR.8 and TOR.11 belong to two other K1a subhaplogroups, different from the ones observed in KEB: TOR.8 belongs to K1a1* and TOR.11 to K1a2a haplogroup. Samples TOR.6 and TOR.7 belong to the same T2b3 haplogroup observed in KEB. Sample TOR.1 belongs to other T2 subclade, T2c1d sub-haplogroup. Samples TOR.5 and TOR.12 belong to J2b1a haplogroup. J is other of the common Neolithic lineages, with an average frequency of 12% at the time[52]. J2 is also a Near Eastern lineage that originated ~37,000 years ago[55]. In summary, the mtDNA profile of BOT and TOR resembles other results obtained from European Neolitihic sites[46,51,56-61].

	Sample ID	Haplogroup	HaploGrep quality	Whole-genome haplotype
IAM	IAM.3	M1b1?	71.40%	73G 94A 152C 195C 263G 569T 750G 813G 1438G 2706G 3107N 3514T 3559T 4769G 4853C 4936T 5385T 6446A 6680C 7028T 8027A 8701G 8860G 10223T 11719A 12705T 13111C 14569T 14766T 15301A 15326G 16129A 16223T 16311C 16519C
	IAM.4	U6a1b	94.30%	73G 263G 750G 1438G 2706G 2887C 3107N 3348G 4769G 7028T 7805A 8670G 8860G 9165C 11467G 11719A 12308G 12372A 14179G 14766T 14927G 15326G 16172C 16219G 16235G 16274A 16278T 16311C 16362C
	IAM.5	U6a1b	92.22%	73G 263G 750G 1438G 2706G 2887C 3107N 3348G 4769G 7028T 7805A 8670G 8860G 9165C 11467G 11719A 12308G 12372A 14179G 14766T 14927G 15326G 16172C 16219G 16235G 16274A 16278T 16311C 16362C
	IAM.6	U6a7	92.53%	73G 263G 750G 955.1C 1438G 2706G 2885C 3107N 3348G 4769G 6986C 7028T 7805A 8860G 8939C 9947A 11467G 11719A 12308G 12372A 14179G 14766T 15043A 15326G 15914G 16145A 16172C 16219G 16278T
	IAM.7	U6a3	94.59%	73G 263G 750G 1438G 2706G 3107N 3348G 4769G 7028T 7805A 8860G 11467G 11719A 11887A 12308G 12372A 14179G 14766T 14891T 15326G 15790T 16172C 16183C 16189C 16219G 16278T 16304C
KEB	KEB.1	X2b+226	99.05%	73G 153G 195C 225A 226C 263G 750G 1438G 1719A 2706G 3107N 4769G 6221C 6371T 7028T 8393T 8860G 11719A 12705T 13708A 13966G 14470C 14766T 15326G 15927A 16183C 16189C 16223T 16274A 16278T 16519C
	KEB.3	K1a1b1	97.86%	73G 114T 152C 263G 497T 750G 1189C 1438G 1811G 2706G 3107N 3480G 4769G 7028T 8860G 9055A 9698C 10398G 10550G 11299C 11467G 11470G 11719A 11914A 12308G 12372A 14167T 14766T 14798C 15326G 15924G 16093C 16224C 16311C 16519C
	KEB.4	K1a1b1	97.86%	73G 114T 152C 263G 497T 750G 1189C 1438G 1811G 2706G 3107N 3107G 3480G 4769G 7028T 8860G 9055A 9698C 10398G 10550G 11299C 11467G 11470G 11719A 11914A 12308G 12372A 14167T 14766T 14798C 15326G 15924G 16093C 16224C 16311C 16519C
	KEB.6	K1a4a1	98.64%	73G 263G 497T 750G 1189C 1438G 1811G 2706G 3107N 3480G 4769G 6260A 7028T 8860G 9055A 9698C 10398G 10550G 11299C 11467G 11485C 11719A 11840T 12308G 12372A 13740C 14167T 14766T 14798C 15326G 16224C 16256T 16311C 16519C
	KEB.7	T2b3+151	100.0%	73G 151T 263G 709A 750G 930A 1438G 1888A 2706G 3107N 4216C 4769G 4917G 5147A 7028T 8697A 8860G 10463C 10750G 11251G 11719A 11812G 13368A 14233G 14766T 14905A 15326G 15452A 15607G 15928A 16126C 16294T 16296T 16304C 16519C
	KEB.8	X2b+226	99.05%	73G 153G 195C 225A 226C 263G 750G 1438G 1719A 2706G 3107N 4769G 6221C 6371T 7028T 8393T 8860G 11719A 12705T 13708A 13966G 14470C 14766T 15326G 15927A 16183C 16189C 16223T 16274A 16278T 16519C
TOR	TOR.1	T2c1d+152	96.67%	73G 146C 152C 263G 279C 709A 750G 1438G 1888A 2706G 3107N 4216C 4769G 4917G 5187T 6261A 7028T 7873T 8697A 8860G 10463C 10822T 11251G 11719A 11812G 13368A 14233G 14766T 14905A 15326G 15452A 15607G 15928A 16126C 16292T 16294T 16519C
	TOR.5	J2b1a	97.89%	73G 150T 152C 263G 295T 489C 750G 1438G 2706G 3107N 4216C 4769G 5633T 7028T 7476T 7647C 8860G 10172A 10398G 11251G 11719A 12612G 13708A 14766T 15257A 15326G 15452A 15812A 16069T 16126C 16193T 16278T 16519C
	TOR.6	T2b3+151	100.0%	73G 151T 263G 709A 750G 930A 1438G 1888A 2706G 3107N 4216C 4769G 4917G 5147A 7028T 8697A 8860G 10463C 10750G 11251G 11719A 11812G 13368A 14233G 14766T 14905A 15326G 15452A 15607G 15928A 16126C 16294T 16296T 16304C 16519C
	TOR.7	T2b3+151	100.0%	73G 151T 263G 709A 750G 930A 1438G 1888A 2706G 3107N 4216C 4769G 4917G 5147A 7028T 8697A 8860G 10463C 10750G 11251G 11719A 11812G 13368A 14233G 14766T 14905A 15326G 15452A 15607G 15928A 16126C 16294T 16296T 16304C 16519C
	TOR.8	K1a1	97.95%	73G 114T 263G 497T 750G 1189C 1438G 1811G 2706G 3107N 3480G 4769G 7028T 8860G 9055A 9698C 10398G 10550G 11299C 11467G 11719A 11914A 12308G 12372A 14167T 14766T 14798C 15326G 16093C 16224C 16311C 16519C
	TOR.11	K1a2a	97.27%	73G 263G 497T 750G 1189C 1438G 1811G 2706G 3107N 3480G 4769G 5773A 7028T 8860G 9055A 9698C 10398G 10550G 11025C 11299C 11467G 11719A 12308G 12372A 14167T 14766T 14798C 15326G 16224C 16311C 16519C
	TOR.12	J2b1a	97.89%	73G 150T 152C 263G 295T 489C 750G 1438G 2706G 3107N 4216C 4769G 5633T 7028T 7476T 8860G 10172A 10398G 11251G 11719A 12612G 13708A 14766T 15257A 15326G 15452A 15812A 16069T 16126C 16193T 16278T
BOT	BOT.1	K1a4a1	95.43%	73G 263G 497T 750G 1189C 1438G 1811G 2706G 3107N 3480G 4769G 6260A 7028T 8860G 9055A 9698C 10398G 10550G 11016A 11299C 11467G 11485C 11719A 11840T 12308G 12372A 13740C 14167T 14766T 14798C 15326G 16224C 16265C 16311C 16519C

Table S4.2 - Detailed haplotype information

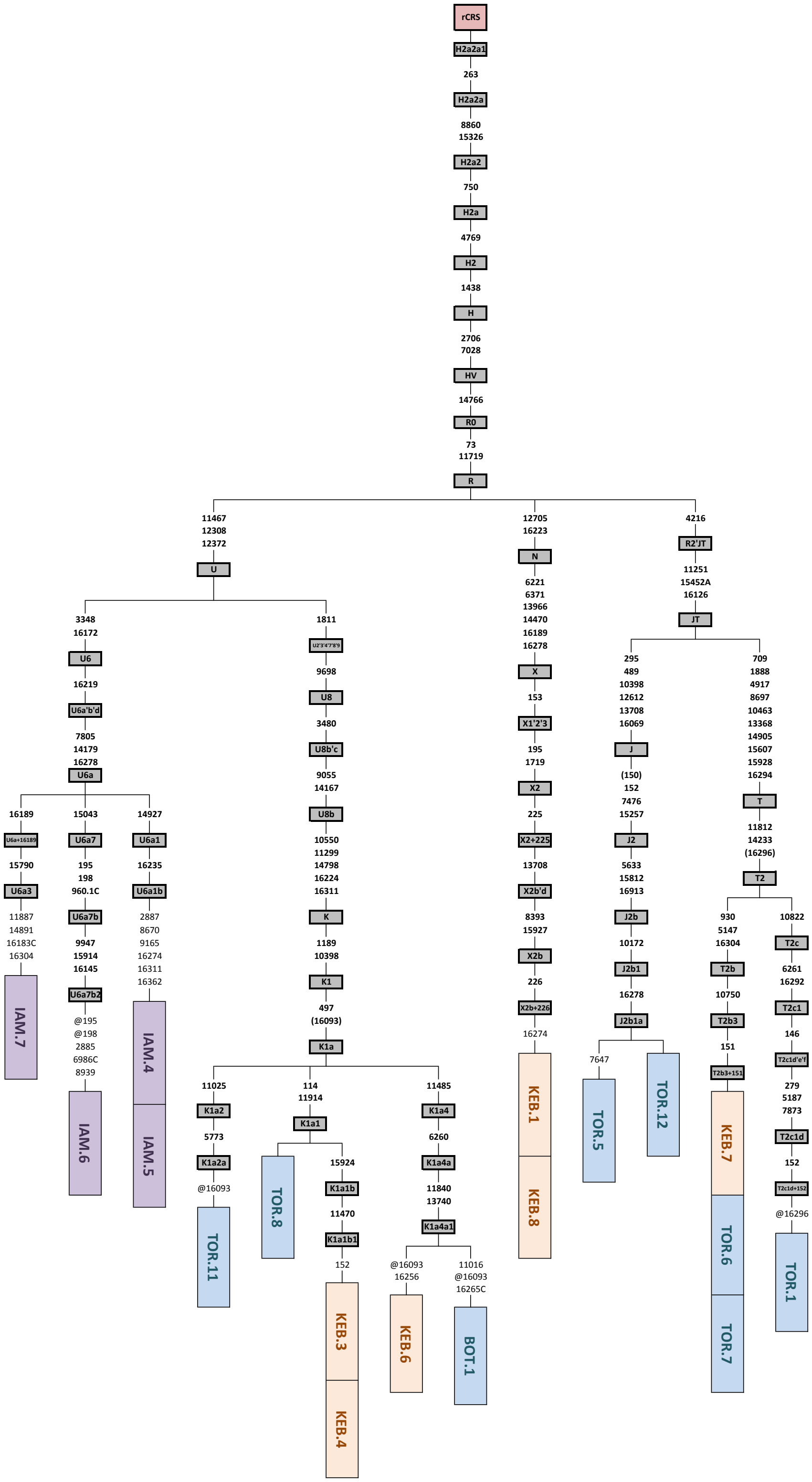


Figure S4.1 - Summarized mtDNA tree

4.1. Haplogroup M1

Sample IAM.3 is most probably classified within haplogroup M1b1 (HaploGrep quality = 71.4%). IAM.3 carries the mutation defining M1b (13111 mutation, present in 3 out of 3 reads) and two out of the four mutations defining M1b1 (4936, 5/6 reads; 8868 1/2 reads). We have also observed that four different reads cover position 813 and all of them carry the A > G mutation, indicating that the sample could also be M1a. However, the other mutation in the M1a branch (6671) is not observed in the only read covering that position. Given that several mutations support the classification of the sample within the M1b clade and, that 813 mutation could have happened on the M1 branch and been lost later in M1b, we consider IAM.3 most probably belongs to a lineage ancestral to M1b1 (Table S4.2). Today, the higher frequency of M1b is observed precisely in Morocco (2.3%)[45]. M1b is considered to have appeared ~20,000 BP in northwest Africa, coinciding with the flourishing of the Iberomaurusian culture. M1b1 arose also in northwest Africa ~10,000 BP and it has been related to the Capsian culture[45].

4.2. Haplogroup U6

Samples IAM.4 and IAM.5 present the same mtDNA genome sequence. They are classified within U6a1b haplogroup. IAM.4 and IAM.5 have several private mutations (2887 8670 9165 16274 16311 16362), but none of them were observed in any of the previously published U6 sequences, so both samples remain classified as U6a1b* (Figure S4.2). This haplogroup is considered to have originated in the Maghreb area around 17,000 years ago, with a later spread to the Mediterranean shores of Europe[44]. Based on our phylogenetic analysis, U6a1b derived haplogroups (U6a1b1, U6a1b2, U6a1b3 and U6a1b4) are today distributed both in North Africa (Algeria and Morocco) and South Europe (Italy, Portugal and Spain).

Sample IAM.7 sample was classified as U6a3*. Our phylogenetic analysis determined that, in spite of IAM.7 having three private mutations (11887 14891 16304), we did not observe any of them in previously published mtDNA genomes (Figure S4.3). U6a3 haplogroup is believed to have arisen around 18,800 years ago[44], although its place of origin is uncertain. Different subclusters of U6a3 are distributed in both in Europe and the Maghreb (U6a3a, U6a3b and U6a3e), the Middle East (U6a3d) and also in West Africa (U6a3c and U6a3f) (Figure S4.3).

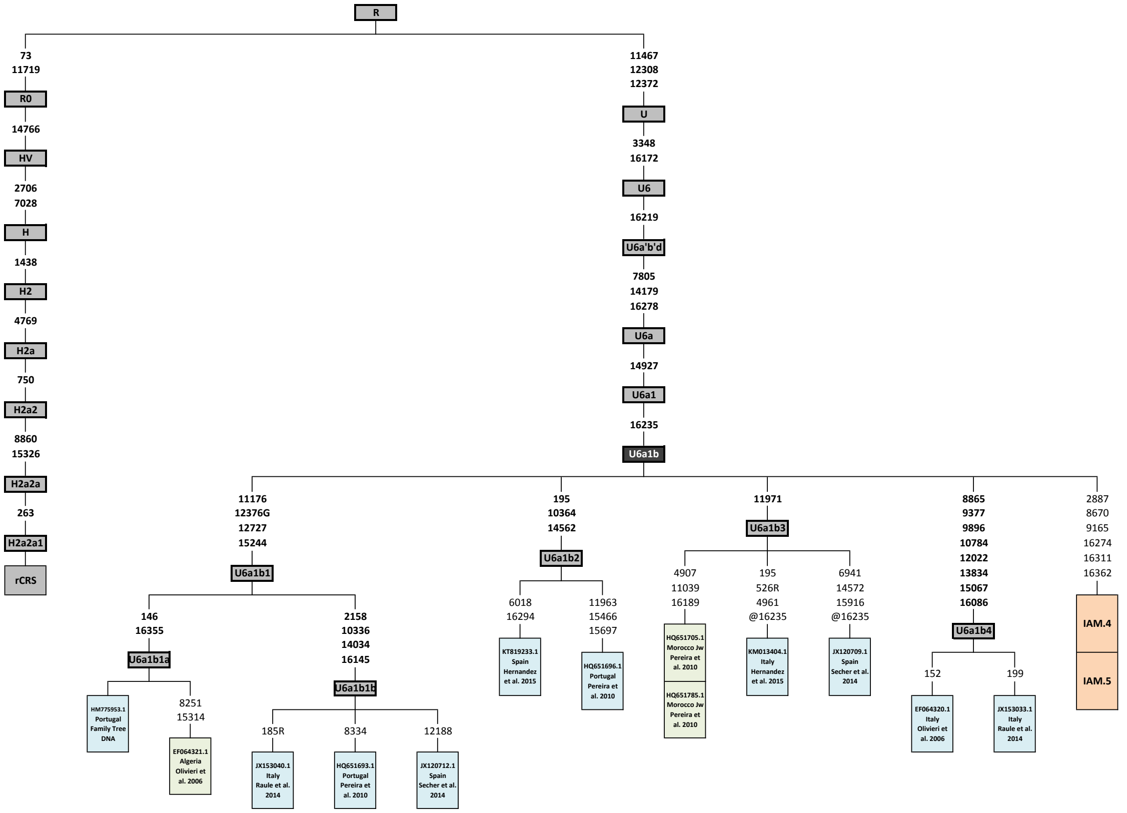

R
73
11719
R0
14766
HV
2706
7028
H
1438
H2
4769
H2a
750
H2a2
8860
15326
H2a2a
263
H2a2a1
rCRS
11467
12308
12372
U
3348
16172
U6
16219
U6a'b'd
7805
14179
16278
U6a
14927
U6a1
16235
U6a1b
11176
12376G
12727
15244
U6a1b1
146
16355
U6a1b1a
HM775953.1 Portugal Family Tree DNA
8251
15314
EF064321.1 Algeria Olivieri et al. 2006
2158
10336
14034
16145
U6a1b1b
185R
JX153040.1 Italy Raule et al. 2014
8334
HQ651693.1 Portugal Pereira et al. 2010
12188
JX120712.1 Spain Secher et al. 2014
195
10364
14562
U6a1b2
6018
16294
KT819233.1 Spain Hernandez et al. 2015
11963
15466
15697
HQ651696.1 Portugal Pereira et al. 2010
11971
U6a1b3
4907
11039
16189
HQ651705.1 Morocco Jw Pereira et al. 2010
HQ651785.1 Morocco Jw Pereira et al. 2010
195
526R
4961
@16235
KM013404.1 Italy Hernandez et al. 2015
6941
14572
15916
@16235
JX120709.1 Spain Secher et al. 2014
8865
9377
9896
10784
12022
13834
15067
16086
U6a1b4
152
EF064320.1 Italy Olivieri et al. 2006
199
JX153033.1 Italy Raule et al. 2014
2887
8670
9165
16274
16311
16362
IAM.4
IAM.5

Figure S4.2 - Summarized phylogenetic tree of the haplogroup U6a1b

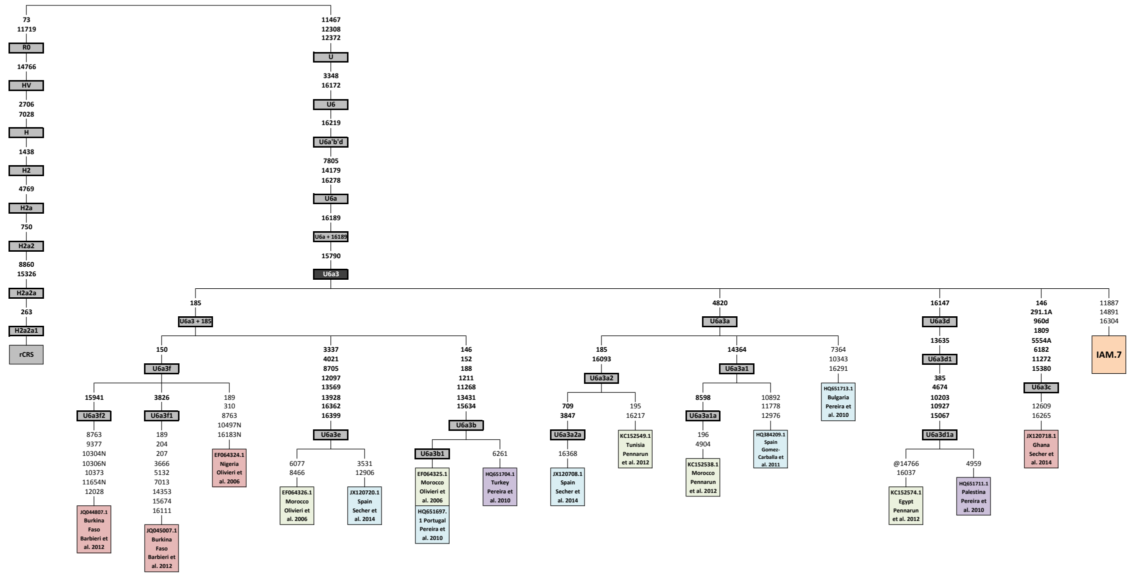


Figure S4.3 - Summarized phylogenetic tree of the haplogroup U6a3

Sample IAM.6 mtDNA lineage was classified within U6a7b haplogroup. Phylogeographic analysis indicated that U6a7b most probably originated in the Maghreb area ~24,000 years ago[44]. IAM.6 shares three of its private mutations with a modern sample from Tunisia[45], defining the new haplogroup U6a7b2 (Figure S4.4), further confirming a temporal continuity in the North African region. The other derived cluster, U6a7b1 is distributed again in North Africa, Europe and the Canary Islands, whose indigenous population has a North African Berber origin[62,63].

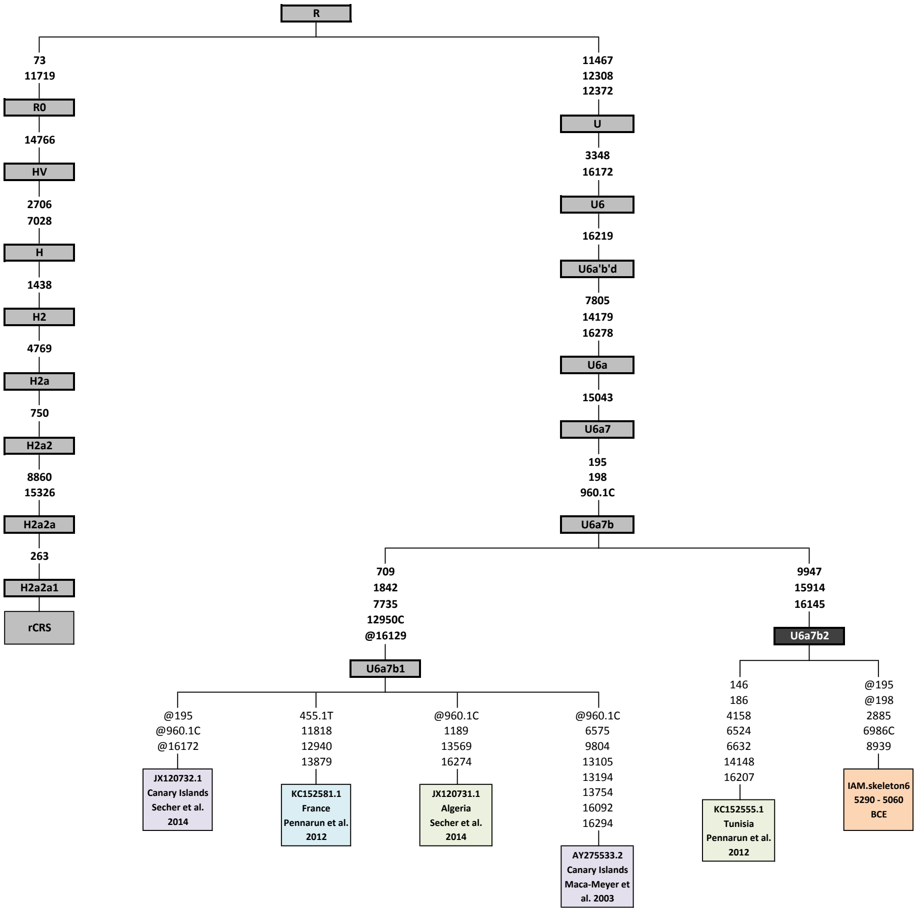


Figure S4.4 - Phylogenetic tree of the haplogroup U6a7b

The only previously published U6 mitogenome obtained from an ancient sample was that of the Pestera Muierii individual. This sample was excavated in Romania and it has been dated around 35,000 BP. The Pestera Muierii mtDNA lineage was identified as U6*, confirming a Eurasian origin for the whole haplogroup[43]. U6 haplogroup was also detected by traditional aDNA techniques (PCR amplification of HVRI and Sanger sequencing) in 23,000 - 10,800 years-old samples from the Taforalt site in Morocco[64], confirming U6 presence in North Africa during the Iberomaurusian period. However, given the age of the samples and the limitations of the HVRI for haplogroup affiliation, the use of whole-genome sequencing would be more appropriate for confirming this result.

4.3. Haplogroup X2

X2b most probably originated in Europe ~20,000 years ago, although a Near Eastern origin is also possible[65]. On our phylogenetic tree (Figure S4.5), KEB.1 and KEB.8 are clustered within X2b+226, with only one private mutation (16274).

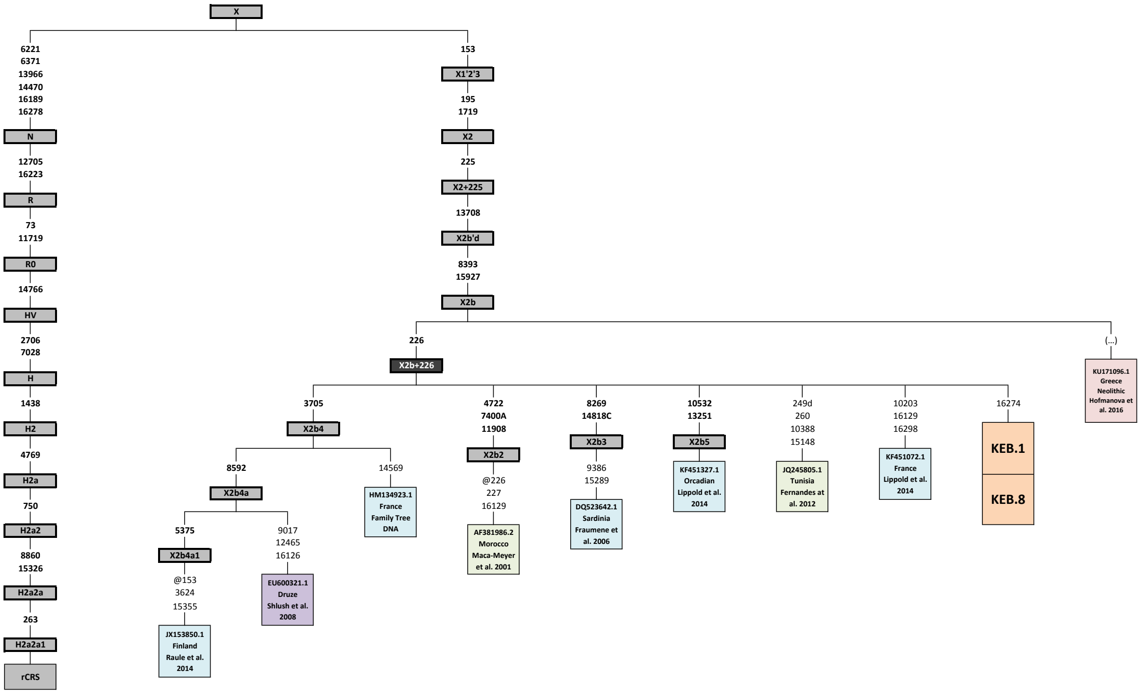

X
6221
6371
13966
14470
16189
16278
N
12705
16223
R
73
11719
R0
14766
HV
2706
7028
H
1438
H2
4769
H2a
750
H2a2
8860
15326
H2a2a
263
H2a2a1
rCRS
153
X1'2'3
195
1719
X2
225
X2+225
13708
X2b'd
8393
15927
X2b
226
X2b+226
3705
X2b4
8592
X2b4a
5375
X2b4a1
@153
3624
15355
JX153850.1 Finland Raule et al. 2014
9017
12465
16126
EU600321.1 Druze Shlush et al. 2008
14569
HM134923.1 France Family Tree DNA
4722
7400A
11908
X2b2
@226
227
16129
AF381986.2 Morocco Maca-Meyer et al. 2001
8269
14818C
X2b3
9386
15289
DQ523642.1 Sardinia Fraumene et al. 2006
10532
13251
X2b5
KF451327.1 Orcadian Lippold et al. 2014
249d
260
10388
15148
JQ245805.1 Tunisia Fernandes at al. 2012
10203
16129
16298
KF451072.1 France Lippold et al. 2014
16274
KEB.1
KEB.8
(...)
KU171096.1 Greece Neolithic Hofmanova et al. 2016

Figure S4.5 - Summarized phylogenetic tree of haplogroup X2b

This haplogroup encompasses modern lineages from Europe, the Near East and also North Africa (Figure S4.4). In fact, X2 frequency in the modern population of Morocco is ~0.8%[49]. X2b have been directly observed in European Neolithic samples. Interestingly, an Early Neolithic sample from Greece (Revenia site; 6,438–6,264 BCE) has been classified as X2b[61]. Within the X2b+226 clade, there are an Early Neolithic sample from Hungary (Garadna site; 5,281–5,132 BCE)[66] and a Chalcolithic sample from Spain (El Mirador Cave; 3,010–2,975 BCE)[51]. HVRI and HVRII data also indicates the presence of X2b+226 in two Neolithic samples from Germany (Salzmünde site; 3,400–3,025 BCE and 4,100–3,950 BCE, respectively)[47]. All these results, imply the affiliation of X2b with Neolithic and Chalcolithic populations of Europe, and suggest the introgression of those populations into North Africa, some time before KEB people inhabited Morocco.

4.4. Haplogroup T2

Samples KEB.7, TOR.6 and TOR.7 are classified as T2b3+151, with no private mutations (See Figure S4.1). T2b, dated ~10,000 years ago, is mainly distributed in Europe. T2b is considered to have dispersed within Europe in early Neolithic times[55], a fact that has been confirmed by direct analysis of aDNA. T2b has been extensively observed in Neolithic remains from Europe and, more recently, the Near East, including sites from Croatia[67], France[68], Germany[47,51,56,58], Hungary[67], Italy[69], Poland[70,71], Spain[72,73], Turkey[51] and Sweden[74]. This haplogroup has also been detected in Chalcolithic and Bronze Age sites in the Czech Republic[75], Denmark[75], Germany[47], Hungary[67,75], Russia[51], Spain[76] and Ukraine[77]. T2b was also unexpectedly discovered in a Mesolithic sample from Sweden[74], pointing to the assimilation of Neolithic lineages into the Mesolithic Pitted Ware culture.

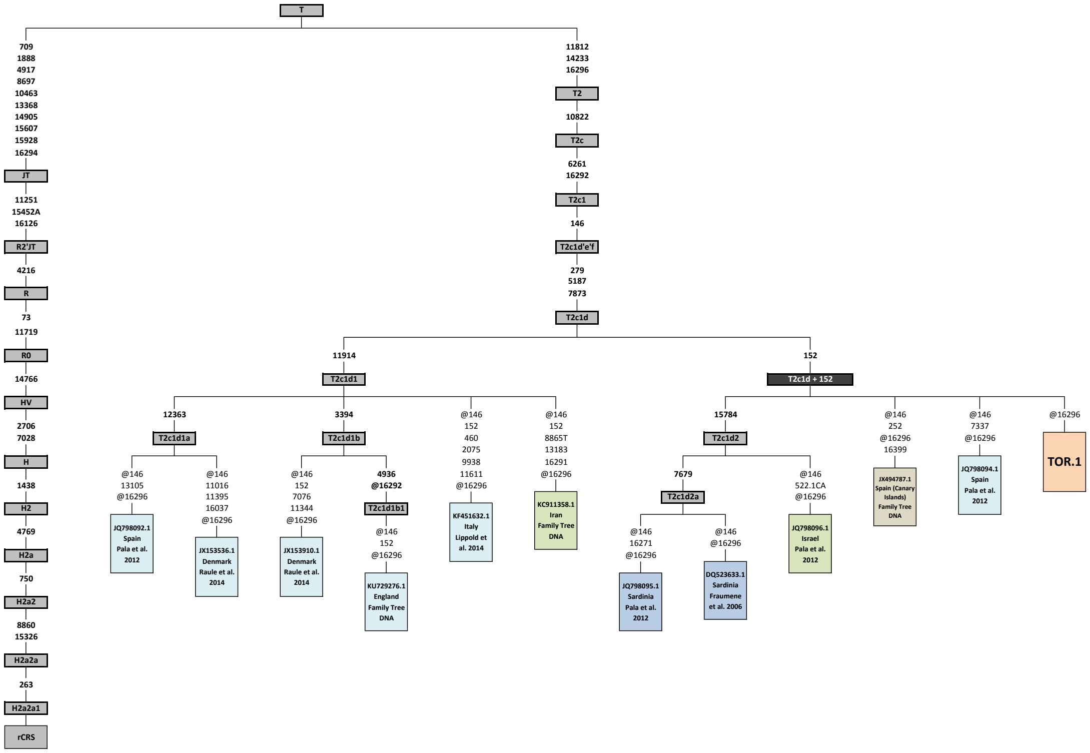


Figure S4.6 - Summarized T2c1d phylogenetic tree

TOR.1 mtDNA is classified as T2c1d+152, with no private mutations (16296 mutation is unstable within the T2 clade). T2c1 dispersed into Europe ~10,000 years ago and it is common in the Levant and in Mediterranean Europe. T2c most probably had an origin in the Near East ~20,000 years ago[55], which has been confirmed by its presence in an Early Neolithic sample from Tepe Abdul Hosein site in Iran (8,204–7,755 BCE)[78]. T2c has also been observed in several European Neolithic sites from Hungary[67], Germany[46,51] and Spain[46,51]. Concretely, a sample from the Early Neolithic site of El Trocs in Spain (5,295–5,066 BCE)[51] and a Linear Pottery culture sample from the Stuttgart site in Germany (~5,000 BCE)[57] have been classified as T2c1d. In our phylogenetic tree (Figure S4.6), it can be observed how T2c1d is distributed in Europe and Near East. Within the T2c1d+152 branch, sample TOR.1 clusters with both European and Near Eastern lineages, including samples for Sardinia and the Canary Islands. This is worth mentioning as Sardinia is considered an isolated European population, which retained a higher Neolithic component. The indigenous people of the Canary Islands, as mentioned before, are related to Berber populations in North Africa. The Canaries were later conquered and colonized by Europeans, leading to extensive admixture in the modern Canarian population, so the exact geographical ascription of this lineage (Europe vs. North Africa) is uncertain.

The presence of T2b3 and T2c1d in TOR is expected, as both haplogroups are frequent in Neolithic Europe. However, the discovery of T2b3 in KEB is suggestive of the influence of Neolithic Anatolian/European people in North Africa.

4.5. Haplogroup J2

Samples TOR.5 and TOR.12 are classified as J2b1a. TOR.5 has one private mutation, not observed in any modern sample, whereas TOR.12 present the basal haplotype of J2b1a (Figure S4.1). Haplogroup J2b1 is considered a European cluster, with a coalescence time of ~15,000 years. J2b1 is mainly distributed in the Mediterranean and Atlantic Europe. In our phylogenetic tree (Figure S4.7), J2b1a is distributed throughout Europe, including several Sardinian lineages. Based on HVRI sequencing, J2b1a (defined by 16069 16126 16193 16278) has been observed in Neolithic populations from France[79] and Sweden[74]. As happens with T2b3 and T2c1d, the presence of J2b1a in TOR indicates its affiliation with other Neolithic European populations.

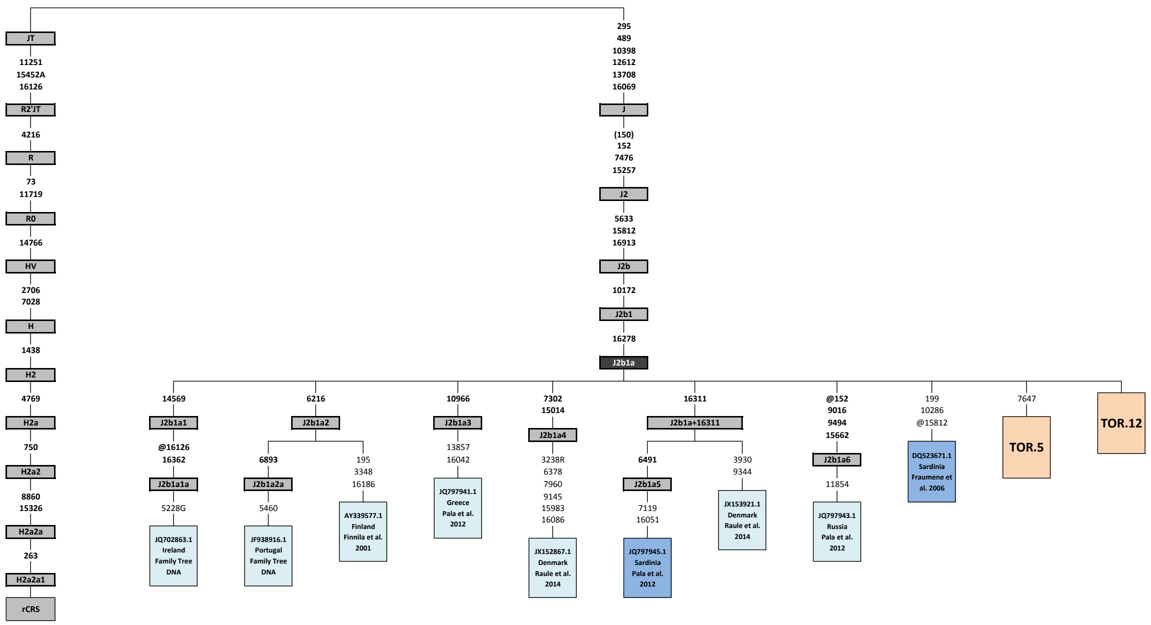


Figure S4.7 - Summarized J2b1a phylogenetic tree

4.6. Haplogroup K1

Six ancient samples from North Africa and southern Iberian are classified within K1a haplogroup. Haplogroup K1a is considered to have expanded ~20,000 years ago, both in the Near East and Europe[53]. Samples classified within K1a cluster and its subhaplogroups have been described from different Neolithic sites in Europe and in the Near East. K1a lineages has been observed in most of the screened archaeological sites, including those of France[68,80], Germany[46,47,51], Hungary[51,67], Poland[70,81], Portugal[82], Spain[51,56,73,83] and Sweden[84]. Haplogroup K1a has also been detected in Neolithic samples from Iran[46], Israel[46] and Turkey[46,51,85]. As for haplogroup T2b, K1a lineages were present in Pitted Ware samples from Sweden[74], further reinforcing the introgression of Neolithic lineages into the Scandinavian Mesolithic people.

One sample from the southern Iberian Neolithic site, TOR.8, belongs to K1a1* haplogroup, with no private mutations. Haplogroup K1a1 originated ~18,000 years ago and it is mostly distributed in Europe[53] and also in the Near East[86].

Two North African Late Neolithic samples, KEB.3 and KEB.4, fall within the K1a1b1 cluster, both with the same private mutation (Figure S4.8). K1a1b1 subclade, with a coalescence age of ~11,000 years, is mostly restricted to Mediterranean Europe and North Africa

(Figure S4.8). The same exact haplogroup has been already observed in a Neolithic sample from Spain excavated in La Mina (3,900–3,600 BCE)[51]. The discovery of K1a1b1 in KEB demonstrate the presence of this lineage in North Africa ~3,000 BCE and rules out an exclusive historical introduction from Mediterranean Europe.

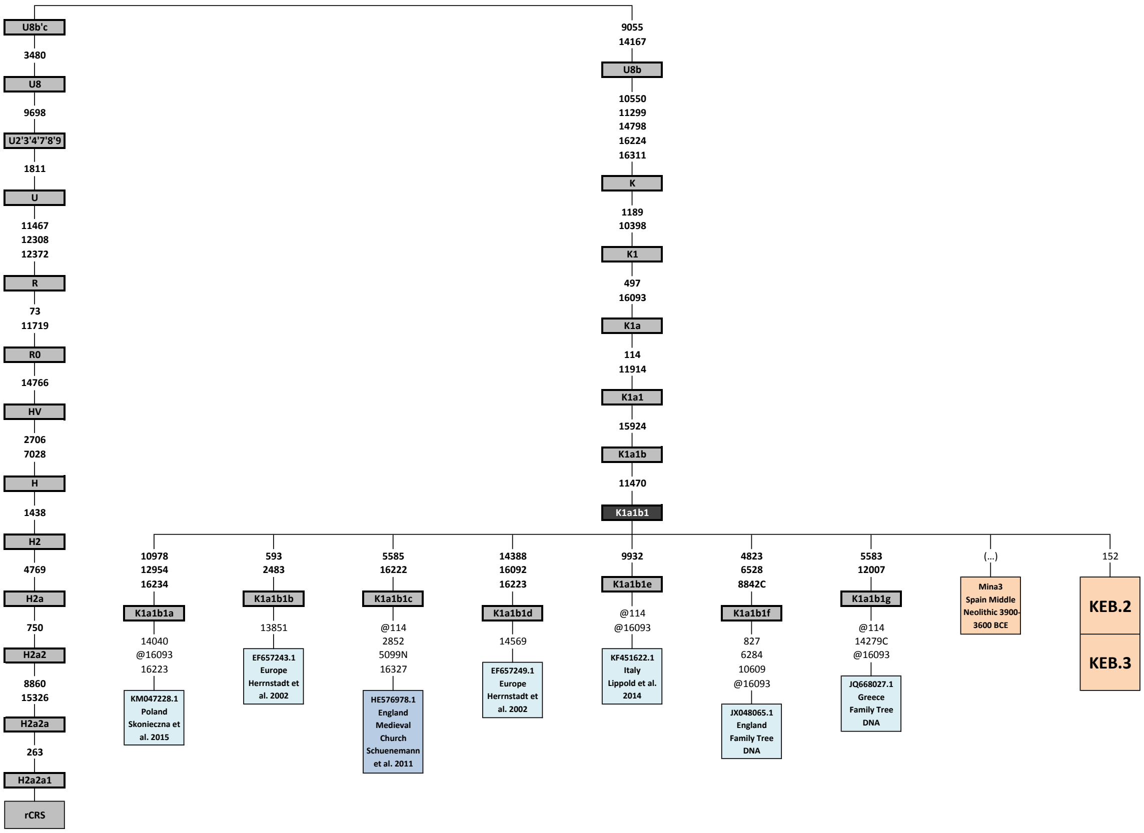


Figure S4.8 - Summarized K1a1b1 phylogenetic tree

Another southern Iberian Neolithic sample, TOR.11, is classified as K1a2a with no private mutations (16093 is unstable within K1a lineages) (Figure S4.9). As before, K1a2a has been observed both in the Near East and the Mediterranean area[53]. K1a2a has already been described in Neolithic samples from Spain (Cova Bonica, 5,470–5,360 BCE; Els Trocs, 5,177–5,068 BCE)[51,83].

Finally, one North African Late Neolithic and one Iberian Neolithic sample, KEB.6 and BOT.1, are classified as K1a4a1. KEB.6 mtDNA lineage has one private mutation, whereas BOT.1 has two private mutations, both different both that of KEB.6. None of these private mutations were observed in modern mtDNA lineages. Haplogroup K1a4 has been observed in modern populations of both in the Near East and South Europe, as well as North Africa[53]. In our phylogenetic tree, we can see how K1a4a1 sublineages are restricted to Europe, although we observed K1a4a1* lineages in the Near East and North Africa, clustering with the two samples from KEB and TOR (Figure S4.10). The same K1a4a1 haplogroup has been detected in other Neolithic sample from Spain (Cova de la Sarsa, 5,321–5,227 BC)[83].

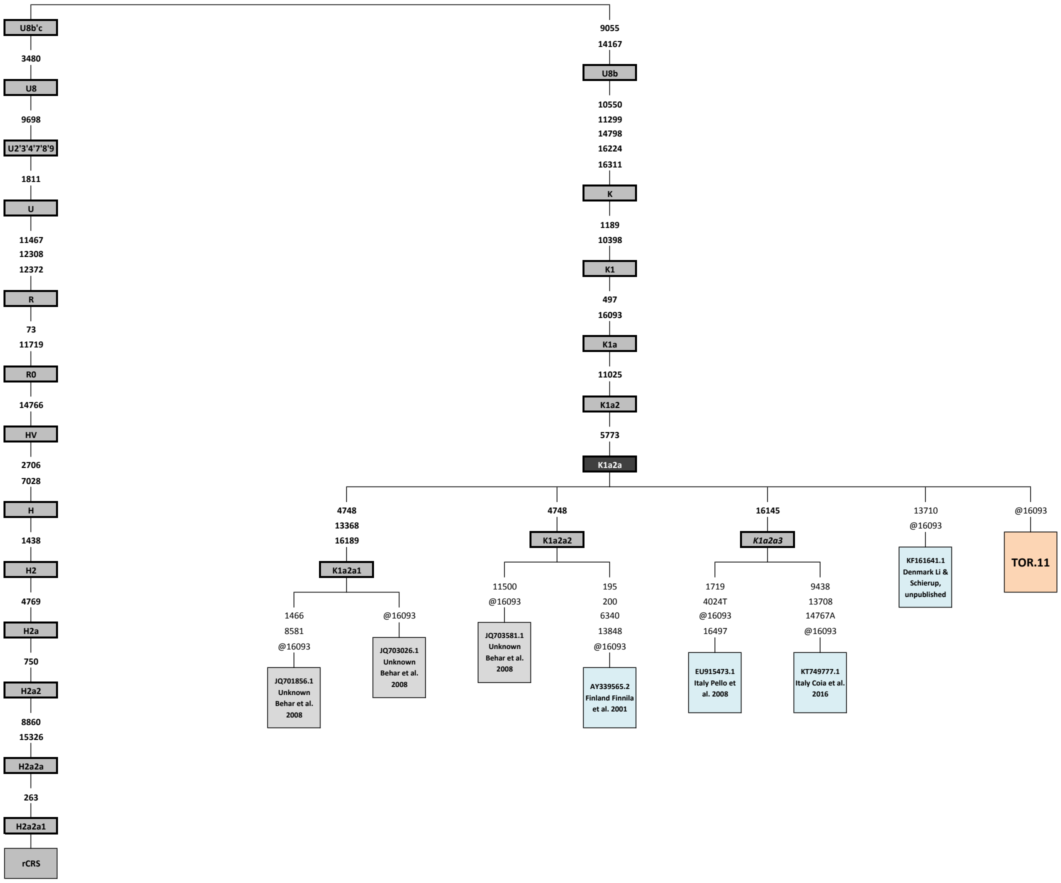


Figure S4.9 - K1a2a phylogenetic tree

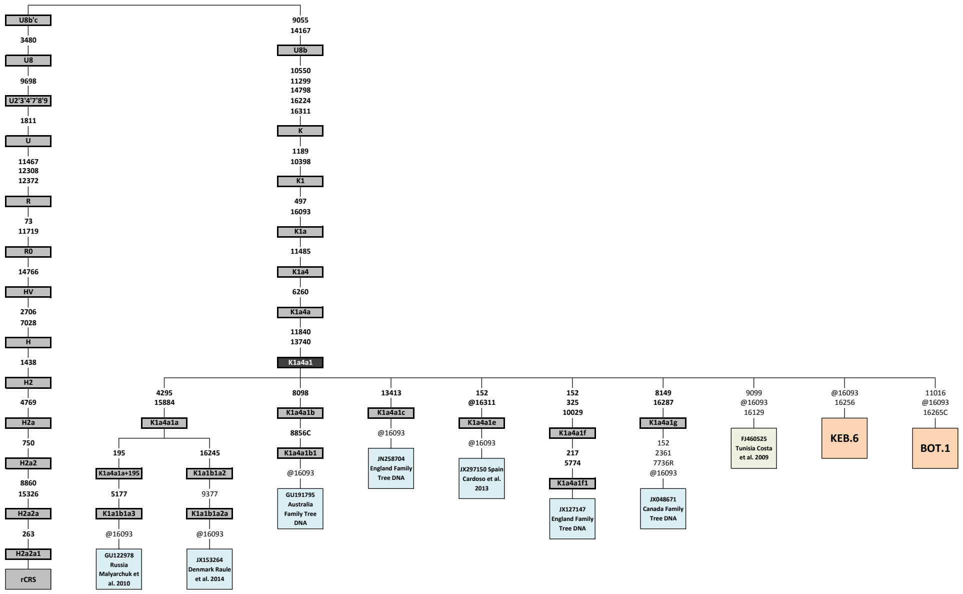


Figure S4.10 - Summarized K1a4a1 phylogenetic tree

Supplementary Note 5: Y-chromosome analysis

Rosa Fregel, Fernando Mendez and Peter A. Underhill

The molecular sex of the samples was identified using the ry estimate proposed by Skoglund et al.[87], based on the ratio of sequences aligning to the X and Y chromosomes.

When comparing ry estimates calculated using both shotgun and capture sequencing data, we observed that ry values were higher when including either WISC or MEGA-capture data. This artifact produced that samples previously identified as females using shotgun sequencing data, were not assigned to any gender when using the merged bam file, containing mainly captured reads. As the baits for capture were prepared using male genomic DNA, pseudo-autosomal regions could be enriched in female aDNA samples and mapped to the Y chromosome, producing a bias in the ry estimate value. Also, we observed that, due to the enrichment on repetitive regions caused by capture, female samples had reads mapping to repetitive regions close to the centromere and the ends of the Y chromosome. After removing pseudo-autosomal and repetitive regions, ry estimate calculations were consistent between pre-capture and post-capture calculations (Figure 5.1).

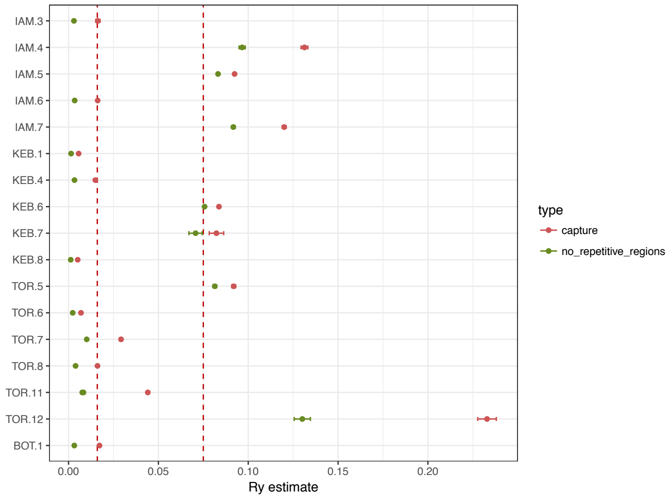


Figure S5.1 - Ry estimate using repetitive regions filtering (green) compared to unfiltered capture (red) sequencing data

Six samples were identified as males based on the ry estimate, two from each main archaeological site: IAM.4, IAM.5, KEB.6, KEB.7, TOR.5 and TOR.12. Y-chromosome analysis was carried out as in Schroeder et al.[88]. Briefly, we retrieved a list of known Y-chromosome variants from the Y-chromosome phylogenetic tree[89] built using 1000 Genomes (http://www.internationalgenome.org) data (1kG data). To simplify the results, we just focused on the main branches of the tree. Then, we obtained the intersected Y-chromosome SNPs on the aDNA samples, using samtools mpileup -l option and filtering bases with BASEQ > 30. To avoid the interference of post-mortem damage, this analysis was performed using the rescaled bam files. Ancient Y-chromosome SNP data was then compared to the ancestral Y-chromosome, and mutations where classified in two groups depending on their ancestral or derived state. For plotting, we additionally classified the observed SNP based on its possible relation with DNA damage (C → T; G → A), following Sikora et al.[90]. Mutations observed for each major branch on the tree were plotted using R software[91] v.3.2.0 (R Core Team, 2013) and ggplot2 package[92].

The two samples from IAM site are both classified as E-M35 (Figure S5.2 and Figure S5.3), based on 123 and 2,409 intersected SNPs, respectively. KEB.6 (n=401 SNPs) is classified within haplogroup T-M184 (Figure S5.4), whereas TOR.5 (n=300 SNPs) belongs to haplogroup G-M201 (Figure S5.5). For samples KEB.7 (n=13 SNPs) and TOR.12 (n=36 SNPs) we did not have enough SNP information for providing a classification into haplogroups. Very few phylogenetic inconsistences are observed on the Y-chromosome plots: 0.83%, 0.29%, 0.24% and 1.33% for IAM.4, IAM.5, KEB.6 and TOR.5, respectively. We paid special attention to TOR.5, as this sample exhibited higher contamination rates and was submitted to post-mortem damage filtering (see Supplementary Note 3). This sample shows a higher number of inconsistencies when compared with the other samples. Although some inconsistent SNPs lead to the Y-chromosome haplogroup observed in the contaminated sample (R-M269), some other ones point to different Y-chromosome branches, and all except for one can be attributed to damage. Given into account that this sample accumulates higher damage rates due to MR score filtering, we do not think contamination is playing a major role and we are confident on the Y-chromosome classification.

For a more detailed analysis, we repeated the analyses as described before, but this time using a SNPs list of all the branches within the main haplogroups observed in our dataset: E-M35, G-M201 and T-M184. Y-chromosome plots were performed as before, but classifying the SNPs based on the branch ID on the complete Y-chromosome sequencing tree provided by Poznik et al.[89]. As an additional resource, we also obtained all the SNPs present in the ISOGG database (http://www.isogg.org/tree/).

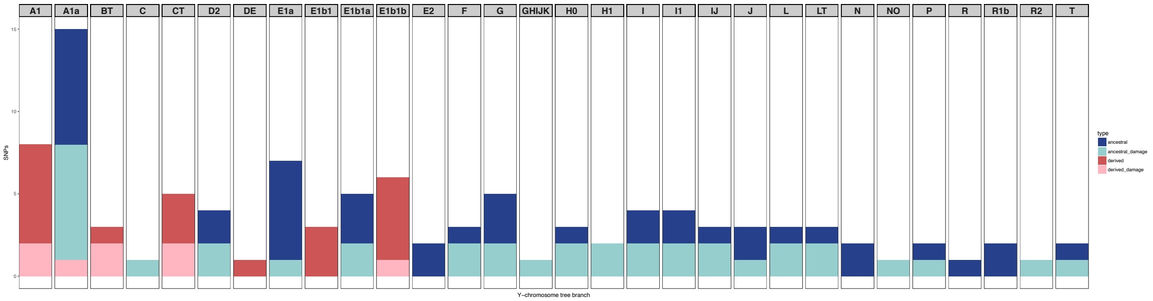


Figure S5.2 - Y-chromosome plot for IAM.4

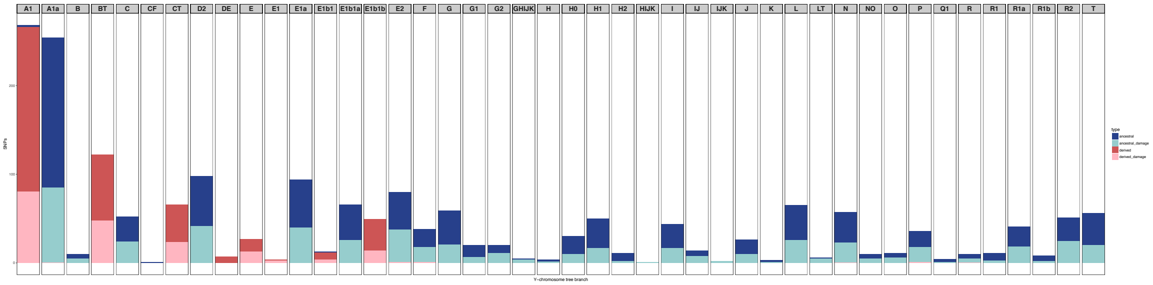


Figure S5.3 - Y-chromosome plot for IAM.5

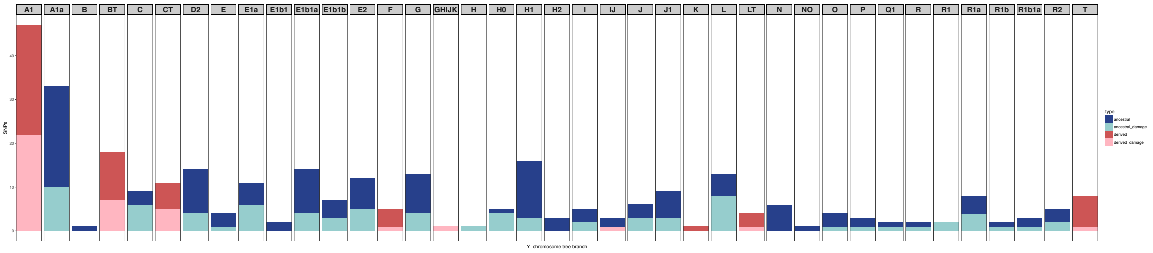


Figure S5.4 - Y-chromosome plot for KEB.6

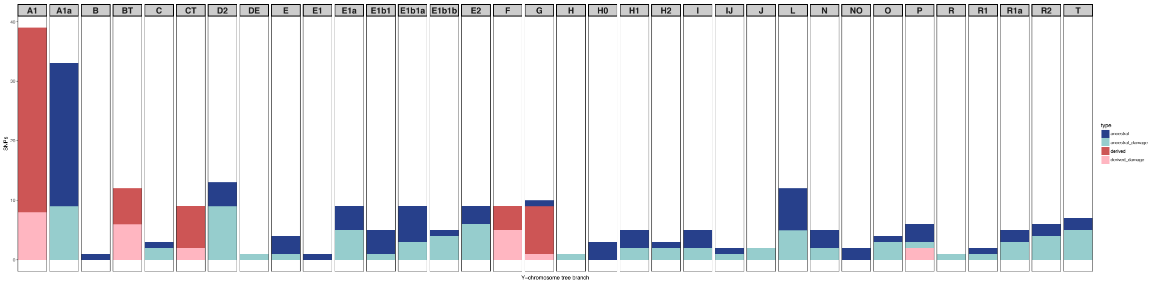


Figure S5.5 - Y-chromosome plot for TOR.

5.1. Haplogroup E-M35

Within the E-M35 cluster (branch 593), sample IAM.4 is only derived for branch 558 (Figure S5.6). This branch of the tree comprises E-L19 (542 branch) and E-M183 (557 branch) samples. As we also have some information on derived branches from 558, it seems IAM.4 does not cluster with either 542 or 557. A similar result is observed for IAM.5 (Figure S5.7), but in this case one out of 30 SNP intersected within 557 branch is derived. Although, we do not have enough information to be certain, this result could mean that IAM.5 (and maybe IAM.4) belong to a Y-chromosome lineage that is ancestral to the 557 branch, which comprises the North African E-M81 haplogroup. This scenario is plausible as E-M81 is younger than IAM samples (~3,700 BCE)[93] and its presence in North Africa has been associated with the spread of Neolithic pastoralism technologies from the Levant[94].

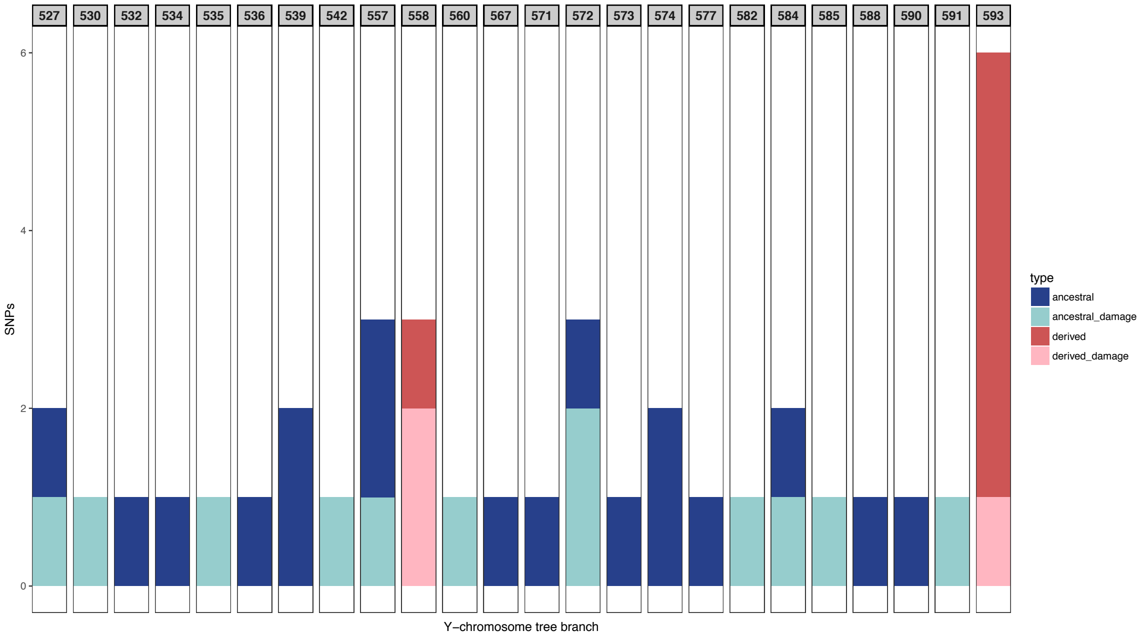


Figure S5.6 - Clade E-M35 plot for IAM.4

Haplogroup E-L19* (E-L19 (xE-M81)) is rather rare in modern populations and it has been spotted in Spain, Corsica, Sardinia, Morocco and Kenya[95]. On the other hand, E-M81 is the most frequent Y-chromosome lineage within North Africa[94], with an overall frequency of ~40%. In North Africa, E-M81 frequency varies in a latitudinal fashion, with the highest frequencies in Morocco and the lowest in Egypt[94,96]. E-M81 is rare outside the North African region, and it is considered to be a Berber autochthonous lineage. Apart from

North Africa, E-M81 is found only with low frequencies in neighboring regions in sub-Saharan Africa, the Near East and Mediterranean Europe. This lineage has been observed also in ancient samples from the indigenous people of the Canary Islands[63,97], so it is clear E-M81 was already present in North Africa at the time when Berber-like populations colonized the islands (~100 BCE).

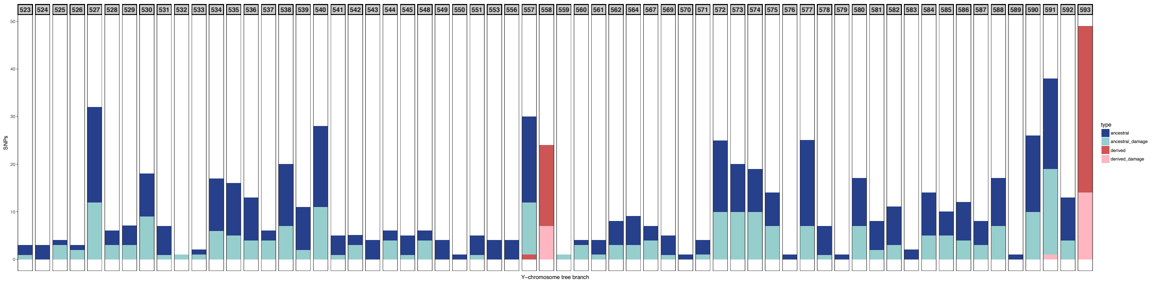


Figure S5.7 - Clade E-M35 plot for IAM.5

5.2. Haplogroup T-M70

KEB.6 is derived for the branches defining haplogroup LT (branch 339) and T (branch 285). Within the T clade, KEB.6 is only derived for branch 284 (Figure S5.8), but not for its sister branch 282. The only 1kG sample[98] on the tree matching that description is a Tuscan individual (NA20520), and it is classified as T-L208* (T1a1a). However, it is worth mentioning that the 1kG database lacks samples from North Africa. In fact, Haplogroup T-M70 (T1a) accounts for 1.16% - 6.22% of the North African Y-chromosome lineages[99].

The presence of haplogroup T in KEB is in agreement with the results observed for the mtDNA, indicating a tight relationship of this people with Near Eastern/European populations. Haplogroup T has been observed in Neolithic samples from Germany[56], as well as, Neolithic samples from Jordan[46]. Current T-M70 frequencies in North Africa are higher in Egypt than in Morocco[99], following an opposite distribution than E-M81[94].

5.3. Haplogroup G-M201

Sample TOR.5 is derived for G-M201 branch (165), and also for internal branches 159 and 153 (Figure S5.9). As the length of the internal branches is small within G-M201, it is complicated to confidently assign TOR.5 to a determined cluster. However, ISOGG data confirms that TOR.5 is derived for the Z39334 marker, being classified as G2a2b2a3a. As we know the source of the contamination for TOR.5, and it is not related to the G haplogroup, we can use the unfiltered data to get additional information. As the filtering option we used was rather stringent, we expect a portion of real ancient reads to be filtered out. When the unfiltered bam file was analyzed, reads with the derived nucleotide for branches 149 and 161 are present. This will corroborate the classification within G2a2b2a3a. This result is congruent with previous data on Neolithic data[46,51,56,61,67,68,72], placing haplogroup G2a as one of the most frequent lineages in ancient European samples[100].

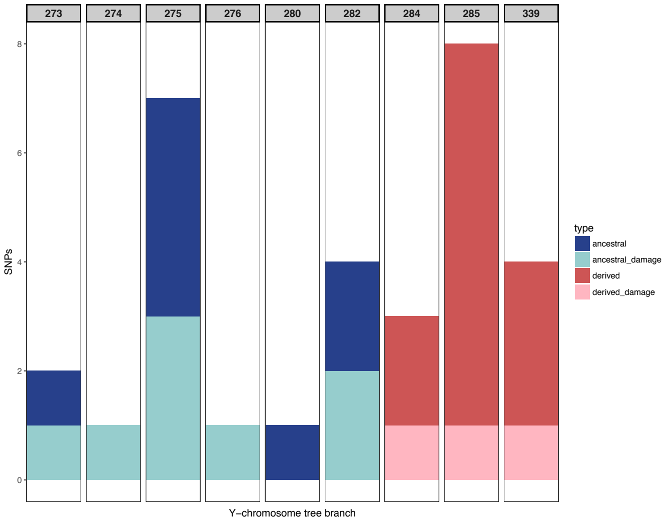


Figure S5.8 - Clade T-M184 plot for KEB.6

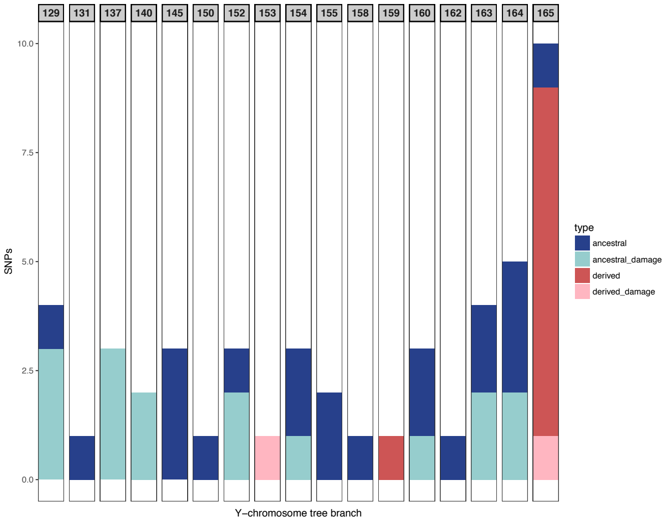


Figure S5.9 - Clade G-M201 plot for TOR.5

Supplementary Note 6: Principal component analysis

Rosa Fregel, Fernando L. Méndez, María C. Ávila-Arcos and Genevieve Wojcik

6.1. Modern DNA datasets

6.1.1. MEGA-HGDP panel

Our ancient samples were first compared to the Human Genome Diversity Project (HGDP)[101], genotyped on Illumina, Inc.'s Multiethnic Genotyping Array (MEGA), Consortium version. This dataset was genotyped at the Center for Inherited Disease Research (CIDR) as part of the Population Architecture using Genomics and Epidemiology (PAGE) Study with NHGRI. Standard initial quality control was conducted at CIDR, including gender discrepancies, Mendelian inconsistencies, unexpected duplication, unexpected non-duplication, poor performance, or observed DNA mixture. The dataset was further cleaned at the University of Washington Genetics Coordinating Center (UWGCC) using methods previously described by Laurie et al.[102]. SNP Quality Control was completed by filtering through exclusion of (1) CIDR technical filters, (2) SNPs with missing call rate >= 2%, (3) SNPs with more than 6 discordant calls in study duplicates, and (4) SNPs with greater than 1 Mendelian error. For our analysis, we subsampled individuals from Europe (Adygei, Basque, French, North Italian, North West Russian, Orcadian, Sardinian, and Tuscan samples), North Africa (Mozabites), Middle East (Bedoins, Druzes, and Palestinians), and sub-Saharan African populations (Bantu, Mandenka, and Yoruba). The modern DNA reference panel was further filtered for minimum allele frequency of 0.01 and genotyping rate of at least 95% using Plink[103] v.1.90. Indels, as well as SNPs on the sexual chromosomes and the mtDNA were filtered out. For PCA, a total of 840,301 SNPs were used. For admixture analyses, the dataset was pruned for linkage disequilibrium using PLINK (466,001 SNPs), with parameters --indep-pairwise 200 25 0.4.

6.1.2. Human Origins panel

The Human Origins panel contains 2,345 present-day humans from 203 populations genotyped for 594,924 SNPs[46]. Three subdatasets were used for analyses: a) all Human Origins populations; b) Sub-Saharan Africa, North Africa, Caucasus, Europe and the Middle East; c) North Africa, Europe and the Middle East. For PCA analysis, we filtered for minimum allele frequency of 0.01 and genotyping rate of at least 95%. For admixture analysis, we pruned for linkage disequilibrium as explained before.

6.2. Ancient DNA datasets

6.2.1. Human Origins panel

We used the following ancient samples from the Human Origins dataset from Lazaridis et al.[46]:

- Anatolian Chalcolithic (Anatolia_ChL): One sample from the Barcın site in Turkey (3,943–3,708 cal BCE).
- Anatolian Neolithic (Anatolia_N): Twenty-four samples from two Neolithic sites in Turkey: Barcın and Mentese. The dates for the samples (based on archaeological context) are 6,500–6,200 BCE for Barcın and 6,400–5,600 BCE for Mentese.
- Armenian Chalcolithic (Armenia_ChL): Five samples from Areni-1 cave complex (southern Armenia), with a radiocarbon date of 4,330–3,985 cal BCE.
- Armenian Early Bronze Age (Armenia_EBA): Three samples from two different sites in Armenia, with radiocarbon dates between 3,347–2,410 cal BCE.
- Armenia Middle/Late Bronze Age (Armenia_MLBA): Three samples from two different sites in Armenia, with radiocarbon dates between 2,619–855 cal BCE.
- Caucasus Paleolithic (CHG): Two samples from hunter-gatherer sites in Georgia, Kotias Klde (7,940–7,600 cal BCE) and Satsurblia (11,430–11,180 cal BCE).
- Eastern European Paleolithic (EHG): One sample from Samara site (5,657–5,541 cal BCE) and two from Karelia (6,850–6,000 BCE; 5,500–5,000 BCE).
- European Early Neolithic (Europe_N): Twenty-nine samples from several Early Neolithic sites in Germany, Hungary and Spain (~5,000 BCE). This group contains

samples from two Iberian sites associated to the Cardial technology (Iberia_EN), with dates between 5,300–5,000 cal BCE.

- European Middle Neolithic and Chalcolithic (Europe_MNChL): Twenty-seven samples from several Middle Neolithic and Chalcolithic sites in Germany, Hungary and Spain (~2,900 BCE). Fourteen samples belong to the Bell-Beaker culture in Spain (Iberia_MNChL).
- European Late Neolithic and Bronze Age (Europe_LNBA): Seventy samples from several sites in the Czech Republic, Denmark, Estonia, Germany, Hungary, Poland and Sweden (~1,700 BCE).
- Iranian Chalcolithic (Iran_ChL): Five samples from the Seh Gabi site (4,839–3,796 cal BCE).
- Iranian Neolithic (Iran_N): Five samples from Ganj Dareh (~8,000 cal BCE).
- Iranian Late Neolithic (Iran_LN): One sample from the Seh Gabi site (5,837–5,659 cal BCE).
- Iranian Mesolithic (Iran_HotuIIIb): A likely Mesolithic sample from the Hotu Cave (9,100–8,600 BCE).
- Levantine Paleolithic (Natufians): Hunter-gatherer samples from the Raqefet Cave in Israel (11,520–11,110 cal BCE).
- Levantine Neolithic (Levant_N): Twelve samples from 'Ain Ghazal site in Jordan, and one sample from Motza in Israel, belonging to the Pre Pottery Neolithic culture. The dates of the samples are 8,000–7,000 cal BCE.
- Levantine Bronze Age (Levant_BA): Three samples from 'Ain Ghazal site in Jordan (2,500–2,000 cal BCE).
- Scandinavian Hunter Gatherers (SHG): Six samples from Motala Cave in Sweden (6,000 - 5,500 cal BCE).
- Steppe Eneolithic (Steppe_Eneolithic): Three samples from Khvalynsk II in Russia (5,200–4,000 BCE).
- Steppe Early Middle Bronze Age (Steppe_EMBA): Twenty-eight samples from different locations in Russia (3,400–1,800 cal BCE).
- Steppe Middle Late Bronze Age (Steppe_MLBA): Twenty-two samples from different sites in Russia and Kazakhstan (2,900–1,600 cal BCE).
- Steppe Scythian (Steppe_IA): Scythian steppe warrior from Russia (375–203 cal BCE).

- Switzerland Paleolithic (Switzerland_HG): One sample from the Grotte du Bichon with a calibrated radiocarbon date of 11,820–11,610 cal BCE.
- Western European Paleolithic (WHG): Three hunter-gatherer samples from Hungary (5780–5640 cal BCE) and Luxemburg (6210–5990 cal BCE) and Spain (5983–5747 cal BCE).

6.2.2. Shotgun data

Shotgun data from previously published data was also integrated into our aDNA dataset, including Neolithic samples from Turkey (Bar8: 6,212–6,030 cal BCE; Bar31: 6,419–6,238 cal BCE) and Greece (Klei10: 4,230–3,995 cal BCE; Pal7: 4,452–4,350 cal BCE; Rev5: 6,438–6,264 cal BCE)[61], Ireland (3,343–3,020 cal BCE)[60] and Iran (7,700–8,000 cal BCE)[104]. These samples were analyzed using the pipeline described before. As our pipeline is designed for low-coverage genomes, we selected 30 million reads at random from each fastq file, matching the reads number of our best sample (IAM.5; Table S6.1). The samples from Turkey and Iran were included in the population label Anatolia_N and Iran_N, respectively. The samples from Greece were labeled as Aegean_N and the sample for Ireland as Ireland_N.

6.3. PCA using HGDP panel

Ancient samples with enough coverage were intersected with the HGDP panel. The average coverage for the MEGA and the Human Origins panels was 20.4% and 11.0%, respectively. The sample with lower coverage was IAM.3 with 4.0% and 1.2%, for HGDP and Human Origins datasets (Table S6.1).

We performed principal components analysis (PCA) based on the reference panel populations only, and then projecting ancient samples using two different methods. First, we tried LASER[105,106], a tool that uses PCA and Procrustes analysis to estimate aDNA samples ancestry. The advantage of this method is that it analyzes aDNA sequencing reads directly without calling genotypes. Briefly, LASER simulates sequence data for each reference individuals, matching the coverage of the aDNA sample, and builds the PCA space based on the simulated reference dataset along with the aDNA sample. Finally, the low-coverage clustering is projected onto the reference alone PCA space, using Procrustes analysis. For intersecting ancient samples we used the filtered bam files, but trimming 4 bp at end, to avoid damage interfering with clustering[46]. As this process is based on a simulation process, stochastic variation can lead to slightly different results for the same sample, LASER analysis was repeated 10 times and the mean coordinates were used for plotting.

ID	Shotgun/WISC reads	Shotgun/WISC coverage	MEGA coverage	Human Origins coverage
IAM.3	140,100	0.20%	3.93%	1.17%
IAM.4	1,173,879	1.74%	9.20%	3.75%
IAM.5	29,204,717	40.15%	54.79%	45.77%
IAM.6	539,234	0.78%	13.20%	4.01%
IAM.7	2,984,474	4.10%	25.61%	9.90%
KEB.1	4,204,676	6.89%	13.58%	8.62%
KEB.4	3,911,647	6.80%	41.38%	17.10%
KEB.6	5,651,915	8.49%	14.16%	11.56%
KEB.8	2,219,193	3.31%	5.18%	4.16%
TOR.6	6,827,793	9.88%	43.24%	22.31%
TOR.7	821,424	1.54%	15.82%	5.42%
TOR.8	1,676,075	2.89%	28.92%	10.17%
TOR.11	918,909	1.81%	11.17%	3.80%
BOT.1	2,949,332	5.75%	5.81%	5.59%

Table S6.1 - Genome-wide coverage

Second, we used smartpca[107] as in Lazaridis et al.[46]. For SNP calling, we obtained a list of the variants present on either HGDP or the Human Origins panels. Then, we retrieved the intersected genome-wide SNPs on the aDNA samples, using samtools mpileup -l option, filtering bases with BASEQ > 30. As in the LASER analysis, we used the bam files with 4 bp trimmed at both ends. Although all our samples produced low-coverage genomes, it is expected that some variants can have two different alleles. In these cases, we chose one allele at random. To avoid bias when comparing with the reference panels, one allele was also chosen at random for the HGDP and the Human Origins reference panels. Finally, all individual aDNA pileup files were merged with the reference datasets using PLINK. PCA was performed using smartpca with default parameters, except for "lsqproject: YES" and "numoutlieriter: 0" options[46]. In that way, PCA space is built on high coverage individuals, while ancient low-coverage samples are projected.

First, ancient samples were projected on a PCA space constructed with the MEGA-HGDP reference panel, using both LASER (Figure S6.1) and the lsqproject option from smartpca (Figure S6.2). Both analyses deliver the same result, with only slight differences. IAM samples cluster with Mozabites, whereas the southern Spain Neolithic samples (TOR and

BOT) are projected close to southern European populations, including Italians, Tuscans and Sardinians. As already suspected from the mtDNA and Y-chromosome data, KEB samples do not cluster with IAM and are placed in an intermediate position between IAM and TOR/BOT.

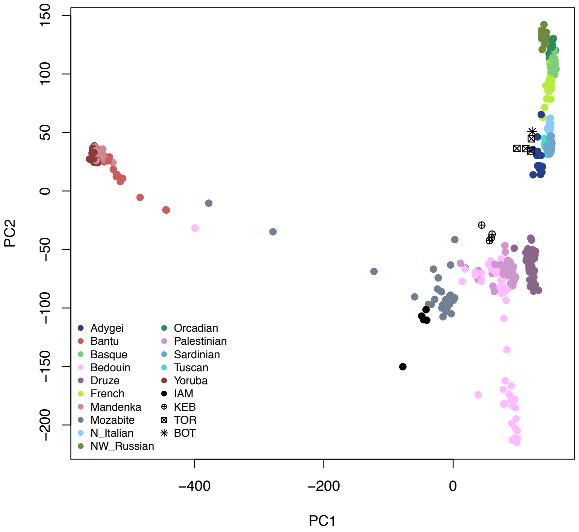


Figure S6.1 - LASER PCA plot for the MEGA-HGDP panel

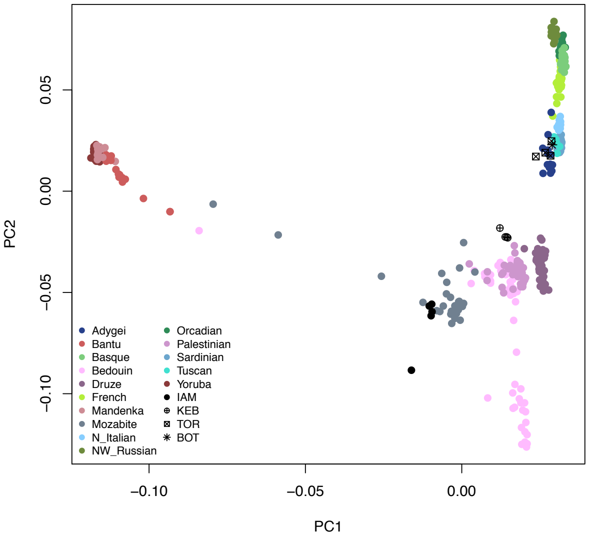


Figure S6.2 - Lsqproject PCA plot for the MEGA-HGDP panel

6.4. PCA using the Human Origins panel

For our PCA analysis, we chose the Human Origins panel containing modern samples, as well as, all the aDNA populations, to determine the relationship of our aDNA samples with other archaeological contexts. When projected on a PCA space built using modern samples from the Caucasus, Europe, the Middle East and North and Sub-Saharan Africa, all IAM, KEB and TOR samples cluster close to North African, Middle Eastern and European populations, respectively (Figure S6.3). We repeated the analysis removing the Caucasus and Sub-Saharan Africa (Figure 2). Lazaridis et al.[46] previously observed that Eurasian populations (North Africa not included) can be basically explained as a mixture of four sources of ancestry: Iranian Neolithic (represented by Iran_N), Levantine Neolithic (Levant_N), western European Paleolithic (WHG) and eastern European Paleolithic (EHG). These four populations are placed on the four corners of their PC1/PC2 plot, and all the other Eurasian populations are distributed based on their affinity with them. In our PCA (Figure 2), we observe the same pattern, with modern North African populations placed on the Levantine corner, close to the Middle Eastern and ancient samples from the Levant. IAM individuals cluster with modern populations from North Africa, and are different from any other ancient samples analyzed so far. TOR and BOT samples are projected close to the modern Sardinian population, together with other Neolithic samples from Europe and Anatolia. Finally, KEB samples are halfway between IAM and the Anatolian/European Neolithic cluster, close to Natufians and Neolithic samples from the Levant.

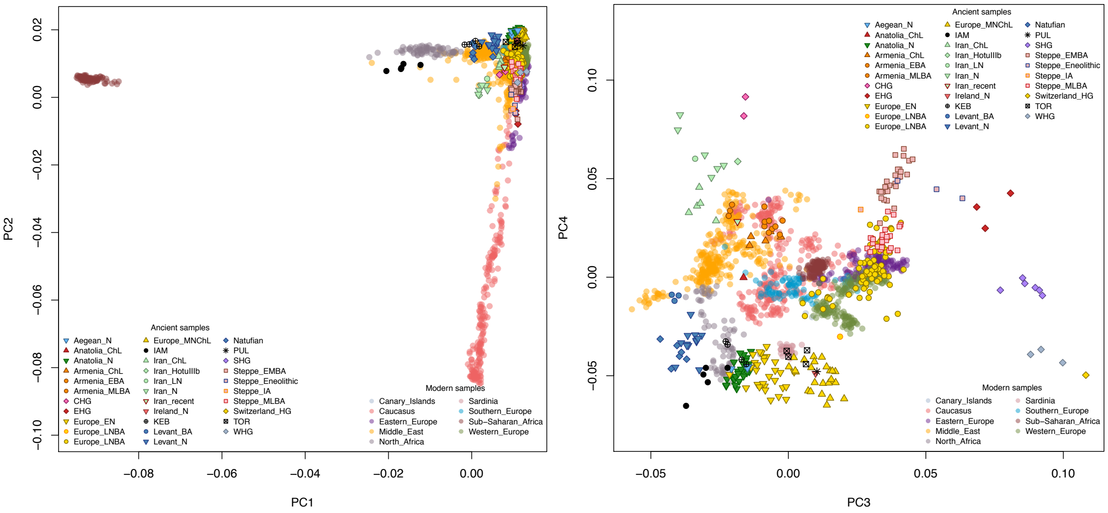


Figure S6.3- Lsqproject PCA plot for the Human Origins panel, including Sub-Saharan African and Caucasus samples

Supplementary Note 7: Global ancestry

Rosa Fregel, Fernando L. Méndez, María C. Ávila-Arcos and Genevieve Wojcik

7.1. HGDP panel

For admixture analysis, we used the same dataset as for lsqproject PCA, but pruning for linkage disequilibrium as explained before. The global ancestry of ancient samples was determined by unsupervised clustering using ADMIXTURE[108]. The analysis was performed in 10 replicates with different random seeds, and only the highest likelihood replicate for each value of K was taken into consideration. In Figure S7.1, we show ADMIXTURE results for K=2 to K=8 ancestral populations. For simplicity, BOT sample has been pooled with TOR samples, and considered together.

At K=2, the ADMIXTURE analysis separates sub-Saharan African (red component) from Eurasian (green) populations, including North Africa. All aDNA samples possess mostly the green component, with IAM showing a higher red component, similar to Mozabites. At K=3, European (green) and Middle East/North African (yellow) populations form different clusters. IAM and KEB consist mostly of the yellow component, while TOR has a major blue ancestral component. At K=4, North Africa (yellow) and the Middle East (violet) separate, with the violet component being higher in the Bedouins. Here, IAM and KEB are similar to Mozabites, although some red African-like ancestry is observed in IAM and some green European-like in KEB. At K=5, the eastern European component (orange) related to the steppe-like ancestry[51] separate from the early European Neolithic component, with Sardinians lacking any other contribution apart from the Neolithic ancestry. Congruently with PCA results, TOR is clustered with Sardinia, IAM with Mozabites, and KEB is placed in an intermediate position, with ~50% of both ancestries. This result is observed also in K=6 to K=8, while other components are further separated: Druze (K=6, grey), Palestinians (K=7, dark blue) and Mandenka (K=8, lilac).

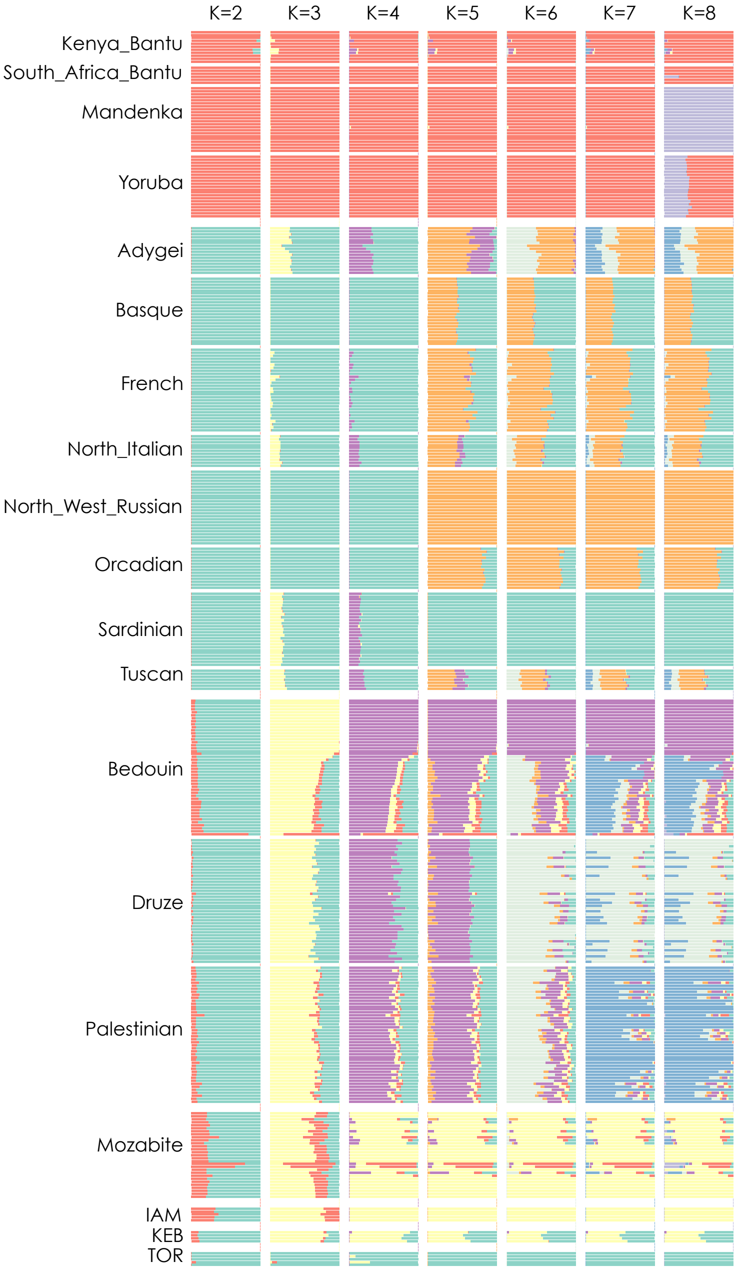


Figure S7.1 - ADMIXTURE plot for MEGA-HGDP panel (K=2 to K=8)

7.2. Human Origins panel

Admixture results for the ancient populations of Human Origins panel and our samples are showed in Figure S7.2, while modern populations from the same analysis are depicted in Figure S7.3. IAM population belongs exclusively to the yellow ancestral population from K=2 to K=8. This yellow component is shared with Levantine populations until K=7, when Natufians and Levant Pre-Pottery Neolithic populations ancestry splits from IAM. In the modern populations, the yellow component is shared at low K values with Middle Eastern populations, but from K=4 on, it is mostly observed in North Africa. Previous analyses on the modern populations of North Africa determined that their ancestry could be explained by the admixture of an autochthonous Maghrebi component with migrants from the Near East, and to a lesser degree from sub-Saharan Africa and Europe[109]. In our unsupervised clustering analysis, IAM fits with this autochthonous component from the Maghreb. Congruently, a clinal distribution of the IAM-like component is observed, with Moroccans and Saharawi having a higher proportion of it, and Egypt having a higher proportion of the Middle-Eastern-like ancestral population. The IAM-like ancestry has been related to the back-to-Africa migration from the Near East, more the 12,000 years ago. This would be on agreement with IAM having mtDNA linages associated to this migration, such us U6 and M1[45,110-114].

TOR samples are similar to other Neolithic samples, especially the Europe_MNChL group. In K values between 5 and 7, some of the European Neolithic samples possess the blue component, which is particularly observed in Western and Southern Europe, but it is dominant in Basques. That blue component is present in some Europe_EN (including Cardial samples from Spain) and Europe_MNChL samples (including Bell-Beaker samples from Spain and Italy). At K=8, a new violet component is majoritarian in Iberian Neolithic_EN and most Europe_MNChL, splitting from the early farmers green component. Europe_MNChL samples that posses 100% of the violet component include Early/Middle Neolithic and Chalcolithic sites from Iberia, Middle Neolithic sites from Germany (Baalberge and Salzmuende cultures) and a Chalcolithic site from Italy (Remedello culture) (Figure S7.4). This result could indicate an, at least partial, Iberian component in Middle Neolithic and Chalcolithic populations in Germany and Italy. It is worth mentioning that several Europe_MNChL samples show a partial early farmer green component, indicating admixture between both groups. As expected for Early Neolithic site in Iberia, TOR samples belong exclusively to the violet component, already present on Early Neolithic samples from Iberia (Cardial culture). Although TOR was associated to the Almagra culture, considered distinct from the Cardial technology and with a possible North African origin, they seem to have a similar genetic background.

Finally, from K=2 to K=8, KEB samples composition is halfway between IAM and TOR, as well as other Early Neolithic sample from Iberia and Chalcolithic samples from Europe. Given the material culture excavated at KEB, it is highly likely that the non-IAM ancestry is related to the expansion of Neolithic people from Iberia through the Mediterranean. This would also explain the presence of the violet component in Bronze Age populations from Armenia, Europe and the Steppe, Iran_ChL and Levant_N. As TOR samples, associated to the Almagra culture, belong to this component, it seems the presence in North Africa of similar pottery is because of European influx rather than the contrary.

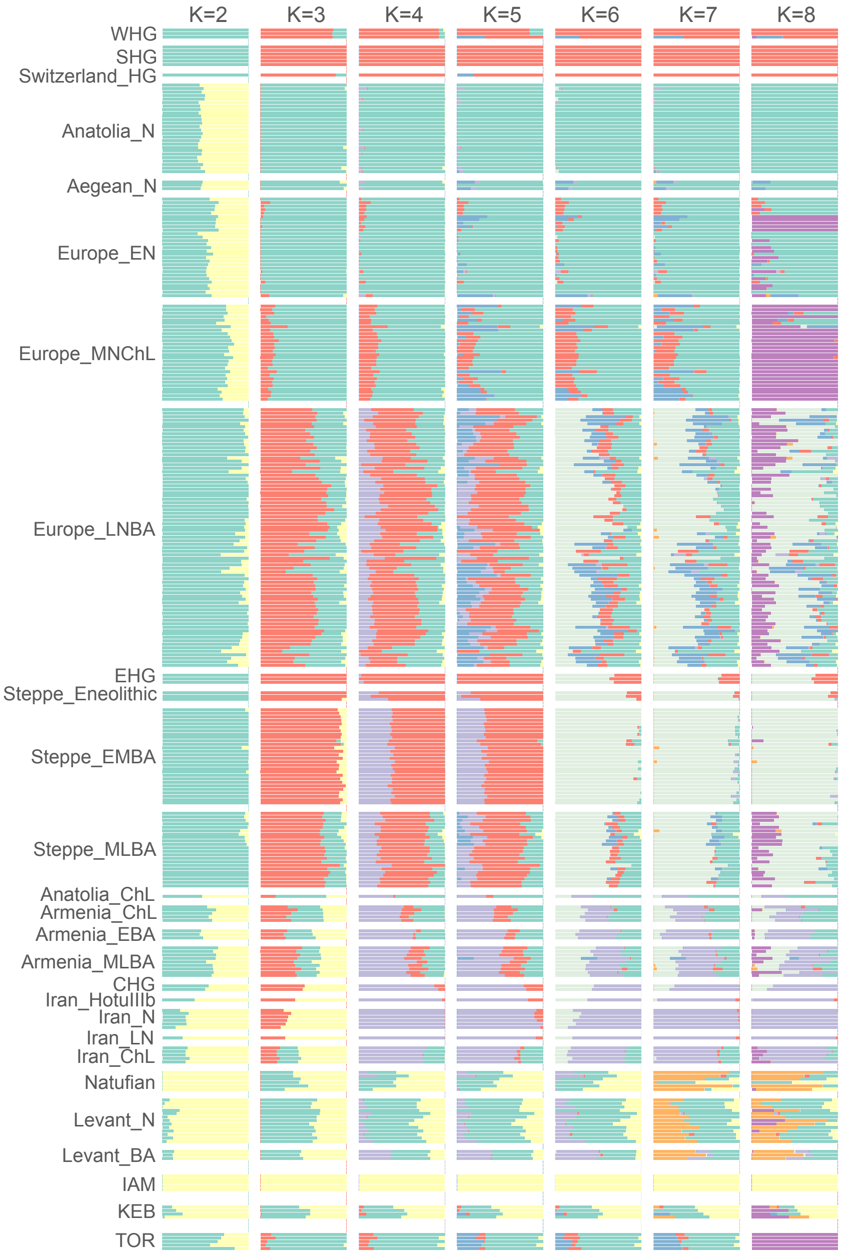


Figure S7.2- Admixture plot for the aDNA samples on the Human Origins panel (K=2 - K=8)

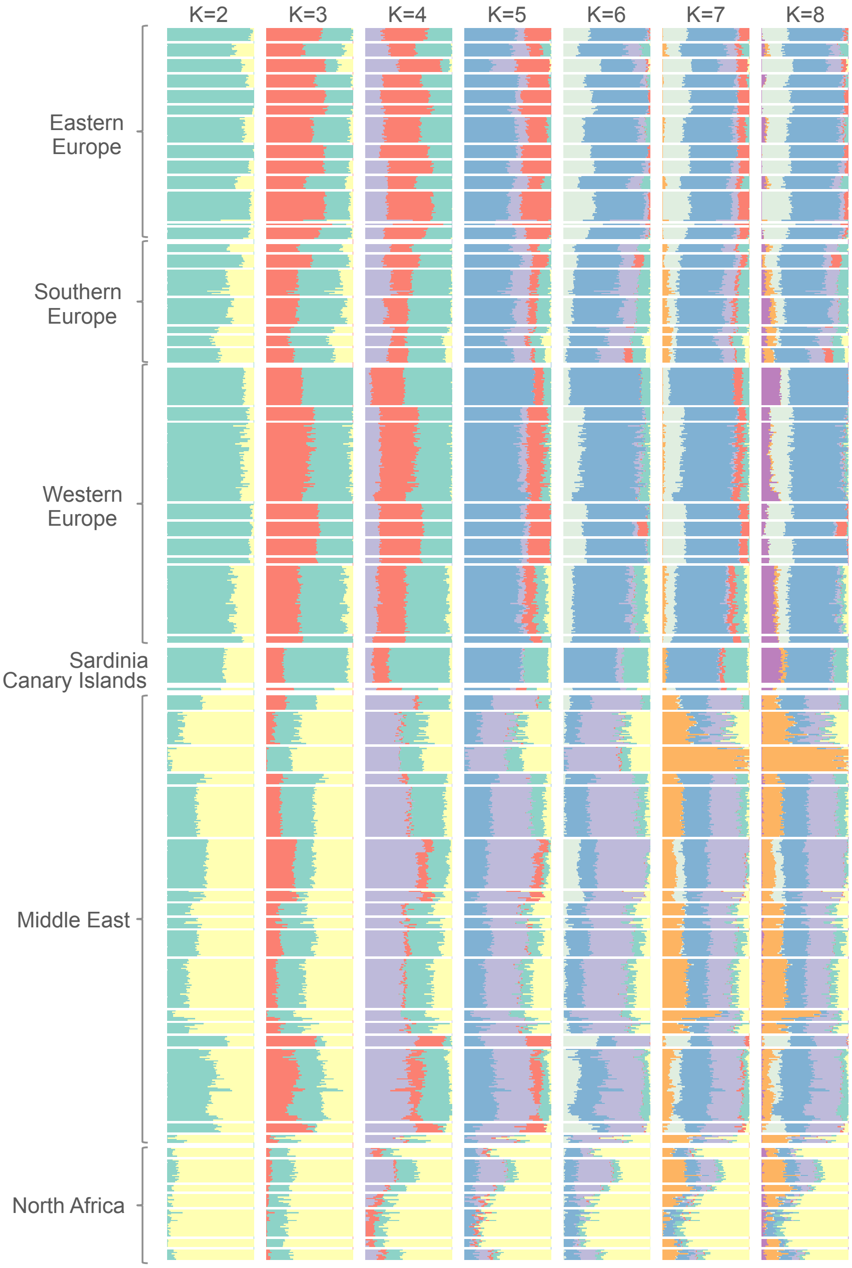


Figure S7.3 - Admixture plot for the modern samples on the Human Origins panel (K=2 - K=8)

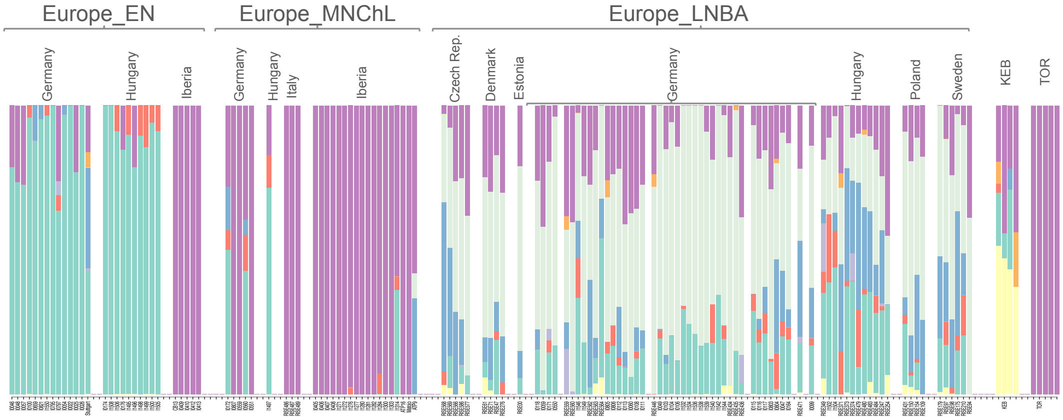


Figure S7.4 - Detail of ADMIXTURE analysis for European populations and KEB (K=8)

Supplementary Note 8: IBD distance and heterozygosity

8.1. Identity-by-descent proportions

Rosa Fregel, Fernando L. Méndez and Genevieve Wojcik

To test for family relationships, we checked if identity by descent (IBD) was higher between ancient samples within the same site. For that, we used the pruned dataset and PLINK to calculate IBD distances using the "--genome" option.

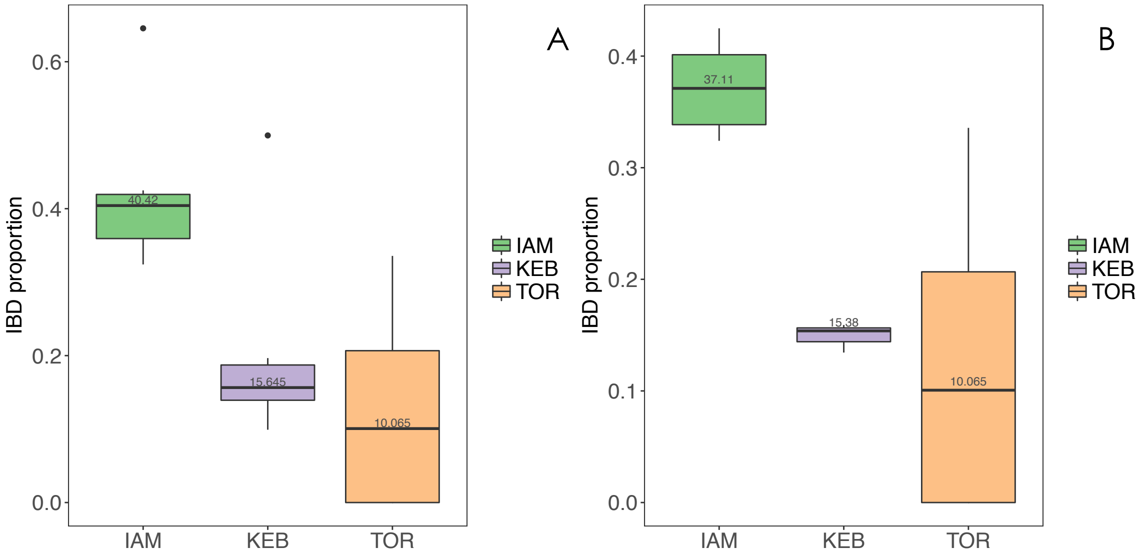


Figure S8.1 - PI_HAT value between samples before (A) and after (B) removing highly related samples

We observe that two outliers on the distribution of the combined IBD measure (PI HAT) within IAM and KEB (Figure S8.1). For IAM, samples IAM.4 and IAM.5 have a PI HAT value of 0.6455 and both share the same mtDNA lineage, and probably the same Y-chromosome lineage. KEB.1 and KEB.8, with a combined IBD distance of 0.4997, belong both to the sample X2b lineage. Samples IAM.4 and KEB.8 were removed from the analysis, and PI HAT values were recalculated (Figure S8.1). Even after removing IAM.4, it is clear that the proportion of relatedness in IAM is higher than that in KEB and TOR. This could be due to inbreeding, but also to isolation and drift.

8.2. Estimation of heterozygosity for genome-wide low coverage ancient DNA sequences

Fernando L. Méndez

To determine if low heterozygosity values are observed in IAM, we focused on the sample with the higher coverage (IAM.5, ~0.7X). For that purpose, we developed a method to estimate heterozygosity in ancient DNA samples, especially when they have intermediate to low coverage (~ 1X). This method takes into account sequencing errors, ancient DNA damage and the probabilities to observe each allele.

8.2.1. Transition matrix

8.2.1.1. General Description

This section presents background, assumptions, and theory behind the method.

8.2.1.1.1. Assumptions

- We assume that all redundant sequences have been removed, that is, all overlapping read pairs have been merged (or trimmed at their overlap) and that there is at most a single read or read pair originating from each DNA fragment extracted from the biological sample.

- We assume that each merged read or read pair has been correctly mapped to its corresponding location in the genome.

- We assume that the probability of (machine) sequencing errors is adequately described by a rate that depends only on the correct base and the reported base.

- We assume that chemically-induced damage in ancient DNA samples occurs only from bases C to T, which could be read as G to A if they occurred in the complementary strand, and that we can adequately describe this rate with a single number.

8.2.1.1.2 Assessment of the validity of these assumptions

The first assumption will hold after a careful process of duplicate removal. Its violation might lead to serious biases in the probability of observing a specific combination of counts for reference and alternative alleles. Duplication may be a special concern when there are sequence capture and PCR amplification steps involved in the preparation of the sequencing libraries.

The second assumption is also very important. A violation of this assumption may be problematic, especially when the genome has low coverage, as the method uses only sites that have coverage of at least 3, and repetitive sequences may exhibit inflated coverage. In addition, they may produce artifactual heterozygous calls. Additional false heterozygous sites may be the result of misalignments. The implementation of a filter that removes typically problematic regions greatly reduces this problem. We removed regions that were tagged as Simple, Satellite, microsatellite, low complexity, or chain self in the hg19 version of the human genome in the UCSC genome browser database. We count the number of bases for each depth of coverage for a range between 1 and 10 and compare with the expected number under a Poisson distribution. Despite known biases in the distribution of sequence coverage for the Illumina platform, our observation suggests that we can approximate the coverage as Poisson distributed for coverage equal to or lower than 6 (not shown). This suggests that our filter effectively removed problematic sequences.

The third assumption is an approximation, and while the sequencing error rates may depend on the sequence context and base qualities, they also depend on the individual run. If the variation in the sequencing error rates is only moderate, the cost of increasing the number of fitted parameters is likely to overwhelm any potential gain in power due to a more precise error model.

The fourth assumption is also an approximation, which seems supported by the overwhelmingly higher error rate in C to T and G to A errors.

8.2.1.2 Statistical model

In human populations, heterozygosity θ (per base probability of being heterozygous) is around 0.001. With sequencing error rates of the same order, it is common that most base calls that do not match the reference correspond to sequencing errors. The frequency with which a C can be read as a T (or a G as an A) due to damage depends both on the past of the DNA sample and the position of the base along the sequence fragment. As a first approximation, we assume that the rate is constant within an individual, but that can vary across samples.

We can write the following model for the probability $R_{i,j}$ of observing base *j* given that the original DNA fragment had base *i* when the individual was alive.

$$R_{i,i} = 1 - \sum_{\substack{j \in \{A,C,G,T\} \\ j \neq i}} R_{i,j}$$

$$R_{i,j} = \epsilon_{i,j} + \delta_{i,j}, \qquad j \neq i$$

where $\varepsilon_{i,j}$ corresponds to the sequencing error from *i* to *j*, $\delta_{i,j}$ corresponds to the damage from *i* to *j*, and $\delta_{C,T} = \delta_{G,A} > 0$, but $\delta_{i,j} = 0$, otherwise.

To simplify the notation, we define *N = {A,C,G,T}*, N_i = *{A,C,G,T}* \ *{i}*, and $N_{i,j}$ = *{A, C, G, T}* / *{i, j}*, so that, for instance,

$$R_{i,i} = 1 - \sum_{j \in N_i} R_{i,j}$$

Let us now consider the probability $B_{i,j}$ that a read without sequencing errors or damage has the base *j* while the reference has a base *i*. Let $f_{i,j}$ be the probability that an individual is homozygous for base *j* when the reference has *i*, and $\theta_{i,j}$ is the probability of having a heterozygous site between bases *i* and *j* when one of the bases is *i*. Then,

$$B_{i,j} = f_{i,j} + \frac{\theta_{i,j}}{2}$$

when $i \neq j$, and

$$B_{i,i} = 1 - \sum_{j \in N_i} \left(f_{i,j} + \frac{\theta_{i,j}}{2} \right)$$

When we combine this result with our error model, we get for the probability $P_{i,j}$ of observing base *j* when the reference has base *i*

$$P_{i,j} = \sum_{k \in N} B_{i,k} \cdot R_{k,j}$$

which we can write

$$P_{i,j} = B_{i,j} \cdot R_{j,j} + B_{i,i} \cdot R_{i,j} + \sum_{k \in N_{i,j}} B_{i,k} \cdot R_{k,j}$$

when $i \neq j$, and

$$P_{i,i} = \sum_{k \in N} B_{i,k} \cdot R_{k,i}$$

As a first approximation, we develop these formulas to order one, to obtain

$$P_{i,j} \sim B_{i,j} \cdot R_{j,j} + B_{i,i} \cdot R_{i,j} \sim \left(f_{i,j} + \frac{\theta_{i,j}}{2} + \epsilon_{i,j} + \delta_{i,j} \right)$$

when $i \neq j$, and

$$P_{i,i} \sim B_{i,i} \cdot R_{i,i} \sim 1 - \sum_{j \in N_i} \left(f_{i,j} + \frac{\theta_{i,j}}{2} + \epsilon_{i,j} + \delta_{i,j} \right)$$

We now consider probabilities for sites that are covered by two reads. Let us consider $P_{i,j,k}$, the probability that the reference has base *i*, and the observed base calls are j and *k*. If we restrict our analysis to sites with at most two alleles, we have

$$
\begin{aligned}
P_{i,i,i} \sim & \left(1-\sum_{j\in N_i} f_{i,j}\right)\cdot\left(1-\sum_{j\in N_i}\left(\epsilon_{i,j}+\delta_{i,j}\right)\right)\cdot \\
\cdot & \left[\left(1-\left(1-\frac{1}{4}\right)\cdot\sum_{j\in N_i}\theta_{i,j}\right)\cdot\left(1-\sum_{j\in N_i}\left(\epsilon_{i,j}+\delta_{i,j}\right)\right)+\sum_{j\in N_i}\left(\epsilon_{j,i}+\delta_{j,i}\right)\theta_{i,j}\right]
\end{aligned}
$$

$$
\begin{aligned}
P_{i,j,j} \sim & \left(1-\sum_{k\in N_j}\left(\epsilon_{j,k}+\delta_{j,k}\right)\right)^2\left[f_{i,j}+\left(1-\sum_{j\in N_i} f_{i,j}\right)\cdot\frac{\theta_{i,j}}{4}\right] \\
+ & \left(1-\sum_{k\in N_j}\left(\epsilon_{j,k}+\delta_{j,k}\right)\right)\left(1-\sum_{j\in N_i} f_{i,j}\right)\left(\epsilon_{i,j}+\delta_{i,j}\right)\frac{\theta_{i,j}}{2}
\end{aligned}
$$

$$
\begin{aligned}
P_{i,i,j} \sim & \left(1-\sum_{k\in N_i} f_{i,k}\right)\cdot\left[\left(1-\sum_{k\in N_j}\left(\epsilon_{i,k}+\delta_{i,k}\right)\right)\cdot\left(1-\sum_{k\in N_{i,j}}\left(\epsilon_{j,k}+\delta_{j,k}\right)\right)\cdot\theta_{i,j}\right] \\
+ & f_{i,j}\cdot\left(\epsilon_{j,i}+\delta_{j,i}\right)
\end{aligned}
$$

The remaining terms involve higher orders, in some cases due to errors in both reads.

$$
\begin{aligned}
P_{i,i,i} & \sim \left(1-\sum_{k\in N_i}\left(f_{i,k}+\epsilon_{i,k}+\delta_{i,k}\right)\right)\cdot\left(1-\sum_{k\in N_i}\left(\frac{3}{4}\theta_{i,k}+\epsilon_{i,k}+\delta_{i,k}\right)\right) \\
& \sim 1-\sum_{k\in N_i}\left(\frac{3}{4}\theta_{i,k}+f_{i,k}+2\epsilon_{i,k}+2\delta_{i,k}\right)
\end{aligned}
$$

$$
\begin{aligned}
P_{i,j,j} & \sim \left(1-2\sum_{k\in N_j}\left(\epsilon_{j,k}+\delta_{j,k}\right)\right)\left(f_{i,j}+\frac{\theta_{i,j}}{4}+\left(\delta_{i,j}+\epsilon_{i,j}\right)^2\right)+\left(1-\sum_{k\in N_j}\left(\epsilon_{j,k}+\delta_{j,k}+f_{i,j}\right)\right)\left(\epsilon_{i,j}+\delta_{i,j}\right)\frac{\theta_{i,j}}{2} \\
& \sim f_{i,j}+\frac{\theta_{i,j}}{4}+\left(\delta_{i,j}+\epsilon_{i,j}\right)^2
\end{aligned}
$$

$$
\begin{aligned}
P_{i,i,j} & \sim \left(1-\sum_{k\in N_i} f_{i,k}\right)\left[\left(1-\sum_{k\in N_{i,j}}\left(\epsilon_{i,k}+\delta_{i,k}+\epsilon_{j,k}+\delta_{j,k}\right)\right)\frac{\theta_{i,j}}{2}+\left(1-\sum_{k\in N_{i,j}}\theta_{i,k}\right)2\left(\delta_{i,j}+\epsilon_{i,j}\right)\right] \\
& \sim 2\left(\delta_{i,j}+\epsilon_{i,j}\right)+\frac{\theta_{i,j}}{2}
\end{aligned}
$$

We make a slight change in the notation to incorporate higher number of base calls at any position. In this new notation $P_{i:n_i,j:n_j}$, the reference has base *i* and the possible alternative allele is *j*. The number of base call agreeing with the reference or the alternative allele are n_i and n_j, respectively. With this notation, $P_{i,i,i}$, $P_{i,i,j}$, and $P_{i,j,j}$ become $P_{i:2,j:0}$, $P_{i:1,j:1}$, and $P_{i:0,j:2}$, respectively. We now have approximations for the probabilities of different configurations of observed bases for different numbers of observed bases.

For 3 reads:

$$P_{i:3,j:0} \sim 1 - \sum_{k \in N_i} \left(\frac{7}{8}\theta_{i,k} + f_{i,k} + 3\epsilon_{i,k} + 3\delta_{i,k} \right)$$

$$P_{i:2,j:1} \sim 3\left(\delta_{i,j} + \epsilon_{i,j}\right) + \frac{3\theta_{i,j}}{8}$$

$$P_{i:1,j:2} \sim 3\left(\delta_{i,j} + \epsilon_{i,j}\right)^2 + \frac{3\theta_{i,j}}{8}$$

$$P_{i:0,j:3} \sim f_{i,j} + \left(\delta_{i,j} + \epsilon_{i,j}\right)^3 + \frac{\theta_{i,j}}{8}$$

For 4 reads:

$$P_{i:4,j:0} \sim 1 - \sum_{k \in N_i} \left(\frac{15}{16}\theta_{i,k} + f_{i,k} + 4\epsilon_{i,k} + 4\delta_{i,k} \right)$$

$$P_{i:3,j:1} \sim 4\left(\delta_{i,j} + \epsilon_{i,j}\right) + \frac{4\theta_{i,j}}{16}$$

$$P_{i:2,j:2} \sim 6\left(\delta_{i,j} + \epsilon_{i,j}\right)^2 + \frac{6\theta_{i,j}}{16}$$

$$P_{i:1,j:3} \sim 4\left(\delta_{i,j} + \epsilon_{i,j}\right)^3 + \frac{4\theta_{i,j}}{16}$$

$$P_{i:0,j:4} \sim f_{i,j} + \left(\delta_{i,j} + \epsilon_{i,j}\right)^4 + \frac{\theta_{i,j}}{16}$$

When *j* >> *i* we also need to consider fixed differences that are observed as heterozygous due to back mutations. In general, we get

$$\begin{aligned} P_{i:n,j:0} &\sim 1 - \sum_{k \in N_i} \left(\frac{2^n - 1}{2^n}\theta_{i,k} + f_{i,k} + n\epsilon_{i,k} + n\delta_{i,k} \right) \\ P_{i:0,j:n} &\sim f_{i,j} + \left(\delta_{i,j} + \epsilon_{i,j}\right)^n + \frac{1}{2^n}\theta_{i,j} \end{aligned}$$

and, if m_1, $m_2 > 0$,

$$P_{i:m_i,j:m_j} \quad \sim \quad \binom{m_i + m_j}{m_j} \left((\delta_{i,j} + \epsilon_{i,j})^{m_j} + f_{i,j} \left(\delta_{j,i} + \epsilon_{j,i} \right)^{m_i} \right) + \frac{\binom{m_i+m_j}{m_j}}{2^{m_i+m_j}} \theta_{i,j}$$

Unless coverage is very high (very unusual for ancient DNA), at most sites at which the individual matches the reference, all base calls also match the reference. Most often, when only one read differs from the reference, the individual matches the reference, and the discrepancy is due either to sequencing error or DNA sequence damage (C to T or G to A). On the other hand, as the read coverage increases, most cases in which all base calls differ from the reference are due to homozygous differences between the individual and the reference sequence.

In all of our analyses, the $\delta_{i,j}$ and $\varepsilon_{i,j}$ appear added together, so the probability distributions for the configuration of read counts depend only $\delta_{i,j}$ + $\varepsilon_{i,j}$. Furthermore, rates for mutations and for the corresponding complemented bases should be equal, because errors and damage can occur in either strand. Therefore, the model has only eighteen parameters: 6 parameters for error and damage, 6 parameters for heterozygosity, and 6 parameters for fixed differences. The probability of having fixed differences is very small and plays an important role only in the probability of observing all base calls as differing from the reference, which leaves only 12 parameters that are important in our model. These parameters can be estimated in pairs (error and heterozygosity for the same pair of bases). For each of these pairs we can further partition our sites according to their coverage.

We estimate the parameters of the model using sites with coverage 4. Given the relatively high rate of damage, the estimation of mutations C to T and G to A would be very noisy if performed using sites with coverage 3 due to the significant contribution of errors to the counts where both bases are observed. On the other hand, sites with coverage 5 or higher are relatively scarce for the coverage of our sample. Because heterygosity and errors are coupled in our equations, we estimate them recursively. For errors, we use the sites where the alternative allele is observed only once, and for heterozygosity sites where the alternative allele is observed three times. If we use in the estimation of heterozygosity also the counts for sites in which the alternative allele is observed twice (except for transitions C to T and G to A) we obtain very similar estimates. To evaluate the uncertainty of our estimation, we also perform a parametric bootstrap analysis with 10,000 replicates and obtain a 95% confidence intervals for heterozygosity and for the fraction of sites that have homozygous differences with the reference. For heterozygosity, we estimate 6.6 x 10^{-4} (95%C.I.: 6.2 x 10^{-4} – 7.0 x 10^{-4}) and for the probability at any site of being homozygous different from the reference 4.6 x 10^{-4} (95%C.I. : 4.4 x 10^{-4} – 4.8 x 10^{-4}). Detailed results are shown in Tables S8.1 - S8.3.

Strength	Same Base	Transitions[a]	Same Strength[a] (Transversions)	Different Strength[a] (Transversions)	No-call
Strong	2,685,517	156,566:4,612:513:1,310	4,486:129:94:296	5,474:146:93:338	477
Weak	4,796,808	4,983:573:419:1,270	4,266:140:84:296	1,019:145:90:337	47

Table S8.1 - Number of observed sites in each state according to the strength of the reference sequence (coverage 4).

Strength	Transitions	Same Strength (Transversions)	Different Strength (Transversions)
Strong	0.01364	0.00038	0.00047
Weak	0.00024	0.00022	0.00005

Table S8.2 - Estimate of sequencing error rate and damage

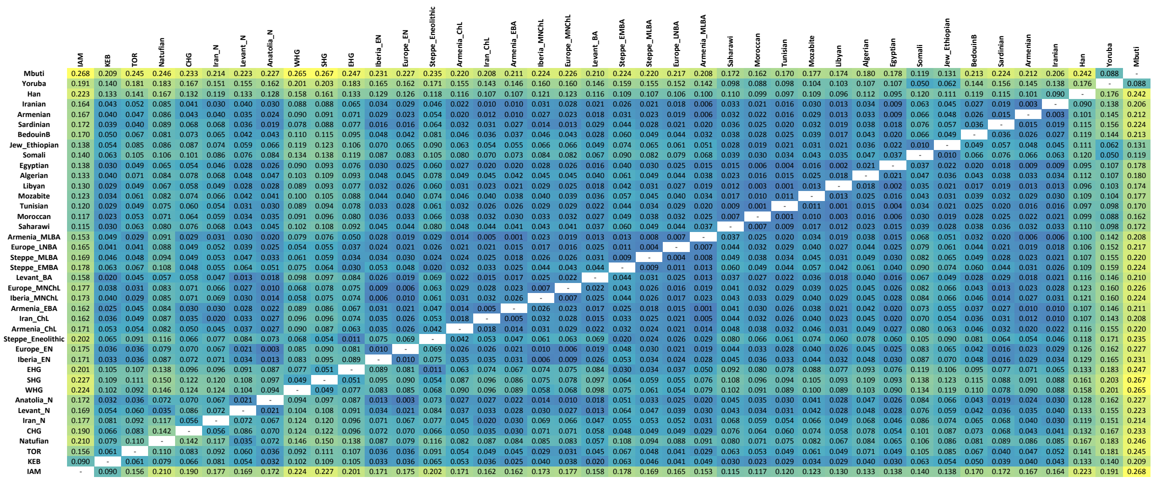
Strength	Transitions	Same Strength (Transversions)	Different Strength (Transversions)
Strong	0.00068	0.00013	0.00013
Weak	0.00035	0.00007	0.00007

Table S8.3 - Estimate of Heterozygosities

Supplementary Note 9: F_{ST} distance analysis

Rosa Fregel, Fernando L. Méndez, María C. Ávila-Arcos and Genevieve Wojcik

For F_{ST} calculations we used the Human Origins dataset filtered as in lsqproject PCA, but containing all populations and the shotgun data detailed in section 6.2.2. F_{ST} distances were obtained using smartpca with default parameters, except for inbreed: YES, and fstonly: YES. For this analysis, only population labels with more than 1 individual were considered, and samples IAM.4 and KEB.8 were removed from the analysis.



	IAM	KEB	TOR	Natufian	CHG	Iran_N	Levant_N	Anatolia_N	WHG	SHG	EHG	Iberia_EN	Europe_EN	Steppe_Eneolithic	Armenia_ChL	Iran_ChL	Armenia_EBA	Iberia_MNChL	Europe_MNChL	Levant_BA	Steppe_EMBA	Steppe_MLBA	Europe_LNBA	Armenia_MLBA	Saharawi	Moroccan	Tunisian	Mozabite	Libyan	Algerian	Egyptian	Somali	Jew_Ethiopian	BedouinB	Sardinian	Armenian	Iranian	Han	Yoruba	Mbuti
Mbuti	0.268	0.209	0.245	0.246	0.233	0.214	0.223	0.227	0.265	0.267	0.247	0.231	0.227	0.235	0.220	0.208	0.211	0.224	0.226	0.210	0.224	0.220	0.217	0.208	0.172	0.162	0.170	0.177	0.174	0.180	0.178	0.119	0.131	0.213	0.224	0.212	0.206	0.242	0.088	-
Yoruba	0.191	0.140	0.181	0.183	0.167	0.151	0.155	0.162	0.201	0.203	0.183	0.165	0.162	0.171	0.155	0.143	0.146	0.160	0.160	0.146	0.159	0.155	0.152	0.142	0.098	0.088	0.098	0.104	0.103	0.107	0.107	0.050	0.062	0.144	0.156	0.145	0.138	0.176	-	0.088
Han	0.223	0.133	0.141	0.167	0.132	0.119	0.133	0.128	0.158	0.161	0.133	0.129	0.126	0.118	0.116	0.107	0.107	0.121	0.123	0.116	0.109	0.107	0.106	0.100	0.110	0.099	0.097	0.109	0.096	0.112	0.095	0.120	0.111	0.119	0.115	0.101	0.090	-	0.176	0.242
Iranian	0.164	0.043	0.052	0.085	0.041	0.030	0.040	0.030	0.088	0.088	0.065	0.034	0.029	0.046	0.022	0.010	0.010	0.031	0.028	0.021	0.026	0.021	0.018	0.006	0.033	0.021	0.016	0.030	0.013	0.034	0.009	0.063	0.045	0.027	0.019	0.003	-	0.090	0.138	0.206
Armenian	0.167	0.040	0.047	0.086	0.043	0.040	0.035	0.024	0.090	0.091	0.071	0.029	0.023	0.054	0.020	0.012	0.010	0.027	0.023	0.018	0.031	0.023	0.019	0.006	0.032	0.022	0.016	0.029	0.013	0.033	0.009	0.066	0.048	0.026	0.015	-	0.003	0.101	0.145	0.212
Sardinian	0.172	0.039	0.040	0.089	0.068	0.068	0.036	0.019	0.078	0.088	0.077	0.016	0.016	0.064	0.032	0.031	0.027	0.014	0.013	0.029	0.044	0.028	0.021	0.020	0.036	0.025	0.020	0.032	0.019	0.038	0.018	0.076	0.057	0.036	-	0.015	0.019	0.115	0.156	0.224
BedouinB	0.170	0.050	0.067	0.081	0.073	0.065	0.042	0.043	0.110	0.115	0.095	0.048	0.042	0.081	0.046	0.036	0.037	0.046	0.043	0.028	0.060	0.049	0.044	0.032	0.038	0.028	0.025	0.039	0.017	0.043	0.020	0.066	0.049	-	0.036	0.026	0.027	0.119	0.144	0.213
Jew_Ethiopian	0.138	0.054	0.085	0.086	0.087	0.074	0.059	0.066	0.119	0.123	0.106	0.070	0.065	0.090	0.063	0.054	0.055	0.066	0.066	0.049	0.074	0.065	0.061	0.051	0.028	0.019	0.021	0.031	0.021	0.036	0.022	0.010	-	0.049	0.057	0.048	0.045	0.111	0.062	0.131
Somali	0.140	0.063	0.105	0.106	0.101	0.086	0.076	0.084	0.134	0.138	0.119	0.087	0.083	0.105	0.080	0.070	0.073	0.084	0.082	0.067	0.090	0.082	0.079	0.068	0.039	0.030	0.034	0.043	0.035	0.047	0.037	-	0.010	0.066	0.076	0.066	0.063	0.120	0.050	0.119
Egyptian	0.138	0.030	0.049	0.065	0.054	0.046	0.028	0.026	0.090	0.093	0.076	0.030	0.025	0.060	0.027	0.020	0.020	0.028	0.026	0.016	0.040	0.030	0.025	0.015	0.015	0.006	0.004	0.016	0.002	0.021	-	0.037	0.022	0.020	0.018	0.009	0.009	0.095	0.107	0.178
Algerian	0.133	0.040	0.071	0.084	0.078	0.068	0.048	0.047	0.103	0.109	0.093	0.048	0.045	0.078	0.049	0.045	0.042	0.045	0.045	0.040	0.061	0.049	0.044	0.038	0.023	0.016	0.015	0.025	0.018	-	0.021	0.047	0.036	0.043	0.038	0.033	0.034	0.112	0.107	0.180
Libyan	0.130	0.029	0.049	0.067	0.058	0.049	0.028	0.028	0.089	0.093	0.077	0.032	0.026	0.060	0.031	0.023	0.021	0.029	0.025	0.018	0.042	0.031	0.027	0.019	0.012	0.003	0.001	0.013	-	0.018	0.002	0.035	0.021	0.017	0.019	0.013	0.013	0.096	0.103	0.174
Mozabite	0.123	0.034	0.061	0.082	0.074	0.066	0.042	0.041	0.100	0.105	0.088	0.044	0.040	0.074	0.046	0.040	0.038	0.040	0.039	0.036	0.057	0.045	0.040	0.034	0.017	0.010	0.011	-	0.013	0.025	0.016	0.043	0.031	0.039	0.032	0.029	0.030	0.109	0.104	0.177
Tunisian	0.120	0.029	0.049	0.075	0.060	0.054	0.031	0.030	0.089	0.094	0.078	0.033	0.028	0.061	0.032	0.026	0.026	0.029	0.029	0.022	0.044	0.034	0.029	0.020	0.009	0.001	-	0.011	0.001	0.015	0.004	0.034	0.021	0.025	0.020	0.016	0.016	0.097	0.098	0.170
Moroccan	0.117	0.023	0.053	0.071	0.064	0.059	0.034	0.035	0.091	0.096	0.080	0.036	0.033	0.066	0.038	0.032	0.030	0.033	0.032	0.027	0.049	0.038	0.032	0.025	0.007	-	0.001	0.010	0.003	0.016	0.006	0.030	0.019	0.028	0.025	0.022	0.021	0.099	0.088	0.162
Saharawi	0.115	0.030	0.063	0.080	0.076	0.068	0.043	0.045	0.102	0.108	0.092	0.045	0.044	0.080	0.048	0.044	0.041	0.043	0.041	0.037	0.060	0.049	0.044	0.037	-	0.007	0.009	0.017	0.012	0.023	0.015	0.039	0.028	0.038	0.036	0.032	0.033	0.110	0.098	0.172
Armenia_MLBA	0.153	0.049	0.029	0.091	0.029	0.031	0.030	0.020	0.079	0.076	0.050	0.028	0.019	0.029	0.014	0.005	0.001	0.023	0.019	0.013	0.013	0.008	0.007	-	0.037	0.025	0.020	0.034	0.019	0.038	0.015	0.068	0.051	0.032	0.020	0.006	0.006	0.100	0.142	0.208
Europe_LNBA	0.165	0.041	0.041	0.088	0.049	0.052	0.039	0.025	0.054	0.055	0.037	0.024	0.021	0.026	0.021	0.021	0.015	0.017	0.016	0.025	0.011	0.004	-	0.007	0.044	0.032	0.029	0.040	0.027	0.044	0.025	0.079	0.061	0.044	0.021	0.019	0.018	0.106	0.152	0.217
Steppe_MLBA	0.169	0.046	0.048	0.094	0.049	0.053	0.047	0.033	0.061	0.059	0.034	0.034	0.030	0.024	0.024	0.025	0.018	0.026	0.026	0.031	0.009	-	0.004	0.008	0.049	0.038	0.034	0.045	0.031	0.049	0.030	0.082	0.065	0.049	0.028	0.023	0.021	0.107	0.155	0.220
Steppe_EMBA	0.178	0.063	0.067	0.108	0.048	0.055	0.064	0.051	0.075	0.064	0.030	0.053	0.048	0.020	0.032	0.033	0.025	0.044	0.043	0.044	-	0.009	0.011	0.013	0.060	0.049	0.044	0.057	0.042	0.061	0.040	0.090	0.074	0.060	0.044	0.031	0.026	0.109	0.159	0.224
Levant_BA	0.158	0.020	0.045	0.057	0.058	0.047	0.013	0.018	0.098	0.097	0.084	0.026	0.019	0.069	0.022	0.015	0.017	0.025	0.022	-	0.044	0.031	0.025	0.013	0.037	0.027	0.022	0.036	0.018	0.040	0.016	0.067	0.049	0.028	0.029	0.018	0.021	0.116	0.146	0.210
Europe_MNChL	0.177	0.038	0.031	0.083	0.071	0.066	0.027	0.010	0.068	0.078	0.075	0.009	0.006	0.063	0.029	0.028	0.023	0.007	-	0.022	0.043	0.026	0.016	0.019	0.041	0.032	0.029	0.039	0.025	0.045	0.026	0.082	0.066	0.043	0.013	0.023	0.028	0.123	0.160	0.226
Iberia_MNChL	0.173	0.040	0.029	0.085	0.071	0.069	0.030	0.014	0.058	0.075	0.074	0.006	0.010	0.061	0.031	0.032	0.026	-	0.007	0.025	0.044	0.026	0.017	0.023	0.043	0.033	0.029	0.040	0.029	0.045	0.028	0.084	0.066	0.046	0.014	0.027	0.031	0.121	0.160	0.224
Armenia_EBA	0.162	0.025	0.045	0.084	0.030	0.030	0.028	0.022	0.089	0.086	0.067	0.031	0.021	0.047	0.014	0.005	-	0.026	0.023	0.017	0.025	0.018	0.015	0.001	0.041	0.030	0.026	0.038	0.021	0.042	0.020	0.073	0.055	0.037	0.027	0.010	0.010	0.107	0.146	0.211
Iran_ChL	0.162	0.036	0.049	0.087	0.035	0.020	0.033	0.027	0.096	0.096	0.074	0.035	0.026	0.053	0.018	-	0.005	0.032	0.028	0.015	0.033	0.025	0.021	0.005	0.044	0.032	0.026	0.040	0.023	0.045	0.020	0.070	0.054	0.036	0.031	0.012	0.010	0.107	0.143	0.208
Armenia_ChL	0.171	0.053	0.054	0.082	0.050	0.045	0.037	0.027	0.090	0.087	0.063	0.035	0.026	0.042	-	0.018	0.014	0.031	0.029	0.022	0.032	0.024	0.021	0.014	0.048	0.038	0.032	0.046	0.031	0.049	0.027	0.080	0.063	0.046	0.032	0.020	0.022	0.116	0.155	0.220
Steppe_Eneolithic	0.202	0.065	0.091	0.116	0.066	0.077	0.084	0.073	0.068	0.054	0.011	0.075	0.069	-	0.042	0.053	0.047	0.061	0.063	0.069	0.020	0.024	0.026	0.029	0.080	0.066	0.061	0.074	0.060	0.078	0.060	0.105	0.090	0.081	0.064	0.054	0.046	0.118	0.171	0.235
Europe_EN	0.175	0.036	0.036	0.079	0.070	0.067	0.021	0.003	0.085	0.090	0.081	0.010	-	0.069	0.026	0.026	0.021	0.010	0.006	0.019	0.048	0.030	0.021	0.019	0.044	0.033	0.028	0.040	0.026	0.045	0.025	0.083	0.065	0.042	0.016	0.023	0.029	0.126	0.162	0.227
Iberia_EN	0.171	0.033	0.036	0.087	0.072	0.071	0.034	0.013	0.083	0.095	0.089	-	0.010	0.075	0.035	0.035	0.031	0.006	0.009	0.026	0.053	0.034	0.024	0.028	0.045	0.036	0.033	0.044	0.032	0.048	0.030	0.087	0.070	0.048	0.016	0.029	0.034	0.129	0.165	0.231
EHG	0.201	0.105	0.107	0.138	0.096	0.096	0.091	0.087	0.077	0.051	-	0.089	0.081	0.011	0.063	0.074	0.067	0.074	0.075	0.084	0.030	0.034	0.037	0.050	0.092	0.080	0.078	0.088	0.077	0.093	0.076	0.119	0.106	0.095	0.077	0.071	0.065	0.133	0.183	0.247
SHG	0.227	0.109	0.111	0.150	0.122	0.120	0.108	0.097	0.049	-	0.051	0.095	0.090	0.054	0.087	0.096	0.086	0.075	0.078	0.097	0.064	0.059	0.055	0.076	0.108	0.096	0.094	0.105	0.093	0.109	0.093	0.138	0.123	0.115	0.088	0.091	0.088	0.161	0.203	0.267
WHG	0.224	0.102	0.092	0.146	0.124	0.124	0.104	0.094	-	0.049	0.077	0.083	0.085	0.068	0.090	0.096	0.089	0.058	0.068	0.098	0.075	0.061	0.054	0.079	0.102	0.091	0.089	0.100	0.089	0.103	0.090	0.134	0.119	0.110	0.078	0.090	0.088	0.158	0.201	0.265
Anatolia_N	0.172	0.032	0.036	0.072	0.070	0.067	0.021	-	0.094	0.097	0.087	0.013	0.003	0.073	0.027	0.027	0.022	0.014	0.010	0.018	0.051	0.033	0.025	0.020	0.045	0.035	0.030	0.041	0.028	0.047	0.026	0.084	0.066	0.043	0.019	0.024	0.030	0.128	0.162	0.227
Levant_N	0.169	0.054	0.060	0.035	0.086	0.072	-	0.021	0.104	0.108	0.091	0.034	0.021	0.084	0.037	0.033	0.028	0.030	0.027	0.013	0.064	0.047	0.039	0.030	0.043	0.034	0.031	0.042	0.028	0.048	0.028	0.076	0.059	0.042	0.036	0.035	0.040	0.133	0.155	0.223
Iran_N	0.177	0.081	0.092	0.117	0.056	-	0.072	0.067	0.124	0.120	0.096	0.071	0.067	0.077	0.045	0.020	0.030	0.069	0.066	0.047	0.055	0.053	0.052	0.031	0.068	0.059	0.054	0.066	0.049	0.068	0.046	0.086	0.074	0.065	0.068	0.040	0.030	0.119	0.151	0.214
CHG	0.190	0.066	0.083	0.142	-	0.056	0.086	0.070	0.124	0.122	0.096	0.072	0.070	0.066	0.050	0.035	0.030	0.071	0.071	0.058	0.048	0.049	0.049	0.029	0.076	0.064	0.060	0.074	0.058	0.078	0.054	0.101	0.087	0.073	0.068	0.043	0.041	0.132	0.167	0.233
Natufian	0.210	0.079	0.110	-	0.142	0.117	0.035	0.072	0.146	0.150	0.138	0.087	0.079	0.116	0.082	0.087	0.084	0.085	0.083	0.057	0.108	0.094	0.088	0.091	0.080	0.071	0.075	0.082	0.067	0.084	0.065	0.106	0.086	0.081	0.089	0.086	0.085	0.167	0.183	0.246
TOR	0.156	0.061	-	0.110	0.083	0.092	0.060	0.036	0.092	0.111	0.107	0.036	0.036	0.091	0.054	0.049	0.045	0.029	0.031	0.045	0.067	0.048	0.041	0.029	0.063	0.053	0.049	0.061	0.049	0.071	0.049	0.105	0.085	0.067	0.040	0.047	0.052	0.141	0.181	0.245
KEB	0.090	-	0.061	0.079	0.066	0.081	0.054	0.032	0.102	0.109	0.105	0.033	0.036	0.065	0.053	0.036	0.025	0.040	0.038	0.020	0.063	0.046	0.041	0.049	0.030	0.023	0.029	0.034	0.029	0.040	0.030	0.063	0.054	0.050	0.039	0.040	0.043	0.133	0.140	0.209
IAM	-	0.090	0.156	0.210	0.190	0.177	0.169	0.172	0.224	0.227	0.201	0.171	0.175	0.202	0.171	0.162	0.162	0.173	0.177	0.158	0.178	0.169	0.165	0.153	0.115	0.117	0.120	0.123	0.130	0.133	0.138	0.140	0.138	0.170	0.172	0.167	0.164	0.223	0.191	0.268

Figure S9.1 - Pair-wise FST distances between populations (after removing related samples)

When we compare pair-wise F_{ST} distances, the most striking result is that IAM presents rather high F_{ST} values with all population except for KEB (Figure S9.1), and even in this case the F_{ST} value is 0.090 (similar to the distance between Yoruba and Mbuti). In fact, IAM is in general as distant to other Eurasians as it is the Yoruba population. In a detailed population-by-population comparison (Figure S9.2), we can see that IAM is closer to modern North African populations, following the west to east trend described before, in such a way Saharawis and Moroccans are closer than Egyptians (Figure S9.3). Outside North Africa, IAM has its higher similarities with the Horn of Africa, the Arabian Peninsula and the Levant. As was deduced from the PCA and ADMIXTURE analysis, IAM is different from any other ancient sampled studied so far, except for KEB (Figure S9.2). Although, ADMIXTURE analysis pointed to some relationship between IAM and Levantine aDNA samples, especially the Natufians, this is not supported by F_{ST} distances.

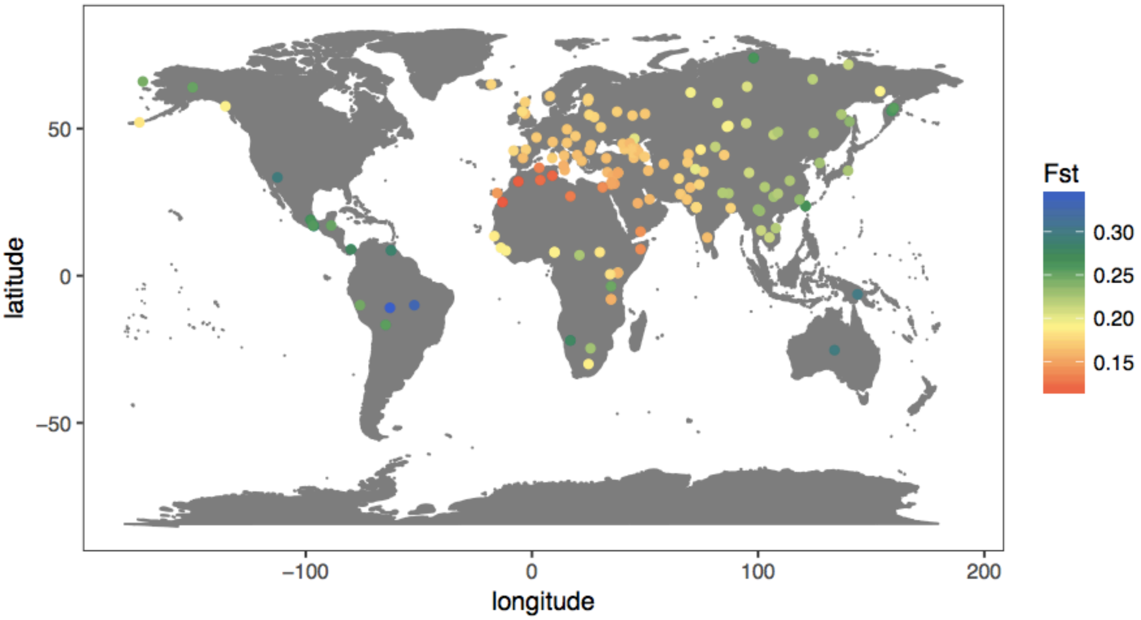


Figure S9.2 - Pair-wise F_{ST} values for IAM and modern populations of the Human Origins panel

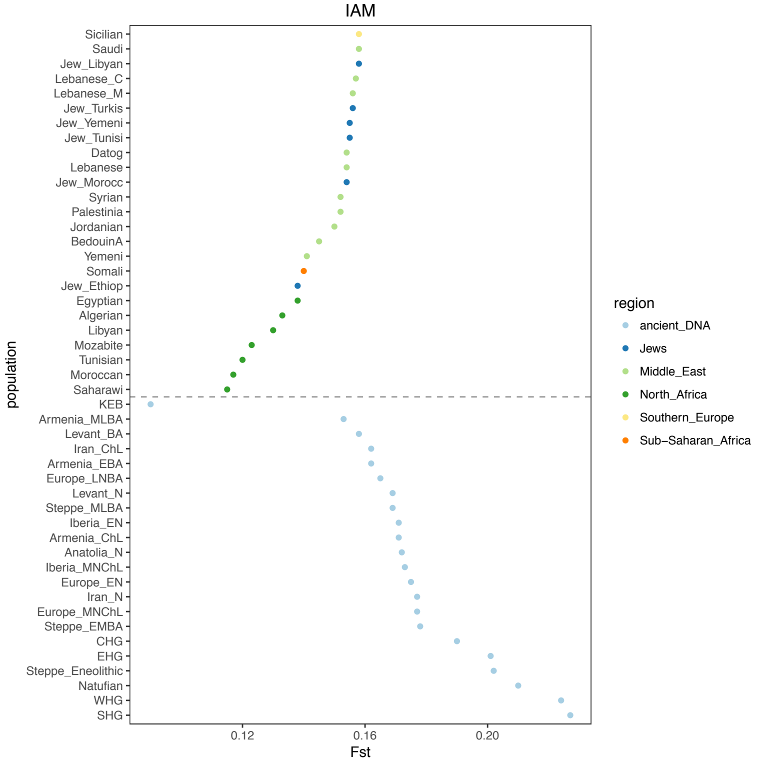


Figure S9.3 - Pair-wise F_{ST} values for IAM and other ancient populations, as well as values for the twenty-five closest modern populations

On the modern DNA reference panel, KEB is similar to North Africans, as well as, European and Middle Eastern populations (Figure S9.4). KEB has lower F_{ST} distances with Moroccans and Canary Islanders, which is an admixed population with both North African and European ancestry. Compared to the aDNA dataset, KEB samples are more similar to Bronze Age samples from Armenia and the Levant (Figure S9.5). Although KEB is the closest population to IAM, the contrary it not true. KEB is closer to any Anatolian, European, Levantine and Iranian sample, than to IAM, with the only exception of European hunter-gatherers (EHG, SHG and WHG).

Regarding the modern populations, TOR is closer to modern populations from Spain, North Italy and Sardinia, and other Southern and Western European populations (Figure S9.6 and S9.7). Compared to ancient populations, TOR is similar to Middle Neolithic/Chalcolithic populations from Europe and Middle/Late Bronze Age populations from Armenia, followed by Early Neolithic populations from Anatolia, Europe and Iberia.

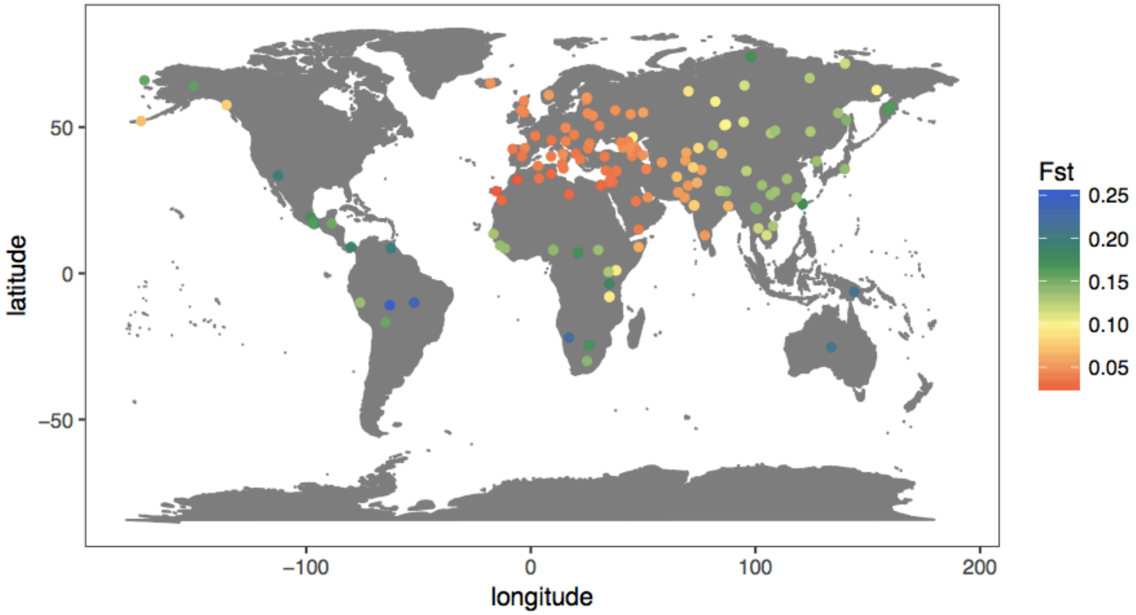


Figure S9.4 - Pair-wise F_{ST} values for KEB and modern populations of the Human Origins panel

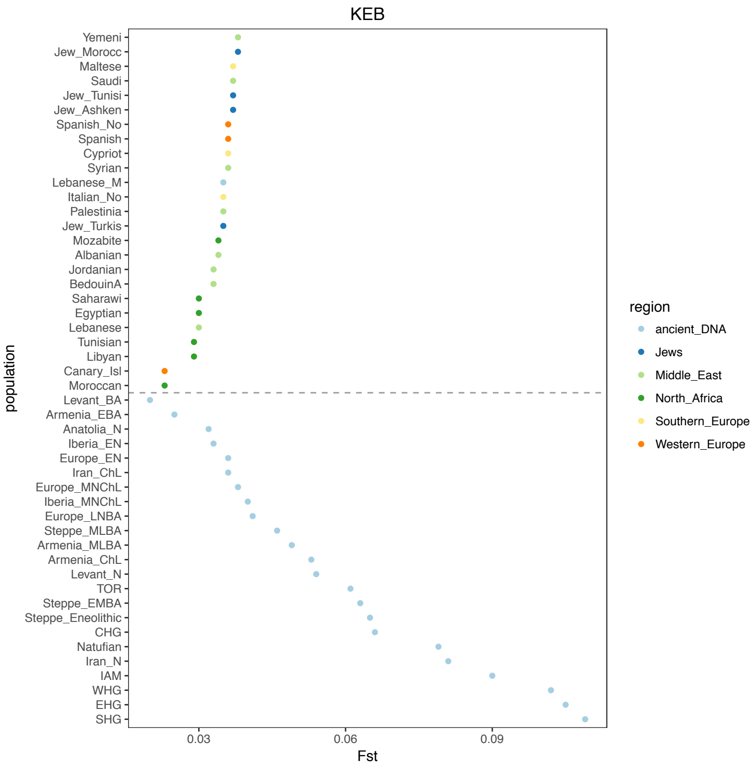


Figure S9.5 - Pair-wise F_{ST} values for KEB and other ancient populations, as well as values for the twenty-five closest modern populations

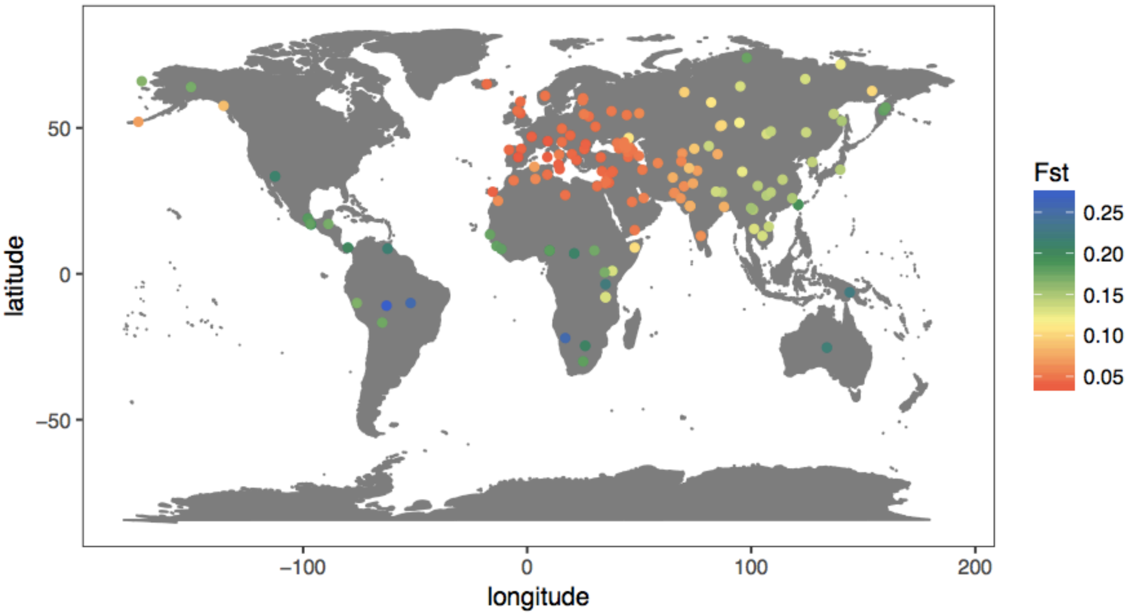


Figure S9.6 - Pair-wise F_{ST} values for TOR and modern populations of the Human Origins panel

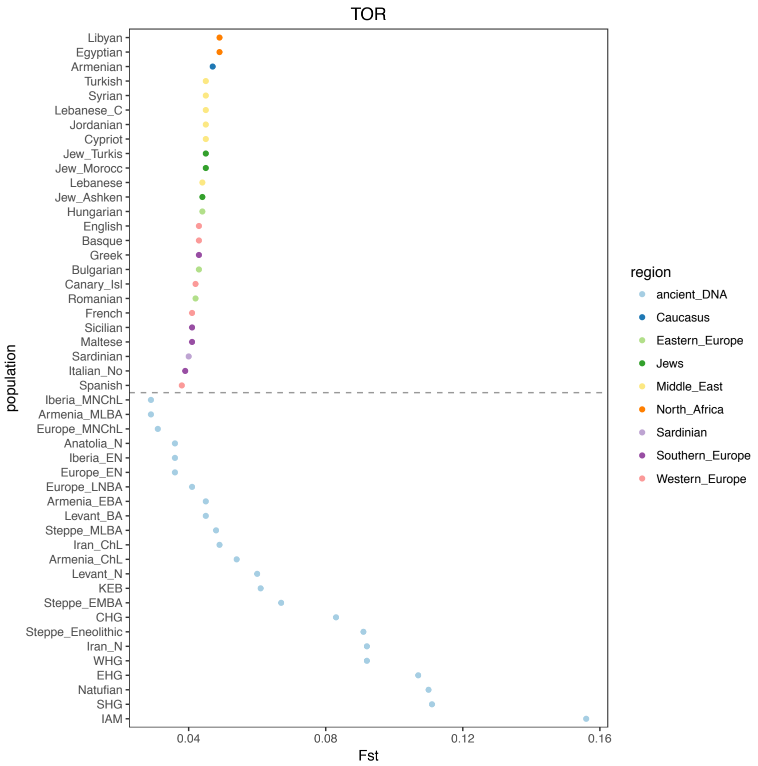


Figure S9.7 - Pair-wise F_{ST} values for TOR and other ancient populations, as well as values for the twenty-five closest modern populations

Supplementary Note 10: f-statistic analyses

Rosa Fregel, Fernando L. Méndez and María C. Ávila-Arcos

We performed analysis of outgroup f3-statistic using qp3Pop program from ADMIXTOOLS, to determine the amount of shared drift between two ancient populations PopA and PopB. For that, we chose an outgroup population that has not experience any post-divergence gene flow with either PopA or PopB as the target population (e.g. Jo'hoan North or Mbuti), and calculate the f3-statistic in the form f3(PopA, PopB; Outgroup). In this way, the result of the f3-statistic will be a positive value proportional to the length of the shared drift path of populations PopA and PopB with respect to the outgroup.

Results obtained for the outgroup f3-statistic are shown in Figure S10.1. As already observed in the PCA and the ADMIXTURE analysis, IAM shares more ancestry with KEB and with Levantine populations, such us Natufians and Levant_N. The fact that IAM could have links with Pre-Pottery Neolithic sites from the Levant had been already proposed based on archaeological evidence. Concretely, the funerary ritual observed in IAM, consisting on placing a millstone on the head of the deceased (Supplementary Note 1), is a uncommon funerary practice that has been only observed in Pre-Pottery Neolithic B sites in Cyprus.

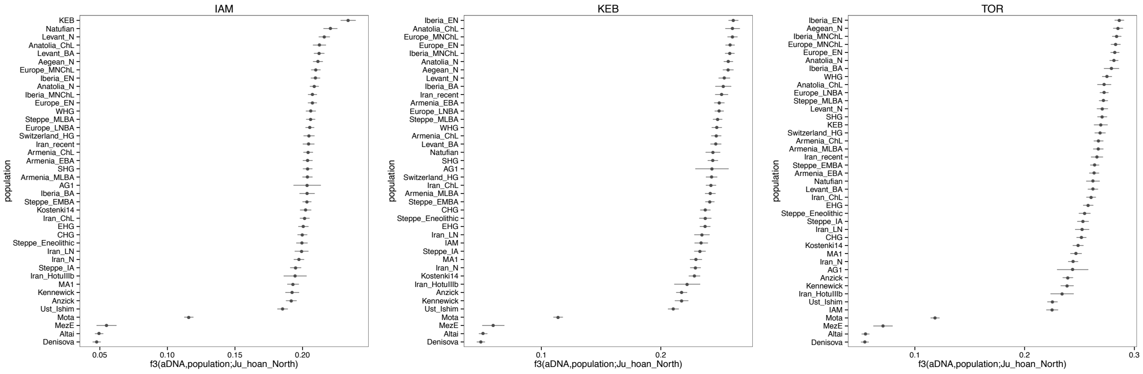


Figure S10.1 - Outgroup f3-statistic results for IAM, KEB and TOR

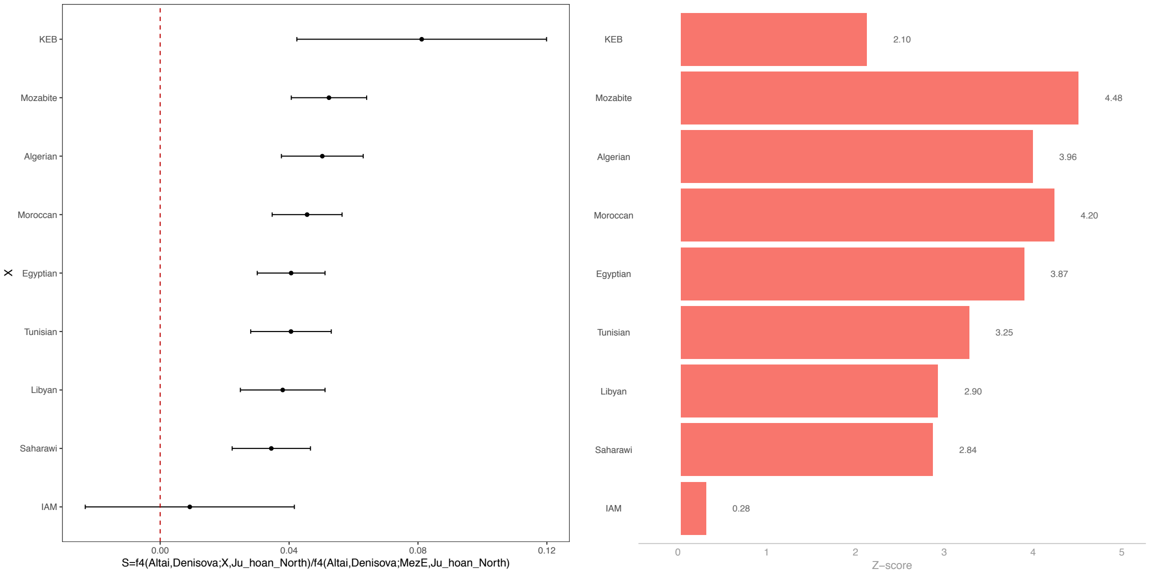
To further explore the connection of IAM with Levantine populations, we performed a f4-statistic test. When calculate it in the form f4(test, Outgroup; PopA, PopB), a positive f4-statistic value would indicate that the test population shares more alleles with PopA than with PopB. To demonstrate that IAM has a Levantine origin, rather than a local origin in Africa, we tested (IAM, Chimp; Levantine population, African population), with the

Levantine population being BedouinB, Levant_N or Natufian, and the African population being Jo’hoan North, Mbuti, Mota or Yoruba.

Test	Out	PopA	PopB	f4	SE	Z-score	SNP
IAM	Chimp	Natufian	Mota	0.0222	0.0014	16.41	58,754
IAM	Chimp	Natufian	Somali	0.0125	0.0010	12.661	58,771
IAM	Chimp	Natufian	Datog	0.0153	0.0011	13.767	58,766
IAM	Chimp	Natufian	Masai	0.0166	0.0010	16.388	58,771
IAM	Chimp	Natufian	Kikuyu	0.0196	0.0011	17.945	58,769
IAM	Chimp	Natufian	Mandenka	0.0229	0.0010	21.944	58,771
IAM	Chimp	Natufian	Gambian	0.0229	0.0011	20.723	58,770
IAM	Chimp	Natufian	BantuKenya	0.0232	0.0011	20.988	58,771
IAM	Chimp	Natufian	Luhya	0.0233	0.0011	21.861	58,771
IAM	Chimp	Natufian	Hadza	0.0237	0.0011	20.908	58,767
IAM	Chimp	Natufian	Esan	0.0238	0.0011	21.522	58,771
IAM	Chimp	Natufian	Yoruba	0.0238	0.0010	22.881	58,771
IAM	Chimp	Natufian	Luo	0.0238	0.0011	22.354	58,771
IAM	Chimp	Natufian	Mende	0.0245	0.0011	23.154	58,771
IAM	Chimp	Natufian	BantuSA	0.0264	0.0011	24.771	58,771
IAM	Chimp	Natufian	Biaka	0.0315	0.0011	28.668	58,771
IAM	Chimp	Natufian	Khomani	0.0315	0.0011	28.69	58,771
IAM	Chimp	Natufian	Mbuti	0.0339	0.0012	29.237	58,771
IAM	Chimp	Natufian	Ju_hoan_North	0.0362	0.0012	31.06	58,771
IAM	Chimp	Levant_N	Mota	0.0213	0.0010	21.22	98,502
IAM	Chimp	Levant_N	Somali	0.0109	0.0007	16.414	98,531
IAM	Chimp	Levant_N	Datog	0.0143	0.0008	19.066	98,527
IAM	Chimp	Levant_N	Masai	0.0153	0.0007	21.652	98,531
IAM	Chimp	Levant_N	Kikuyu	0.0181	0.0008	23.507	98,528
IAM	Chimp	Levant_N	Mandenka	0.0213	0.0008	28.31	98,531
IAM	Chimp	Levant_N	Gambian	0.0216	0.0008	27.292	98,530
IAM	Chimp	Levant_N	Luhya	0.0218	0.0008	28.868	98,531
IAM	Chimp	Levant_N	BantuKenya	0.0220	0.0008	27.941	98,531
IAM	Chimp	Levant_N	Esan	0.0222	0.0008	28.658	98,531
IAM	Chimp	Levant_N	Yoruba	0.0223	0.0007	30.772	98,531
IAM	Chimp	Levant_N	Luo	0.0225	0.0008	29.498	98,531
IAM	Chimp	Levant_N	Hadza	0.0225	0.0008	27.82	98,525
IAM	Chimp	Levant_N	Mende	0.0233	0.0008	30.102	98,531
IAM	Chimp	Levant_N	BantuSA	0.0248	0.0008	31.857	98,531
IAM	Chimp	Levant_N	Biaka	0.0300	0.0008	38.023	98,531
IAM	Chimp	Levant_N	Khomani	0.0303	0.0008	38.461	98,531
IAM	Chimp	Levant_N	Mbuti	0.0325	0.0008	39.127	98,531
IAM	Chimp	Levant_N	Ju_hoan_North	0.0349	0.0009	40.413	98,530
IAM	Chimp	BedouinB	Mota	0.0202	0.0008	26.07	127,834
IAM	Chimp	BedouinB	Somali	0.0092	0.0004	24.92	127,877
IAM	Chimp	BedouinB	Datog	0.0123	0.0005	24.101	127,866
IAM	Chimp	BedouinB	Masai	0.0136	0.0004	32.648	127,877
IAM	Chimp	BedouinB	Kikuyu	0.0164	0.0005	31.581	127,873
IAM	Chimp	BedouinB	Mandenka	0.0199	0.0005	42.177	127,877
IAM	Chimp	BedouinB	Gambian	0.0201	0.0005	38.179	127,876
IAM	Chimp	BedouinB	Luhya	0.0203	0.0005	41.058	127,877
IAM	Chimp	BedouinB	BantuKenya	0.0204	0.0005	38.124	127,877
IAM	Chimp	BedouinB	Esan	0.0206	0.0005	42.247	127,877
IAM	Chimp	BedouinB	Hadza	0.0207	0.0006	37.432	127,869
IAM	Chimp	BedouinB	Luo	0.0207	0.0005	42.52	127,877
IAM	Chimp	BedouinB	Yoruba	0.0208	0.0005	46.113	127,877
IAM	Chimp	BedouinB	Mende	0.0218	0.0005	44.201	127,877
IAM	Chimp	BedouinB	BantuSA	0.0232	0.0005	45.715	127,877
IAM	Chimp	BedouinB	Biaka	0.0286	0.0005	56.754	127,877
IAM	Chimp	BedouinB	Khomani	0.0292	0.0005	55.454	127,877
IAM	Chimp	BedouinB	Mbuti	0.0311	0.0006	55.898	127,877
IAM	Chimp	BedouinB	Ju_hoan_North	0.0337	0.0006	56.291	127,876

Table S10.1 - Results for the f4-statistic test for IAM, comparing Levantine and African populations

All comparisons are positive, with high significant Z scores, indicating IAM is more related to Levantine than to African populations (Table S10.1). We also added additional evidence of IAM being related to the out-of-Africa migration by testing if it has Neanderthal introgression, an event that happened after modern humans left Africa. It has been demonstrated that modern North Africa populations have admixture with Neandethals. However, it is possible that IAM lacked Neanderthal admixture and the signal observed today is due to migration influx into North Africa from the Middle East and Europe in historical times. We estimate Neanderthal admixture as in Lazaridis et al.[57] using the S-statistic: f4(Neanderthal_1, Denisova; IAM, Ju_hoan_North) / f4(Neanderthal_1, Denisova; Neanderthal_2, Ju_hoan_North), being "Denisova" the high-coverage genome from the Denisovan archaic sample, "Neandethal_1" the high-coverage genome from Altai (52X)[115] and "Neanderthal_2" the combined low-coverage genome from three individuals from Vindija Cave (1.3X)[116] or the low-coverage Mezmaiskaya genome (0.5X)[115]. We also included KEB and modern samples from North Africa in the analysis for comparison.



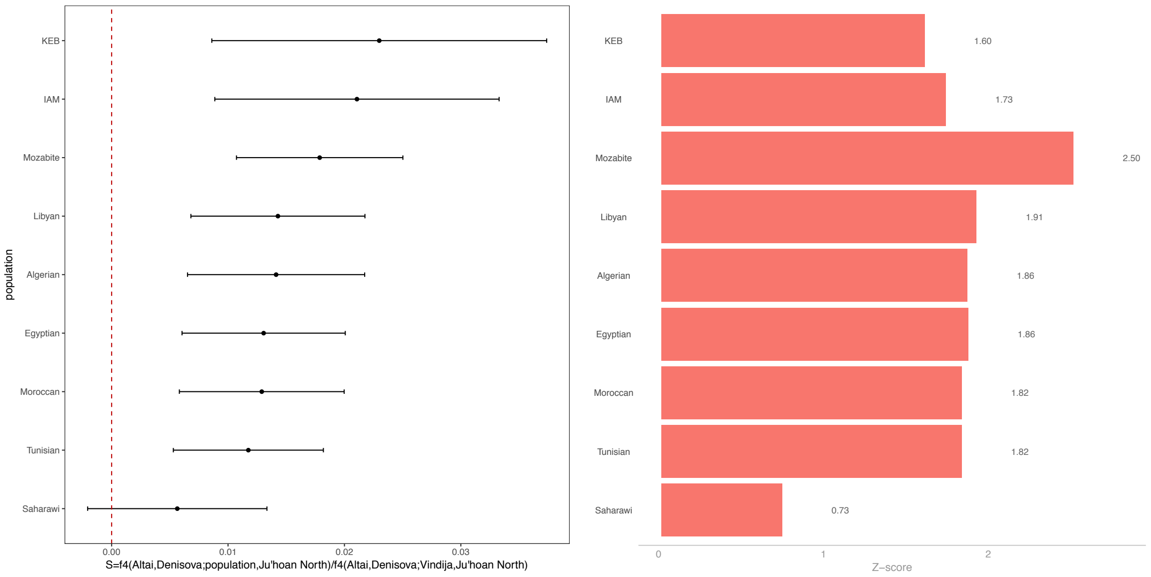
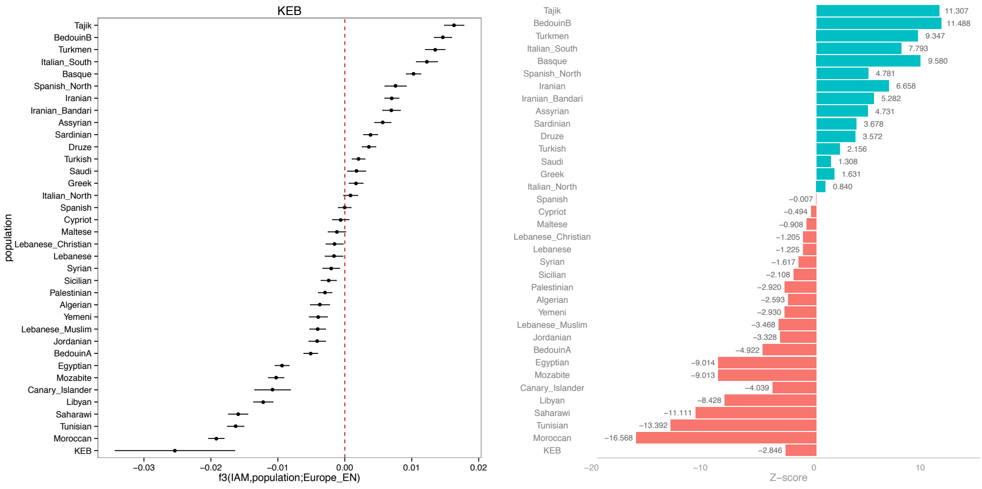
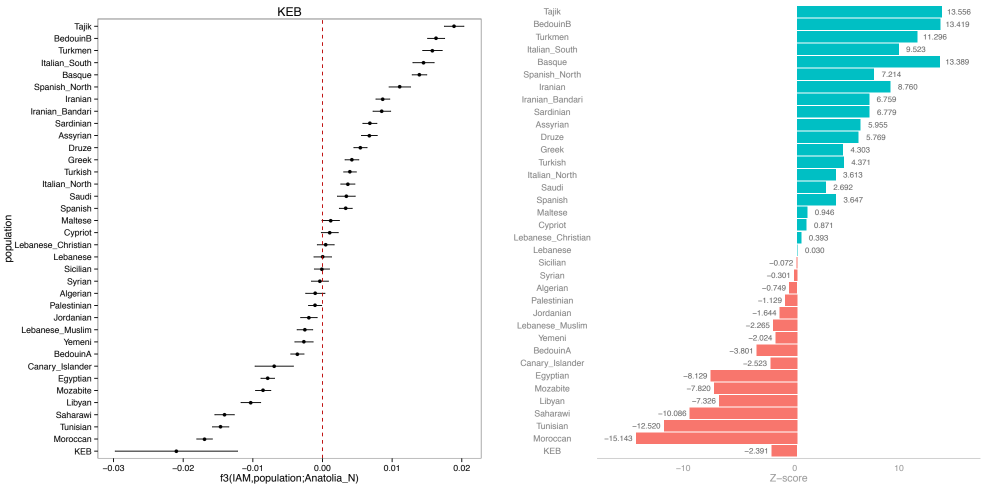


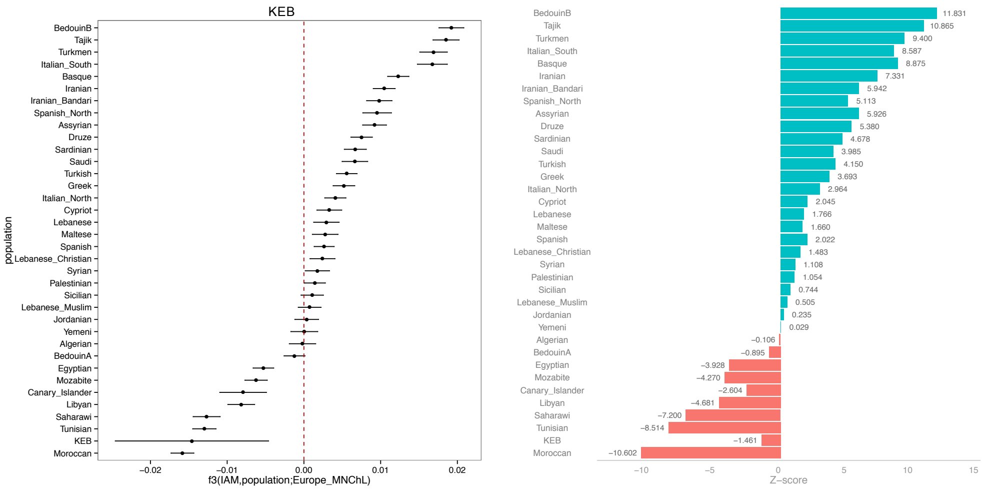
Figure S10.2 - S-statistic result for testing the admixture with Neanderthal in IAM and other populations

Results of the S-statistic test for identifying Neanderthal admixture are shown in Figure S10.2. When using the Altai and Mezmaiskaya Neanderthal genomes, it is possible to detect Neanderthal introgression into KEB, but it is impossible to say so in IAM, as the error bars overlaps with 0. However, when we used Altai, with the Vindija Cave genome with a higher coverage than Mezmaiskaya, the introgression signal in IAM and KEB is clear, pointing to an ancestry outside of Africa.

Several evidence point to KEB being a mixture of IAM-like component and a European Neolithic component. The unsupervised clustering place the non-IAM ancestry of KEB as being related to the purple component, present in Early Neolithic from Iberia and Middle Neolithic/Chalcolithic site from Europe. Although KEB shows shared ancestry with IAM, the outgroup-f3 analysis indicates that KEB shares more drift with Neolithic and Chalcolithic populations from Anatolia and Europe, with highest f3-statistic value for Iberia_N (Figure S10.1). To test if KEB population can be explained as a mixture of IAM and other populations, we used an admixture f3 test. We calculated the f3 in the form f3(PopA, PopB; Target), with PopA being IAM, PopB being Iberian Early Neolithic (Iberia_EN) or Iberian Middle Neolithic/Chalcolithic (Iberia_MNChL) and Target being KEB or any modern populations of North Africa and Middle East.


KEB
population
Tajik
BedouinB
Turkmen
Italian_South
Basque
Spanish_North
Iranian
Iranian_Bandari
Assyrian
Sardinian
Druze
Turkish
Saudi
Greek
Italian_North
Spanish
Cypriot
Maltese
Lebanese_Christian
Lebanese
Syrian
Sicilian
Palestinian
Algerian
Yemeni
Lebanese_Muslim
Jordanian
BedouinA
Egyptian
Mozabite
Canary_Islander
Libyan
Saharawi
Tunisian
Moroccan
KEB
−0.03
−0.02
−0.01
0.00
0.01
0.02
f3(IAM,population;Europe_EN)
Tajik 11.307
BedouinB 11.488
Turkmen 9.347
Italian_South 7.793
Basque 9.580
Spanish_North 4.781
Iranian 6.658
Iranian_Bandari 5.282
Assyrian 4.731
Sardinian 3.678
Druze 3.572
Turkish 2.156
Saudi 1.308
Greek 1.631
Italian_North 0.840
Spanish −0.007
Cypriot −0.494
Maltese −0.908
Lebanese_Christian −1.205
Lebanese −1.225
Syrian −1.617
Sicilian −2.108
Palestinian −2.920
Algerian −2.593
Yemeni −2.930
Lebanese_Muslim −3.468
Jordanian −3.328
BedouinA −4.922
Egyptian −9.014
Mozabite −9.013
Canary_Islander −4.039
Libyan −8.428
Saharawi −11.111
Tunisian −13.392
Moroccan −16.568
KEB −2.846
−20
−10
0
10
Z−score


KEB
population
Tajik
BedouinB
Turkmen
Italian_South
Basque
Spanish_North
Iranian
Iranian_Bandari
Sardinian
Assyrian
Druze
Greek
Turkish
Italian_North
Saudi
Spanish
Maltese
Cypriot
Lebanese_Christian
Lebanese
Sicilian
Syrian
Algerian
Palestinian
Jordanian
Lebanese_Muslim
Yemeni
BedouinA
Canary_Islander
Egyptian
Mozabite
Libyan
Saharawi
Tunisian
Moroccan
KEB
−0.03
−0.02
−0.01
0.00
0.01
0.02
f3(IAM,population;Anatolia_N)
Tajik 13.556
BedouinB 13.419
Turkmen 11.296
Italian_South 9.523
Basque 13.389
Spanish_North 7.214
Iranian 8.760
Iranian_Bandari 6.759
Sardinian 6.779
Assyrian 5.955
Druze 5.769
Greek 4.303
Turkish 4.371
Italian_North 3.613
Saudi 2.692
Spanish 3.647
Maltese 0.946
Cypriot 0.871
Lebanese_Christian 0.393
Lebanese 0.030
Sicilian −0.072
Syrian −0.301
Algerian −0.749
Palestinian −1.129
Jordanian −1.644
Lebanese_Muslim −2.265
Yemeni −2.024
BedouinA −3.801
Canary_Islander −2.523
Egyptian −8.129
Mozabite −7.820
Libyan −7.326
Saharawi −10.086
Tunisian −12.520
Moroccan −15.143
KEB −2.391
−10
0
10
Z−score


KEB
population
BedouinB
Tajik
Turkmen
Italian_South
Basque
Iranian
Iranian_Bandari
Spanish_North
Assyrian
Druze
Sardinian
Saudi
Turkish
Greek
Italian_North
Cypriot
Lebanese
Maltese
Spanish
Lebanese_Christian
Syrian
Palestinian
Sicilian
Lebanese_Muslim
Jordanian
Yemeni
Algerian
BedouinA
Egyptian
Mozabite
Canary_Islander
Libyan
Saharawi
Tunisian
KEB
Moroccan
−0.02
−0.01
0.00
0.01
0.02
f3(IAM,population;Europe_MNChL)
BedouinB 11.831
Tajik 10.865
Turkmen 9.400
Italian_South 8.587
Basque 8.875
Iranian 7.331
Iranian_Bandari 5.942
Spanish_North 5.113
Assyrian 5.926
Druze 5.380
Sardinian 4.678
Saudi 3.985
Turkish 4.150
Greek 3.693
Italian_North 2.964
Cypriot 2.045
Lebanese 1.766
Maltese 1.660
Spanish 2.022
Lebanese_Christian 1.483
Syrian 1.108
Palestinian 1.054
Sicilian 0.744
Lebanese_Muslim 0.505
Jordanian 0.235
Yemeni 0.029
Algerian −0.106
BedouinA −0.895
Egyptian −3.928
Mozabite −4.270
Canary_Islander −2.604
Libyan −4.681
Saharawi −7.200
Tunisian −8.514
KEB −1.461
Moroccan −10.602
−10
−5
0
5
10
15
Z−score

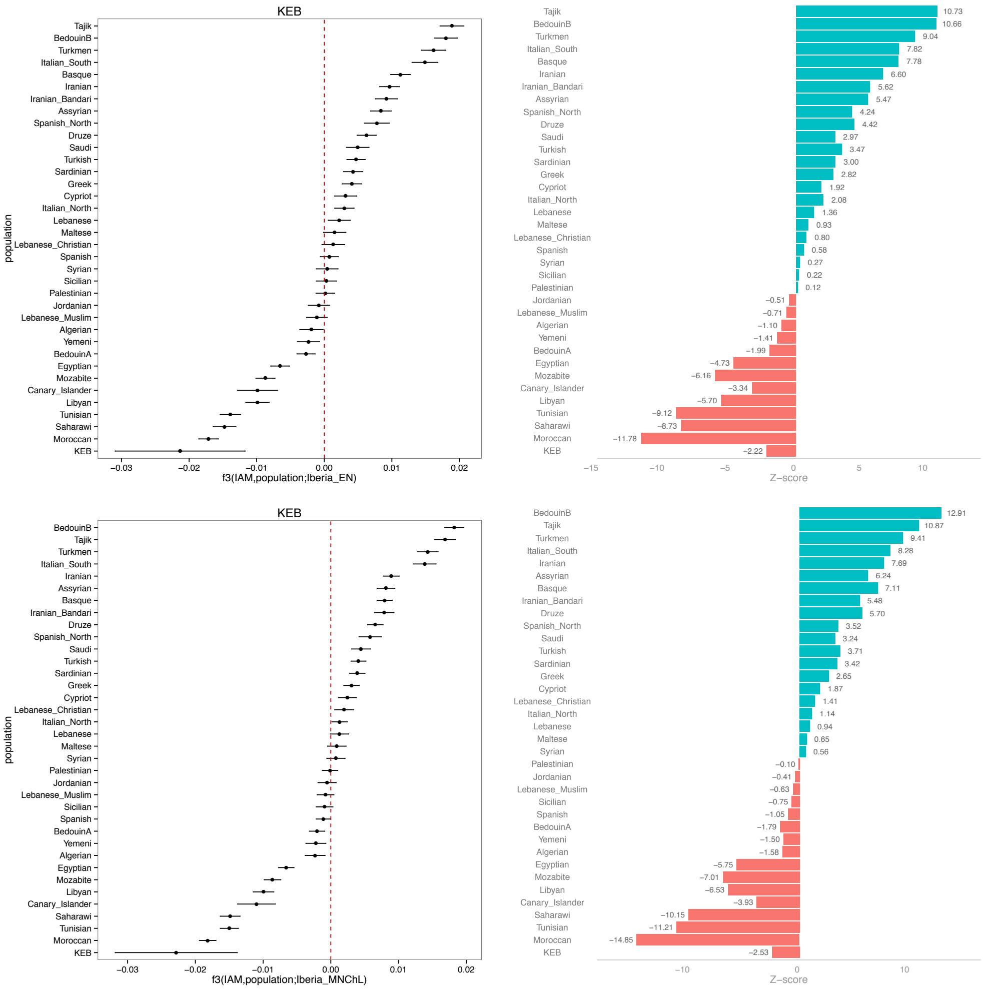


Figure S10.3 - Admixture f3-statistic results for KEB and modern populations from North Africa and the Middle East

For KEB, both comparisons produced a f3-statistic negative value and the Z scores were negative and close to -3 (Figure S10.3). Although this result does not reach significance, we have to take into account we are comparing populations with low-coverage samples. Additionally, this result is also in accordance with the archaeological record, with Late Neolithic sites in North Africa presenting pottery and ivory tools similar to those associated to Iberian Neolithic. Although this is a too simplistic model for modern populations, which were later affected by historical migrations that diluted the prehistorical European component, f3-statistic values are negative and Z scores significant for most of the North African populations, with a higher signal in those populations of the west.

As expected, TOR has more shared ancestry with Iberia_EN as well as other Neolithic and Chalcolithic population from Europe (Figure S10.1). Archaeological work in southern Iberia has pointed out a specific Early Neolithic culture, previous to the Cardial expansion, having similarities with farmer traditions in the Maghreb, and suggesting a North African source. From our results, we observed that TOR has a similar genetic composition than other Neolithic populations from Iberia. Based on the results observed in KEB, we know similarities observed between the Almagra pottery and farmer traditions in the Maghreb, most probably respond to European Neolithic influence in North Africa before 3,000 BCE. However, it is also possible that farmer in North Africa migrated to Europe. From the genetic point of view, the idea of this prehistorical North African influx in Iberia is sustained by the presence of old European-specific lineages of North African origin. We also observed IAM-like ancestry in modern populations from southern Europe, although that can be related to the historic Arab expansion. For testing that, we analyzed a f3-statistic test in the form f3(IAM, Anatolia_N; Test), being Test all Neolithic populations from Europe, modern southern Europe populations with IAM-like ancestry and KEB and the Canary Islands as positive controls for comparison.

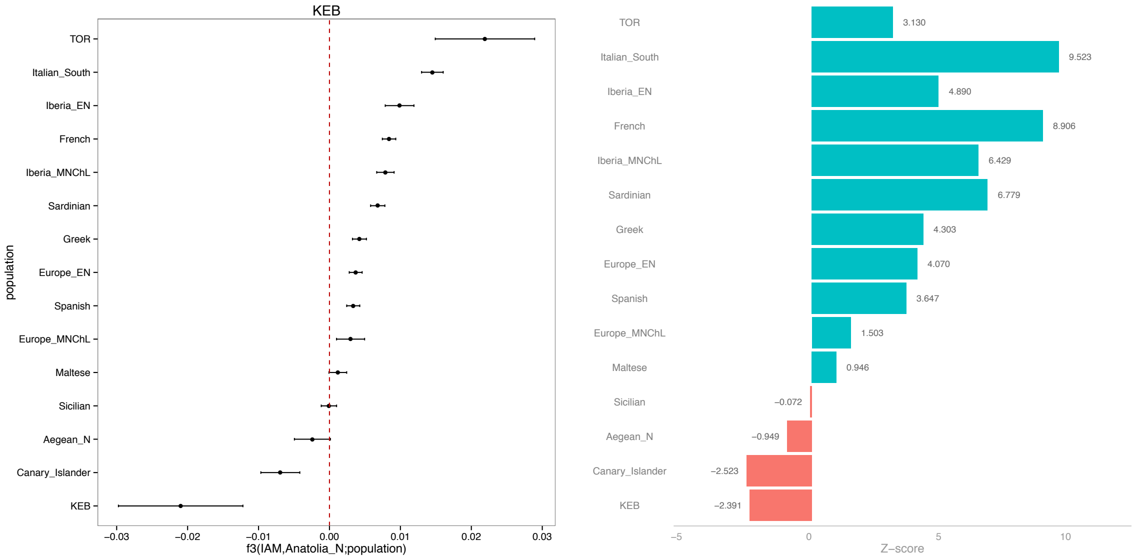


Figure S10.4 - Admixture f3-statistic results for TOR and modern populations with IAM-like ancestry

We only obtain values close to significance for KEB and for the Canary Islands (Figure S10.4). For Neolithic samples from the Aegean region the f3-statistic value is negative but the Z score is practically zero. For all the other comparisons including TOR, f3 is positive.

Supplementary Note 11: Phenotype analyses

Rosa Fregel, Fernando L. Méndez and María C. Ávila-Arcos

For analyzing phenotypic variants in our aDNA samples, we obtained a list of SNPs related to different phenotypes. As we are dealing with low-coverage genomes, we merged all the bam files from the same site together (IAM, KEB and TOR), and analyzed the alleles present at a population level. We obtained the SNPs of interest using samtools mpileup -l option, filtering bases with BASEQ > 30 and using the bam files with 4 bp trimmed at both ends.

The SNP rs1426654 of the SLC24A5 gene is related to light-skinned pigmentation in individuals with European ancestry. The derived allele was already fixed in Anatolian Early Neolithic populations, suggesting that its high frequency in Neolithic Europe was related to demic diffusion from the Middle East[51]. In our sample set, TOR presents the derived A allele (3 reads), while IAM has the ancestral G allele (2 reads) related to dark-skinned phenotype. However, in line with previous results, KEB is similar to European Neolithic samples, and presents the derived allele (2 reads). Other SNP associated to skin pigmentation in Europe is rs16891982 of the SLC45A2 gene. In this case, the having both derived G alleles will determine light skin. Both IAM and KEB show the C allele (3 and 1 reads, respectively), while TOR has the G allele (4 reads). The rs16891982 SNP of the OCA2 gene is also related to pigmentation. This SNP is one of several variants associated with increased probability of having dark eye color in Caucasians. Again, IAM present the T allele (1 read) related to dark pigmentation, while KEB and TOR presents the C allele (3 reads both) with increased probability of having light colored eyes. Another SNP of the OCA2 gene, rs12913832, was fixed in Europe during the Mesolithic and it is the major determinant for blue eye color. In our sample set, only IAM has coverage for that position and all the three reads show the ancestral allele, indicating IAM people had dark eyes.

Lactose persistence occurs when humans are able to digest milk as adults. Different MCM6-gene SNPs are responsible for lactase persistence in different geographic areas: while rs4988235 is mostly responsible for lactase persistence in Europe, rs145946881 is related to lactase tolerance in Sub-Saharan Africa. Although it was thought that lactase persistence in Europe was acquired at early stages of the Neolithic revolution, it has been determined that it only dates to the last 4,000 years ago and appeared for the first time in Chalcolithic samples from Central Europe[51]. Regarding the African variant, haplotype analysis indicates an eastern African origin[117], although it was absent in the 4,500-years-old Mota genome from Ethiopia[118]. We only have coverage for rs4988235 in TOR samples and for rs145946881 in IAM. Samples from TOR present the ancestral allele (3 reads) for the rs4988235 SNP, which is congruent giving their radiocarbon date of ~5,000 BCE. On the other hand, IAM presents the reference allele for rs145946881, indicating they did not have the variant linked to lactose persistence. Sadly, IAM did not have coverage for the two SNPs associated to lactase persistence in the Middle East: rs41525747 and rs41380347.

REFERENCES:

1 Bokbot, Y. & Ben-Ncer, A. in *Bell Beaker in Everyday Life* (eds M. Baioni *et al.*) 327–330 (Museo Fiorentino di Preistoria 'Paolo Graziosi', 2008).

2 Laviano, F. *La faune néolithique du site d'Ifri n'Amr o'Moussa (Oued Beth, plateau de Zemmour, Maroc) : méthodologie appliquée à une stratigraphie perturbée*, Université Paul Valéry Montpellier, (2015).

3 Ben-Ncer, A., Bokbot, Y., Amani, F. & Ouachi, M. in *Proceedings of the II Meeting of African Prehistory* (eds M Sahnouni, S. Semaw, & J. Rios-Garaizar) (Centro Nacional de Investigación sobre la Evolución Humana, 2015).

4 Ben-Ncer, A., Bokbot, Y., Amani, F. & Ouachi, M. in *Around the Petit-Chasseur Site in Sion (Valais, Switzerland) and New Approaches to the Bell Beaker Culture* (ed M. Besse) 251-258 (Archaeopress Archaeology, 2011).

5 Martínez-Sánchez, R. M., Vera-Rodríguez, J. C., Pérez-Jordà, G., Peña-Chocarro, L. & Bokbot, Y. The beginning of the Neolithic in northwestern Morocco. *Quaternary International* **In press** (2017).

6 Bailloud, G. & Mieg de Boofzheim, P. La nécropole néolithique d'El Kiffen, près des Tamaris (province de Casablanca, Maroc). *Lybica* **XII**, 95-171 (1964).

7 Lacombe, J. P., El Hajraoui, A. & Daugas, J. P. Etude antropologique préliminaire des sépultures néolithiques de la grotte d'El Mnasra (Témara, Maroc). *Bulletin de la Societé d'Anthropologie du Sud-Ouest* **XVVI**, 163-176 (1991).

8 Ben-Ncer, A., Oujaa, A., El Hajraoui, M. A. & Nespoulet, R. Etude archéothanatologique de La Sépulture 4 d'El Harhoura II. *Antropo* **27**, 87-96 (2012).

9 Vassos, K. in *Bulletin de correspondance hellénique* Vol. 105 967-1024 (1981).

10 Reimer, P. J. *et al.* IntCal13 and Marine13 radiocarbon age calibration curves 0–50,000 years cal BP. *Radiocarbon* **55**, 1869-1887 (2013).

11 Bronk-Ramsay, C. Bayesian analysis of radiocarbon dates. *Radiocarbon* **51**, 337-360 (2009).

12 Mikdad, A. in *Beiträge zur Allgemeinen und Vergleichenden Archäologie* Vol. 18 (ed KAVA) 243-252 (1998).

13 De Wailly, A. Le Kef el Baroud et l'ancienneté de l'introduction du cuivre au Maroc. *Bulletins et Mémoires de l'Archéologie Marocaine* **10**, 49 (1976).

14 Martin-Socas, D., Camalich-Massieu, M. D. & Gonzalez, P. *La Cueva de El Toro (Sierra de El Torcal-Antequera-Málaga). Un modelo de ocupación ganadera en el territorio andaluz entre el VI y II milenios A.N.E. Arqueología. Monografías*. (Consejería de Cultura Junta de Andalucía, 2004).

15 Harris, E. C. *Principles of archaeological stratigraphy*. (Academic Press, 1979).

16 Eguez-Gordon, N., Mallol-Duque, C., Martin-Socas, D. & Camalich-Massieu, M. D. Radiometric dating and micromorphological evidence for domestic activity and sheep penning in a Neolithic cave: Cueva del Toro (Málaga, Antequera, Spain). *Archaeological and Anthropological Sciences* **8**, 107-123 (2016).

17 Martin-Socas, D., Camalich-Massieu, M. D., Caro-Herrero, J. L. & Rodriguez-Santos, F. J. The beginning of the Neolithic in Andalusia. *Quaternary International* **in press**, doi:10.1016/j.quaint.2017.06.057 (2017).

18 Camalich-Massieu, M. D. & Martin-Socas, D. Los inicios del Neolítico en Andalucía. Tradición e Innovación. *Menga. Revista de Prehistoria de Andalucía* **4**, 103-129 (2013).

19 Martin-Socas, D., Camalich-Massieu, M. D. & Rodriguez-Santos, F. J. in *Antequera Milenaria: Historia de una Tierra* Vol. Volumen 1. La Prehistoria (ed L. García-Sanjuán) (Real Academia de Nobles Artes de Antequera, 2017).

20 Robb, J. *et al.* Cleaning the dead: Neolithic ritual processing of human bone at Scaloria Cave, Italy. *Antiquity* **89**, 39–54 (2014).

21 Olaria, C. *Las Cuevas de los Botijos y de la Zorrera en Benalmadena*. (Museo Arqueologico de Benalmadena, 1977).

22 Hansen, H. B. *et al.* Comparing Ancient DNA Preservation in Petrous Bone and Tooth Cementum. *PLoS One* **12**, e0170940, doi:10.1371/journal.pone.0170940 (2017).

23 Adler, C. J., Haak, W., Donlon, D., Cooper, A. & Consortium, T. G. Survival and recovery of DNA from ancient teeth and bones. *Journal of Archaeological Science* **38**, 956–964 (2011).

24 Dabney, J. *et al.* Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc Natl Acad Sci U S A* **110**, 15758-15763, doi:10.1073/pnas.1314445110 (2013).

25 Korlevic, P. *et al.* Reducing microbial and human contamination in DNA extractions from ancient bones and teeth. *Biotechniques* **59**, 87-93, doi:10.2144/000114320 (2015).

26 Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor protocols* **2010**, pdb prot5448, doi:10.1101/pdb.prot5448 (2010).

27 Pinhasi, R. *et al.* Optimal Ancient DNA Yields from the Inner Ear Part of the Human Petrous Bone. *PLoS One* **10**, e0129102, doi:10.1371/journal.pone.0129102 (2015).

28 Mundorff, A. Z., Bartelink, E. J. & Mar-Cash, E. DNA preservation in skeletal elements from the World Trade Center disaster: recommendations for mass fatality management. *J Forensic Sci* **54**, 739-745, doi:10.1111/j.1556-4029.2009.01045.x (2009).

29 Carpenter, M. L. *et al.* Pulling out the 1%: Whole-Genome Capture for the Targeted Enrichment of Ancient DNA Sequencing Libraries. *American Journal of Human Genetics* **93**, 852-864, doi:10.1016/j.ajhg.2013.10.002 (2013).

30 Lindgreen, S. AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC research notes* **5**, 337, doi:10.1186/1756-0500-5-337 (2012).

31 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

32 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

33 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-498, doi:10.1038/ng.806 (2011).

34 Ginolhac, A., Rasmussen, M., Gilbert, M. T., Willerslev, E. & Orlando, L. mapDamage: testing for damage patterns in ancient DNA sequences. *Bioinformatics* **27**, 2153-2155, doi:10.1093/bioinformatics/btr347 (2011).

35 Fu, Q. *et al.* Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**, 445-449, doi:10.1038/nature13810 (2014).

36 Andrews, R. M. *et al.* Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet* **23**, 147 (1999).

37 Kloss-Brandstatter, A. *et al.* HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Hum Mutat* **32**, 25-32, doi:10.1002/humu.21382 (2011).

38 van Oven, M. & Kayser, M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum Mutat* **30**, E386-394 (2009).

39 Milne, I. *et al.* Using Tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics* **14**, 193-202 (2013).

40 Hall, T. A. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic acids symposium series* **41**, 95-98 (1999).

41 Fregel, R. & Delgado, S. HaploSearch: a tool for haplotype-sequence two-way transformation. *Mitochondrion* **11**, 366-367 (2011).

42 Bandelt, H. J., Forster, P. & Rohl, A. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* **16**, 37-48 (1999).

43 Hervella, M. *et al.* The mitogenome of a 35,000-year-old Homo sapiens from Europe supports a Palaeolithic back-migration to Africa. *Sci Rep* **6**, 25501, doi:10.1038/srep25501 (2016).

44 Secher, B. *et al.* The history of the North African mitochondrial DNA haplogroup U6 gene flow into the African, Eurasian and American continents. *BMC evolutionary biology* **14**, 109, doi:1471-2148-14-109 [pii] (2014).

45 Pennarun, E. *et al.* Divorcing the Late Upper Palaeolithic demographic histories of mtDNA haplogroups M1 and U6 in Africa. *BMC Evol Biol* **12**, 234, doi:10.1186/1471-2148-12-234.; eng; ID: 4230 (2012).

46 Lazaridis, I. *et al.* Genomic insights into the origin of farming in the ancient Near East. *Nature* **536**, 419-424, doi:10.1038/nature19310 (2016).

47 Brandt, G. *et al.* Ancient DNA reveals key stages in the formation of central European mitochondrial genetic diversity. *Science* **342**, 257-261, doi:10.1126/science.1241844 (2013).

48 Soares, P. *et al.* Correcting for purifying selection: an improved human mitochondrial molecular clock. *Am J Hum Genet* **84**, 740-759 (2009).

49 Reidla, M. *et al.* Origin and diffusion of mtDNA haplogroup X. *Am J Hum Genet* **73**, 1178-1190 (2003).

50 Shlush, L. I. *et al.* The Druze: a population genetic refugium of the Near East. *PLoS One* **3**, e2105, doi:10.1371/journal.pone.0002105 (2008).

51 Mathieson, I. *et al.* Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**, 499-503, doi:10.1038/nature16152 (2015).

52 Fu, Q., Rudan, P., Paabo, S. & Krause, J. Complete mitochondrial genomes reveal neolithic expansion into Europe. *PLoS One* **7**, e32473, doi:10.1371/journal.pone.0032473 (2012).

53 Costa, M. D. *et al.* A substantial prehistoric European ancestry amongst Ashkenazi maternal lineages. *Nat Commun* **4**, 2543, doi:10.1038/ncomms3543 (2013).

54 Plaza, S. *et al.* Joining the pillars of Hercules: mtDNA sequences show multidirectional gene flow in the western Mediterranean. *Ann Hum Genet* **67**, 312-328 (2003).

55 Pala, M. *et al.* Mitochondrial DNA signals of late glacial recolonization of Europe from near eastern refugia. *Am J Hum Genet* **90**, 915-924, doi:10.1016/j.ajhg.2012.04.003.; eng; ID: 4019 (2012).

56 Haak, W. *et al.* Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207-211, doi:10.1038/nature14317 (2015).

57 Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409-413, doi:10.1038/nature13673 (2014).

58 Haak, W. *et al.* Ancient DNA from European Early Neolithic Farmers Reveals Their Near Eastern Affinities. *Plos Biology* **8**, e1000536-e1000536, doi:10.1371/journal.pbio.1000536 (2010).

59 Haak, W. *et al.* Ancient DNA from the first European farmers in 7500-year-old Neolithic sites. *Science* **310**, 1016-1018 (2005).

60 Cassidy, L. M. *et al.* Neolithic and Bronze Age migration to Ireland and establishment of the insular Atlantic genome. *Proc Natl Acad Sci U S A* **113**, 368-373, doi:10.1073/pnas.1518445113 (2016).

61 Hofmanova, Z. *et al.* Early farmers from across Europe directly descended from Neolithic Aegeans. *Proc Natl Acad Sci U S A* **113**, 6886-6891, doi:10.1073/pnas.1523951113 (2016).

62 Maca-Meyer, N. *et al.* Ancient mtDNA analysis and the origin of the Guanches. *Eur J Hum Genet* **12**, 155-162 (2004).

63 Fregel, R. *et al.* Demographic history of Canary Islands male gene-pool: replacement of native lineages by European. *BMC Evol Biol* **9**, 181 (2009).

64 Kéfi, R., Stevanovitch, A., Bouzaid, E. & Colomb, B. Diversité mitochondriale de la population de Taforalt (12.000 ans bp - Maroc): une approche génétique à l'étude du peuplement de l'Afrique du Nord. *Anthropologie* **43**, 1-11 (2005).

65 Fernandes, V. *et al.* The Arabian cradle: mitochondrial relicts of the first steps along the southern route out of Africa. *American Journal of Human Genetics* **90**, 347-355, doi:10.1016/j.ajhg.2011.12.010 [doi] (2012).

66 Gamba, C. *et al.* Genome flux and stasis in a five millennium transect of European prehistory. *Nat Commun* **5**, 5257, doi:10.1038/ncomms6257 (2014).

67 Szécsényi-Nagy, A. *et al.* Tracing the genetic origin of Europe's first farmers reveals insights into their social organization. *Proceedings of the Royal Society B: Biological Sciences* **282**, doi:10.1098/rspb.2015.0339 (2015).

68 Lacan, M. *et al.* Ancient DNA reveals male diffusion through the Neolithic Mediterranean route. *Proceedings of the National Academy of Sciences* **108**, 9788-9791, doi:10.1073/pnas.1100723108 (2011).

69 De Benedetto, G. *et al.* Mitochondrial DNA sequences in prehistoric human remains from the Alps. *Eur J Hum Genet* **8**, 669-677, doi:10.1038/sj.ejhg.5200514 (2000).

70 Lorkiewicz, W. *et al.* Between the Baltic and Danubian Worlds: the genetic affinities of a Middle Neolithic population from central Poland. *PLoS One* **10**, e0118316, doi:10.1371/journal.pone.0118316 (2015).

71 Witas, H. W. *et al.* Hunting for the LCT-13910*T allele between the Middle Neolithic and the Middle Ages suggests its absence in dairying LBK people entering the Kuyavia region in the 8th millennium BP. *PLoS One* **10**, e0122384, doi:10.1371/journal.pone.0122384 (2015).

72 Lacan, M. *et al.* Ancient DNA suggests the leading role played by men in the Neolithic dissemination. *Proc Natl Acad Sci U S A* **108**, 18255-18259, doi:10.1073/pnas.1113061108 (2011).

73 Alt, K. W. *et al.* A Community in Life and Death: The Late Neolithic Megalithic Tomb at Alto de Reinoso (Burgos, Spain). *PLoS One* **11**, e0146176, doi:10.1371/journal.pone.0146176 (2016).

74 Malmström, H. *et al.* Ancient mitochondrial DNA from the northern fringe of the Neolithic farming expansion in Europe sheds light on the dispersion process. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370**, doi:10.1098/rstb.2013.0373 (2015).

75 Allentoft, M. E. *et al.* Population genomics of Bronze Age Eurasia. *Nature* **522**, 167-172, doi:10.1038/nature14507 (2015).

76 Gomez-Sanchez, D. *et al.* Mitochondrial DNA from El Mirador cave (Atapuerca, Spain) reveals the heterogeneity of Chalcolithic populations. *PLoS One* **9**, e105105, doi:10.1371/journal.pone.0105105 (2014).

77 Wilde, S. *et al.* Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. *Proc Natl Acad Sci U S A* **111**, 4832-4837, doi:10.1073/pnas.1316513111 (2014).

78 Broushaki, F. *et al.* Early Neolithic genomes from the eastern Fertile Crescent. *Science* **353**, 499-503, doi:10.1126/science.aaf7943 (2016).

79 Rivollat, M. *et al.* When the waves of European Neolithization met: first paleogenetic evidence from early farmers in the southern Paris Basin. *PLoS One* **10**, e0125521, doi:10.1371/journal.pone.0125521 (2015).

80 Rivollat, M. *et al.* Ancient mitochondrial DNA from the middle neolithic necropolis of Obernai extends the genetic influence of the LBK to west of the Rhine. *Am J Phys Anthropol* **161**, 522-529, doi:10.1002/ajpa.23055 (2016).

81 Juras, A. *et al.* Investigating kinship of Neolithic post-LBK human remains from Krusza Zamkowa, Poland using ancient DNA. *Forensic Sci Int Genet* **26**, 30-39, doi:10.1016/j.fsigen.2016.10.008 (2017).

82 Goncalves, D., Granja, R., Alves-Cardoso, F. & Carvalho, A. F. All different, all equal: Evidence of a heterogeneous Neolithic population at the Bom Santo Cave necropolis (Portugal). *Homo* **67**, 203-215, doi:10.1016/j.jchb.2015.12.004 (2016).

83 Olalde, I. *et al.* A Common Genetic Origin for Early Farmers from Mediterranean Cardial and Central European LBK Cultures. *Mol Biol Evol* **32**, 3132-3142, doi:10.1093/molbev/msv181 (2015).

84 Skoglund, P. *et al.* Genomic diversity and admixture differs for Stone-Age Scandinavian foragers and farmers. *Science* **344**, 747-750, doi:10.1126/science.1253448 (2014).

85 Kilinc, G. M. *et al.* The Demographic Development of the First Farmers in Anatolia. *Curr Biol* **26**, 2659-2666, doi:10.1016/j.cub.2016.07.057 (2016).

86 Behar, D. M. *et al.* Counting the founders: the matrilineal genetic ancestry of the Jewish Diaspora. *PLoS ONE* **3**, e2062, doi:10.1371/journal.pone.0002062.; eng; ID: 4392 (2008).

87 Skoglund, P., Stora, J., Gotherstrom, A. & Jakobsson, M. Accurate sex identification of ancient human remains using DNA shotgun sequencing. *Journal of Archaeological Science* **40**, 4477-4482 (2013).

88 Schroeder, H. *et al.* Genome-wide ancestry of 17th-century enslaved Africans from the Caribbean. *Proc Natl Acad Sci U S A* **112**, 3669-3673, doi:10.1073/pnas.1421784112 (2015).

89 Poznik, G. D. *et al.* Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. *Nat Genet* **48**, 593-599, doi:10.1038/ng.3559 (2016).

90 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 genetics* **10**, e1004353, doi:10.1371/journal.pgen.1004353 (2014).

91 A language and environment for statistical computing (R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org, 2008).

92 Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer Publishing Company, Incorporated, 2009).

93 Fadhlaoui-Zid, K. *et al.* Sousse: extreme genetic heterogeneity in North Africa. *J Hum Genet* **60**, 41-49, doi:10.1038/jhg.2014.99 (2015).

94 Arredi, B. *et al.* A predominantly neolithic origin for Y-chromosomal DNA variation in North Africa. *Am J Hum Genet* **75**, 338-345 (2004).

95 Trombetta, B., Cruciani, F., Sellitto, D. & Scozzari, R. A new topology of the human Y chromosome haplogroup E1b1 (E-P2) revealed through the use of newly characterized binary polymorphisms. *PLoS One* **6**, e16073, doi:10.1371/journal.pone.0016073 (2011).

96 Pereira, L. *et al.* Linking the sub-Saharan and West Eurasian gene pools: maternal and paternal heritage of the Tuareg nomads from the African Sahel. *Eur J Hum Genet* **18**, 915-923, doi:10.1038/ejhg.2010.21. Epub 2010 Mar 17.; eng; ID: 4297 (2010).

97 Ordóñez, A. C. *et al.* Genetic studies on the prehispanic population buried in Punta Azul cave (El Hierro, Canary Islands). *Journal of Archaeological Science* **78**, 20-28, doi:https://doi.org/10.1016/j.jas.2016.11.004 (2017).

98 Poznik, G. D. *et al.* Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. *Science* **341**, 562-565, doi:10.1126/science.1237619.; eng; ID: 4271 (2013).

99 Bekada, A. *et al.* Introducing the Algerian mitochondrial DNA and Y-chromosome profiles into the North African landscape. *PLoS ONE* **8**, e56775, doi:10.1371/journal.pone.0056775. Epub 2013 Feb 19.; eng; ID: 4125 (2013).

100 Batini, C. *et al.* Large-scale recent expansion of European patrilineages shown by population resequencing. *Nat Commun* **6**, 7152, doi:10.1038/ncomms8152 (2015).

101 Cann, H. M. *et al.* A human genome diversity cell line panel. *Science* **296**, 261-262 (2002).

102 Laurie, C. C. *et al.* Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet Epidemiol* **34**, 591-602, doi:10.1002/gepi.20516 (2010).

103 Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81**, 559-575, doi:10.1086/519795 (2007).

104 Gallego-Llorente, M. *et al.* The genetics of an early Neolithic pastoralist from the Zagros, Iran. **6**, 31326, doi:10.1038/srep31326 https://http://www.nature.com/articles/srep31326 - supplementary-information (2016).

105 Wang, C. *et al.* Ancestry estimation and control of population stratification for sequence-based association studies. *Nat Genet* **46**, 409-415, doi:10.1038/ng.2924 (2014).

106 Wang, C., Zhan, X., Liang, L., Abecasis, G. R. & Lin, X. Improved ancestry estimation for both genotyping and sequencing data using projection procrustes analysis and genotype imputation. *Am J Hum Genet* **96**, 926-937, doi:10.1016/j.ajhg.2015.04.018 (2015).

107 Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-909, doi:10.1038/ng1847 (2006).

108 Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome research* **19**, 1655-1664, doi:10.1101/gr.094052.109 [doi] (2009).

109 Henn, B. M. *et al.* Genomic Ancestry of North Africans Supports Back-to-Africa Migrations. *Plos Genetics* **8**, e1002397-e1002397, doi:10.1371/journal.pgen.1002397 (2012).

110 Maca-Meyer, N., Gonzalez, A. M., Larruga, J. M., Flores, C. & Cabrera, V. M. Major genomic mitochondrial lineages delineate early human expansions. *BMC Genet* **2**, 13 (2001).

111 Maca-Meyer, N. *et al.* Mitochondrial DNA transit between West Asia and North Africa inferred from U6 phylogeography. *BMC Genet* **4**, 15 (2003).

112 Olivieri, A. *et al.* The mtDNA legacy of the Levantine early Upper Palaeolithic in Africa. *Science* **314**, 1767-1770 (2006).

113 Gonzalez, A. M. *et al.* Mitochondrial lineage M1 traces an early human backflow to Africa. *BMC Genomics* **8**, 223 (2007).

114 Pereira, L. *et al.* Population expansion in the North African late Pleistocene signalled by mitochondrial DNA haplogroup U6. *BMC Evol Biol* **10**, 390, doi:10.1186/1471-2148-10-390.; eng; ID: 4231 (2010).

115 Prufer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43-49, doi:10.1038/nature12886 [doi] (2014).

116 Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710-722 (2010).

117 Ranciaro, A. *et al.* Genetic origins of lactase persistence and the spread of pastoralism in Africa. *Am J Hum Genet* **94**, 496-510, doi:10.1016/j.ajhg.2014.02.009 (2014).

118 Gallego Llorente, M. *et al.* Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent. *Science* **350**, 820-822, doi:10.1126/science.aad2879 (2015).