doi:10.1038/nature12736

TABLE OF CONTENTS

SI 1. The archaeological sites of Mal'ta and Afontova Gora-2

S1 1.1 Mal'ta

Mal'ta is a multi-component site located along the Belaya River, tributary of the Angara River in the Pre-Baikal region of southern Siberia, about 86 km northwest of Irkutsk (Figure 1a). It is most noted for its rich middle Upper Palaeolithic (MUP) archaeological materials, including the famous double-child burial and "Venus" figurines. The site was originally excavated during the early 20th century (1928-1958) by Gerasimov[1,2], and later during the 1990s by Medvedev and colleagues[3]. Original excavations were extensive, covering 1348 m^2, where Gerasimov reported a single cultural layer for the Upper Palaeolithic found 1.5 m below the surface[1].

Recent geoarchaeological investigations have shown site stratigraphy to be more complex than originally thought[3]. The profile is ~2.5 m deep with 10 lithostratigraphic units grouped into two sections. Ten cultural layers were found in the upper section. The basal section is not well preserved and not present uniformly across the site. The upper 2 m section consists of alternating alluvial and aeolian stratigraphic units. Neolithic and Final Palaeolithic artifacts were found in two cultural layers in lithostratigraphic units 10 and 9, respectively. Middle Upper Palaeolithic (MUP) artifacts were found in seven cultural layers. Stratum 8 contained four MUP layers (3 and 4/5/6), where cultural layers 4/5/6 were only identified as separate layers in the profile after excavation and therefore not separated during excavation. Cultural layers 3 and 4/5/6 provided most of the excavated MUP materials, indicating this is the same stratigraphic context from where the materials excavated by Gerasimov[1,2] came, including the Mal'ta skeletal remains. Additional Upper Palaeolithic cultural layers, 7-10, were found in lithostratigraphic units 7 through to 4, respectively. Cryoturbation affected the entire Mal'ta profile, and large ice wedges originating in stratum 8 were found to penetrate down through stratum 3[3]. Our focus here is on the MUP materials from cultural layers 3-6.

Medvedev and colleagues[3] reported several radiocarbon (^{14}C) dates on bones from their excavations. Twelve reported from stratum 8 are fairly consistent, mostly overlapping at 2σ and ranging in age from 21,700 ± 160 (OxA-6191) to 19,900 ± 800 (GIN-7705) ^{14}C BP. Despite the effects of cryoturbation, three additional bone dates from overlying stratum 9 (14,720 ± 190 [GIN-8476] ^{14}C BP) and underlying stratum 6 (43,100 ± 2400 [OxA-6189] ^{14}C BP) and stratum 5 (41,100 ± 1500 [GIN-7707] ^{14}C BP) neatly bracket the age range, 22,000-20,000 ^{14}C (26,000-23,500 cal) BP, of the MUP cultural layers (^{14}C dates were calibrated using Calib 6.1.1 software and the Intcal09 curve). Additionally, Richards and colleagues[4] presented a direct date of 19,880 ± 160 (OxA-7129) ^{14}C BP on one of the children from the double burial. Generally, these dates correspond well with the date obtained in this study (see Table SI 1).

Remarkably, Gerasimov found a huge assemblage of cultural materials in addition to the famous double-child burial[2]. Tens of thousands of artifacts were found. The lithic economy is based on production of tools made of small blades and flakes. Osseous (bone, antler, ivory) artifacts number more than 500, including projectile points, billets, needles, awls, undecorated and decorated pendants and beads, rectangular and disc-shaped plaques, several enigmatic polished and decorated pieces, zoomorphic figures, and at least 30 anthropomorphic "Venus" figurines were also found[3,5]. Faunal remains were extremely well preserved and include reindeer, Arctic fox, woolly rhinoceros, mammoth, bison, wolverine, red fox, horse, bighorn sheep, red deer, cave lion, brown bear, wolf, birds, large fish, and rodents. Dwelling features were large, circular-shaped, lined with large stone slabs and contained central hearth and storage-pits[6]. This spectacular array of cultural materials directly impacted the definition of MUP archaeology in Siberia for decades and is termed by many as the Mal'ta Culture or archaeological complex[3,7-12]. Due to the wide array and large number of artifacts coupled with the burial and dwelling features that reportedly came from a single cultural layer, Gerasimov concluded the site must have served as long-term residential base[1,2,6].

Most important to this study is the burial feature. Gerasimov reported an oval-shaped feature that originated in the MUP cultural layer and extended down into the lowest section of site deposits[13]. A stone slab reportedly covered the remains of what Gerasimov originally thought to be a single *Homo sapiens* child skeleton of 3-4 years in age, but closer look revealed a second child of about 1-2 years old among the remains[14]. The older child was found "wearing" a necklace of beads, several pendants and an ivory diadem. Other grave goods included a decorated plaque, bird-shaped pendant, ivory bracelet, stone tools and an ivory rod. Remains of the older child include much of the cranium, parts of the mandible and maxilla, several vertebrae and ribs, one humerus, the fragment of a lower limb bone, two phalanges, fragments of two femora, and fragments of two tibiae. The second child is represented by a second set of teeth[14]. The cranial fragments and teeth are currently kept at the Kunstkamera Museum of Anthropology, St. Petersburg, Russia. The other remains are permanently housed at Hermitage State Museum, St. Petersburg. Our humerus sample (MA-1) came from the older individual, housed at the Hermitage State Museum.

Christy Turner studied the Mal'ta teeth in the early 1980's and based on morphological characteristics concluded they were most closely related to Upper Palaeolithic Europeans such as those represented at the Sungir' site near Moscow[15] (Figure 1a). Based on slight shoveling of the incisors, however, Alekseev[16] concluded the Mal'ta children were likely of the "Mongoloid" type, or East Asian in origin. Our results certainly support Turner's interpretations, especially coupled with the presence of the mitochondrial DNA (mtDNA) haplogroup (hg) U in Mal'ta and other European Upper Palaeolithic peoples sampled for ancient DNA[17] (Figure 1a). Similar to Mal'ta, several MUP sites from the Don River valley on the Russian Plain to the Dordogne River valley in France have carved ivory "Venus" figurines (Figure 1a). Though it has no female figurine forms, the earlier Yana RHS site (ca. 28,000 ^{14}C BP) in western Beringia also contains an elaborate ivory-carved art[18] (Figure 1a).

S1 1.2 Afontova Gora

Afontova Gora-2 is located within the city limits of Krasnoiarsk, Krasnoiarsk Krai, Russia (Figure 1a). The site was first excavated from 1912-1914 by V. I. Gromov and subsequently by G. P. Sosnovskii, N. K. Auerbakh, and G. M. Mergart from 1919-1925[19-21]. Initial geological work there was undertaken by Gromov during these early 20th century expeditions. In the 1960s the site was revisited and monitored by archaeologists Z. A. Abramova and S. N. Astakhov[22]. Archaeological excavations covered an area of about 200 m^2.

The site is positioned on the western bank of the Enisei River on a 14-16 m high terrace-like feature mantled by alluvial and colluvial deposits[23]. The profile is approximately 12 m thick with 5 stratigraphic units. The lowest stratum consisted of alluvial cobbles and coarse sands, making up streambed deposits. Overlying stratum 4 (90-100 cm thick) consisted of sandy loams with alternating bands of clay and sand. Stratum 3 was a 350 cm-thick set of pale-yellow-to-gray sandy loams banded with fine silts near the top and evidence of carbonate development, probably reflecting soil formation. Stratum 2 was a 450 cm-thick set of pale-yellow-brown sandy loams again with evidence of carbonate development and few dark mottles. Within this stratum, four cultural layers, C_3, C_0, C_2, C_1, were found. Overlying C_0, C_2, and C_1 were observed dipping west and underlying C_3 was found to be concave in cross-section. Atop the profile was stratum 1, a light yellow-gray loess with snail shells and two cultural layers, B_2 and B_1. Of concern here is stratum 2 and its lowest cultural layer C_3.

Because all the work at Afontova Gora-2 was done before the advent of radiocarbon dating, no ^{14}C dates were obtained from the stratigraphic profile; however, Sosnovskii collected charcoal from the concave-shaped horizon of stratum 2, corresponding with cultural layer C_3 because he thought this horizon represented a dwelling feature. Askhakov asserted the "feature" resulted from landslide deformation and was not a cultural feature[23]. Due to evidence of soil development in this stratum, he argued it must date to an interstadial period of the late glacial. Graf sampled and reported dates obtained on three wood-charcoal pieces (Identified as probable *Salix* sp.) from Sosnovskii's collection of C_3[24,25]. These ages of 13,970±80 (AA-68663), 13,870±80 (AA-68664), and 12,280±80 (AA-68662) ^{14}C BP suggest a Bølling-Allerød (17,000-14,000 cal BP) age. The first two dates overlap at 2σ and correspond nicely with the date obtained in this study (see Table SI 1).

Archaeological materials from cultural layer C_3 are numerous. Artifacts from this layer number over 20,000 with more than 450 tools, 60 cores, and 19,500 debitage pieces. The lithic economy is based on production of tools made on blades and flakes, but also production of microblades to be inserted into slotted osseous points and knives. Osseous implements number at least 250 and include slotted and unslotted points, rods, wrenches or "*baton de commandements*," awls, needles, slotted knives, beads, pendants (some on teeth), and miscellaneous worked pieces of ivory. Faunal remains include reindeer, Arctic fox, hare, birds, mammoth, Argali sheep, red deer, wolverine, horse, bison, roe deer, Siberian mountain goat, Saiga antelope, wolf/dog, red fox, mollusks, and rodents (in order of abundance) and suggest a mixed open forest-steppe environment.

Osteological remains of two *Homo sapiens* individuals were also found in cultural layer C_3. These include the second upper premolar of a juvenile (11-15 years old) and fragments of a left radius, ulna, humerus, a phalanx, and part of the frontal of an adult[23,26,27]. Based on the little data available, Russian palaeoanthropologists have contended these individuals were East Asian in origin[26-28]. These remains are housed at the Hermitage State Museum, St. Petersburg, from where we sampled the humerus (AG-2).

References for SI 1

1. Gerasimov, M.M. *Mal'ta – paleoliticheskaiia stoianka.* IGOM, Irkutsk (1931)

2. Gerasimov, M.M. Paleoliticheskaiia stoianka Mal'ta: Raskopki 1956-1958 gg. *Sov. Etnografiia* **3**, 82-52 (1958)

3. Medvedev, G.N., Cauwe, G., Vorob'eva, D., Coupe, L., Claes, E., Lipnina, S., Modrie, S., Mukharamov, S., Osadchii, P., Pettitt, P., Rebrikov, E., Rogovskii, V., Sitlivyi, L., Sulerzhitskii, L.D., Khenzykhenova, F. *Paleoliticheskoye Mestonakhozhdenie Mal'ta*. ARCOM, Irkutsk (1996)

4. Richards, M.P., Pettitt, P.B., Stiner, M.C., Trinkaus, E. Stable isotope evidence for increasing dietary breadth in the European mid-Upper Paleolithic. *PNAS* **98**, 6528-6532 (2001)

5. Abramova, Z.A. *L'art Paléolithique d'Europe Orientale et de Sibérie*. Jerome Millon, Grenoble (1995)

6. Gerasimov, M.M. The Paleolithic site of Mal'ta: excavation of 1956-7. *The Archaeology and Geomorphology of Northern Asia: Selected Works*. University of Toronto Press, Toronto. 5-32 (1964)

7. Derevianko, A.P. *The Paleolithic of Siberia: New Discoveries and Interpretations*. University of Illinois Press, Urbana (1998)

8. Medvedev, G.I. Art from central Siberian Paleolithic sites. *The Paleolithic of Siberia: New Discoveries and Interpretations*. University of Illinois Press, Urbana. 132-137 (1998)

9. Medvedev, G.I. Upper Paleolithic Sites in South-Central Siberia. *The Paleolithic of Siberia: New Discoveries and Interpretations*. University of Illinois Press, Urbana. 122-132 (1998)

10. Okladnikov, A.P. Buret', Novaia Paleoliticheskaia Stoianka na Angare. *Sovetsakaia Arkheologiia* **5**:290-293 (1940)

11. Vasil'ev, S.A. The Late Paleolithic of the Yenisei: A New Outline. *Journal of World Prehistory* **6**, 337-383 (1992)

12. Vasil'ev, S.A. Faunal Exploitation, Subsistence Practices and Pleistocene Extinctions in Paleolithic Siberia. *Deinsea* **9**, 513-556 (2003)

13. Gerasimov, M.M. Raskopki plaeoliticheskoi stoianki v sele Mal'ta. *Izvestiia Gosdarstvennoi Akademii Istorii Material'noi Kul'tuy* **118**, 78-124 (1935)

14. Alekseev, V.P., Gokhman, I.I. Kostnye ostatki detskikh skeletov iz pogrebeniia na paleoliticheskoi stoianke Mal'ta. *Izvestiia SO AN SSSR, Seriia, filologii I filosofii* **16**, 54-57 (1987)

15. Turner, C.G. Rebenok verkhnepaleoliticheskoi stoianki Mal'ta (Sibir'). *Izvestiia Sibirskogo Otdeleniia Akademiia Nauk* SSSR **2**, 70-71 (1990)

16. Alekseev, V. The Physical Specificities of Paleolithic Hominids in Siberia. *The Paleolithic of Siberia: New Discoveries and Interpretations*. University of Illinois Press, Urbana. 329-335 (1998)

17. Fu, Q. *et al.* A Revised Timescale for Human Evolution Based on Ancient Mitochondrial Genomes. *Curr. Biol.* **23**, 553-559. (2013)

18. Pitulko, V.V., Pavlova, E.I.U. *Geoarkheologiia i Radiouglerodnaia Khronologiia Kamennogo Veka Severo-Vostochnoi Azii*. Nauka, St. Petersburg (2010)

19. Auerbakh, N.K., Sosnovskii, G.P. Ostatki Drevneishei Kul'tury Cheloveka v Sibiri. *Zhizn' Sibiri* **5**, 199-241 (1924)

20. Sosnovskii, G.P. Paleoliticheskie Stoianki Severnoi Azii. *TRUDY: II Mezhdunarodnoi Konferentsii Assotsiatsii po Izucheniiu Chetvertichnogo Perioda Evropy*, Vypusk V. Gosudarstvenoie Hauchno-Tekhnicheskoe Gorno-Geologo-Neftianoe Izdatel'stvo, Leningrad. 246-292 (1934)

21. Sosnovskii, G.P. Poslenie na Afontovoi Gore. *Izvestiia Gosudarstvennoi Akademii Istorii Material'noi Kul'tury* **118**, 152-218 (1935)

22. Abramova, Z.A., Astakhov, S.N., Vasil'ev, S.A., Ermalova, N.M., Lisitsyn, N.F. *Paleolit Eniseia*. Nauka, Leningrad. (1991)

23. Astakhov, S.N. *Paleolit Eniseia: Paleoliticheskie Stoianki Afontovoi Gore v G. Krasnoiarske*. RAN, St. Petersburg (1999)

24. Graf, K.E. Is it really that old? Dating the Siberian Upper Paleolithic site of Afontova Gora-2. *Current Research in the Pleistocene* **25**, 19-21 (2008)

25. Graf, K.E. "The Good, the bad, and the ugly": evaluating the radiocarbon chronology of the middle and late Upper Paleolithic in the Enisei River valley, south-central Siberia. *Journal of Archaeological Science* **36**, 694-707 (2009)

26. Alekseev, V.P., Gokhman, I.I. *Antropologiia Aziatskoi Chasti SSSR*. Izdatel'stvo Nauka, Moscow (1984)

27. Gerasimova, M.M., Astakhov, S.N., Velichko, A.A. *Paleoliticheskii Chelovek, Ego Material'naia Yul'tura I Prirodnaia Sreda Obitaniia*. Nester Istoria, St. Petersburg (2007)

28. Debets, G.F. Fragment lobnoi kosti cheloveka iz kul'turnogo sloia stoianki "Afontova Gora II" pod Krasnoiarskom. *Biull. Komissii po Izucheniiu Chetvertichnogo Perioda* **8**, 73-77 (1946)

SI 2. Radiocarbon dating

Accelerator Mass Spectrometry (AMS) ^{14}C chemistry followed standard protocol[1,2] with the following modifications. The two bone samples from Mal'ta (MA-1) and Afontova Gora-2 (AG-2) were physically cleaned by removing the outer 1 mm of cortex, followed by washing in acetone and methanol and drying under vacuum. Each of the bone samples was broken into approximately 3-4 mm fragments and decalcified in 4^{o} C, 0.5 *N* HCl over 3 days. After washing to neutrality in deionized (DI) water, the decalcified collagen was extracted with 0.1% KOH at 4^{o} C for 2 days, followed by washing to neutrality with DI water. The KOH-extracted, decalcified collagen's percent pseudomorph was recorded and freeze-dried to determine percent yield of collagen relative to modern bone. The decalcified, KOH-extracted collagen was heated at 90^{o} C in 0.02 *N* HCl to dissolve (gelatinize) the collagen. Heating continued until the collagen dissolved after 15-30 minutes. After filtering the gelatin solution through 0.45 µm Millex Durapore filters, the solution was freeze-dried. Gelatin was hydrolyzed for 22 hours at 110^{o} C in 6 *N* HCl. The hydrolyzate, which contained free amino acids, fulvic acids, and insoluble inorganic and organic detritus was passed through a 1 cm long X 5 mm diameter bed of XAD-2 resin in a solid phase extraction (SPE) column attached to a 0.45 µm Millex filter. The XAD column contained 100-200 µm diameter, research grade XAD-2 from Serva Biochemicals (Cat. No 42825). The bulk resin was initially wetted with acetone, followed by DI water and finally multiple washes with 1 *N* HCl made from distilled HCl. Individual SPE columns were packed with the XAD-2 as a slurry of resin and HCl. The columns were each equilibrated with 50 ml of distilled 6 *N* HCl and the washings discarded. The collagen hydrolyzate as approximately 1 ml of solution was pipetted onto the SPE XAD column and eluted into a glass tube. Following the initial sample aliquot, the column was washed with 10 ml of 6 *N* HCl that was added to the original eluate.

The XAD-purified collagen hydrolyzate was dried by passing UHP N_2 gas over the glass tube heated to 50^{o} C. The dried amino acids formed a viscous syrup. The dried hydrolyzate was diluted with DI water and 2-4 mg of amino acids were transferred to a 6 mm ID X 20 mm quartz tube and dried under vacuum. Approximately 40 mg of purified CuO wire and 5 mg of Ag^{o} powder were added to the quartz tube. CuO wire was combusted in crucibles at 900^{o} C and stored in Pyrex tubes that were combusted at 570^{o} C immediately before each use. Aesar 99.995%, 100 µm silver powder was used without additional purification. After evacuation to < 20 millitorr by vacuum pumping across a LN trap, the quartz tubes were sealed with a H_2/O_2 torch. The tubes were combusted at 850^{o} C for 2 hours and cooled from 850^{o} C to 250^{o} C at 30^{o} C per hour. Following purification of the combustion products to remove water and N_2, ~2 milligrams of carbon as CO_2 were converted into graphite by the Fe-H_2 method at the UC-Irvine AMS laboratory. Contemporary ^{14}C standards included National Bureau of Standards Oxalic Acid-I and ANU sucrose. Respective chemistry and combustion backgrounds were determined by using >70ka collagen isolated from the fossil *Eschrichtius robustus* (gray whale)[2,3] and Sigma Aldrich L-alanine (Catalog number A-7627). The graphitized samples and standards were analyzed at the University of California-Irvine WM Keck Carbon Cycle Accelerator Mass Spectrometry Laboratory (UCIAMS) under the direction of Dr. John Southon. The ^{14}C and calibrated dates, the latter obtained using OxCal 4.2[4] and the INTCAL09 dataset[5], are shown in Table SI 1.

References for SI 2

1. Stafford, T.W., Jr., Jull, A.J.T., Brendel, K., Duhamel, R. and Donahue, D. Study of bone radiocarbon dating accuracy at the University of Arizona NSF accelerator facility for radioisotope analysis. *Radiocarbon* **29**, 24-44 (1987)

2. Stafford, T.W., Jr., Brendel, K. and Duhamel, R. Radiocarbon, ^{13}C, and ^{15}N analysis of fossil bone: removal of humates with XAD-2 resin. *Geochimica Cosmochimica Acta* **52**, 2257-2267 (1988)

3. Stafford, T.W., Jr., Hare, P.E., Currie, L.A., Jull, A.J.T. and Donahue, D. Accelerator radiocarbon dating at the molecular level. *Journal of Archaeological Sciences* **18**: 35-72 (1991)

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

5. Reimer, P.J. *et al.* INTCAL09 and MARINE09 radiocarbon age calibration curves, 0-50,000 years cal BP. *Radiocarbon* **51**, 1111-1150 (2009)

Table SI 1: AMS ^{14}C measurements on human bone from Afontova Gora-2 and Mal'ta. AMS radiocarbon measurements on XAD-2 purified bone collagen. Percent collagen yields are milligrams (mg) of dry KOH-extracted collagen per milligram of clean, dry bone. Modern bones yield 20-22% protein by weight. Percent pseudomorph is the physical appearance of collagen relative to modern collagen that has a 100% pseudomorph. Radiocarbon measurements are in radiocarbon years (RC yr) before present (1950 AD) and are calibrated in calendar years before present (cal BP) with OxCal 4.2[4] using INTCAL09[5].

Site	Description	Bone amount (mg)	Collagen (mg)	Collagen yield (%)	Pseudo-morph (%)	^{14}C age ± SD (RC yr)	Cal BP (2 sigma)	AMS Lab No.
Afontova Gora-2 (AG-2)	SR-7866 Humerus	112.1	9.8	8.7	92	13,810±35	17,075-16,750	UCIAMS-79661
Mal'ta (MA-1)	SR-7912 Humerus, 1574-89	74.5	13.9	18.7	99-100	20,240±60	24,423-23,891	UCIAMS-79666

SI 3. Extractions, libraries and sequencing

SI 3.1 Ancient samples: MA-1 and AG-2

SI 3.1.1 Extractions

149 mg and 119 mg of bone powder from a humerus from MA-1 and a humerus from AG-2, respectively, were obtained using a Dremel drill. Both samples were extracted using a modified silica spin-column protocol[1,2,3]. Briefly, samples were incubated overnight at 55^o C in 1 ml of digestion buffer (1M urea, 0.45M EDTA, 0.1 mg/ml proteinase K). Following digestion, samples were concentrated through 30 kDa centrifugal filter units (Millipore, Billerica, MA) down to 250 µl and purified through MinElute columns (Qiagen MinElute PCR Purification Kit, Qiagen, Hilden, Germany), following manufacturer's protocol with the following modification. In the elution step, spin columns were incubated in 40 µl buffer EB at 37^o C for 10 minutes, spun down, and repeated once more. The eluates from both rounds of elution were pooled. These extracts were labeled MA-1_supernatant and AG-2_supernatant. Since both samples had undigested pellet after the overnight digestion, 1 ml of the digestion buffer described above was added to the pellet, except that the concentration of proteinase K was increased to 0.2 mg/ml. The digestion was left to incubate overnight at 55^o C after which time little to no pellet was found left-over. The supernatant obtained was concentrated and purified following the same protocol outlined above. These extracts were labeled MA-1_1stextraction and AG-2_1stextraction.

SI 3.1.2 Library preparation and sequencing

Illumina libraries were constructed for each of the four extracts from SI 3.1.1. A-tailed libraries were prepared with 16 µl of the extracts using the GS FLX Titanium Rapid Library Preparation Kit (454 Life Sciences, Roche, Branford, CO) and according to the manufacturer's protocol, with the following modifications. The extracts were not nebulized since ancient DNA is fragmented in nature. Ligation was performed with 0.4 µM of Index PE Adapter Oligo Mix (Illumina Multiplexing Sample Preparation Oligonucleotide Kit) at 25^o C for 25 minutes. The libraries were purified through MinElute columns and eluted in 25 µl of Qiagen Buffer EB, following a 10-minute incubation at 37^o C. The purified libraries were amplified as follows: 25 µl DNA library, 1X High Fidelity PCR buffer, 2 mM $MgSO_4$, 200 µM dNTPs each (Invitrogen, Carlsbad, CA), 200 nM Illumina Multiplexing PCR primer inPE1.0 (5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT), 4 nM Illumina Multiplexing PCR primer inPE2.0 (5'-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT), 200 nM Illumina Index PCR primer (5'-CAAGCAGAAGACGGCATACGAGATNNNNNNGTGACTGGAGTTC, where N's correspond to a 6 nucleotide index tag), 1 U of Platinum *Taq* DNA Polymerase (High Fidelity) (Invitrogen, Carlsbad, CA) and water to 50 µl. Cycling conditions were: initial denaturing at 94°C for 4 minutes, 8 cycles of: 94°C for 30 seconds, 60°C for 30 seconds, 68°C for 40 seconds, and a final extension at 72°C for 7 minutes. PCR products were purified through MinElute spin columns and eluted in 10 µl of Qiagen Buffer EB, following a 10-minute incubation at 37°C**.** A second round of PCR (two parallel reactions for each library) was set up as follows: 5 µl of purified product from first PCR round, 1X High Fidelity PCR buffer, 2 mM $MgSO_4$, 200 µM dNTPs each, 500 nM Illumina Multiplexing PCR primer 1.0, 10 nM Illumina Multiplexing PCR primer 2.0, 500 nM Illumina Index PCR primer, 1 U of Platinum *Taq* DNA Polymerase (High Fidelity), and water to 50 µl. Cycling conditions included an initial denaturing at 94°C for 4 minutes, 10 cycles of: 94°C for 30 seconds, 60°C for 30 seconds, 68°C for 40 seconds, and a final extension at 72°C for 7 minutes. Both PCR products originating from one library were purified through a MinElute spin column and eluted in 20 µl of Qiagen Buffer EB, following a 10-minute incubation at 37°C. All four libraries were run on Agilent 2100 Bioanalyzer High Sensitivity DNA chips. The library for AG-2_supernatant had an adapter dimer peak corresponding to 119 base pairs (bp), and hence was run on a 2% E-Gel SizeSelect Gel (Invitrogen, Carlsbad, CA). The band corresponding to the dimer peak was discarded, while the smear corresponding to the library was recovered with repeated refilling of the collection well with 25 µl aliquots of water. The collection was concentrated down to 20 µl using a SpeedVac (HetoVac VR-1, Birkerød, Denmark). AG-2_1stextraction did not yield a library product; hence one blunt-end library was constructed on 21.25 µl of the DNA extract using NEBNext DNA Sample Prep Master Mix Set 2 (New England Biolabs, E6070). The protocols outlined in the kit manual and Orlando *et al.* (2013)[4] were followed with the following modification. Reaction volumes were cut down from the manufacturer's protocol by a quarter in the end-repair step and by half in the ligation and fill-in steps. After the end-repair and ligation incubations, the reaction was purified through MinElute spin columns and eluted in 15 µl and 21 µl, respectively, after a 5-minute incubation at 37°C with Qiagen EB. Ligation reaction was performed for 25 minutes at 20° C using Illumina-specific adapters specified in Meyer & Kircher (2010)[5]. Fill-in reaction was performed for 20 minutes at 65° C. Library was amplified in two rounds as described above. The two PCR products originating from the library were pooled together, purified through a MinElute column and eluted in 20 µl EB, following a 10-minute incubation at 37°C. The size selected AG-2_supernatant and the blunt-end AG-2_1stextraction libraries were visualized on Agilent 2100 Bioanalyzer High Sensitivity DNA chip.

Equimolar pools of all libraries were sequenced to near-saturation over six lanes on the llumina HiSeq 2000 (100 cycles, single read mode) at the Danish National High-Throughput DNA Sequencing Centre. Sequencing was carried out a period of several months in 2012-2013.

SI 3.2 Modern Individuals: Tajik, Avar, Mari, Indian

SI 3.2.1 Sample background and collection

Tajik, Avar and Mari

Blood sample was collected from the Tajik individual in 2006 from west-central Pamir in the Gorno Badakhshan region of Tajikistan. This region is populated by several tribal groups such as the Shugnans, Rushans, Vanchs, Rins, Gorans, Wakhans and Bartangs. Many researchers consider them ethnic rather then sub-ethnic groups, as they all speak different, though related, languages. The individual's father and his paternal grandparents were ethnically Rushan from Rushan village, while his mother and maternal grandparents were ethnically Shugnan from Porshnev village. Blood sample from the Avar individual was collected in 2012. He was born in the mono-ethnic Avar settlement of Gergebil in Dagestan. His parents and grandparents were from the same settlement. Blood sample from the Mari individual was collected in 2009 in the Mishkinsky district of Bashkortostan, where Maris comprise up to 71.5% of the population according to census (2010). He was born in Novoakbulatovo where Maris are also the main ethnic group (79% according to the census in 2002).

Indian

Saliva sample from the researcher who performed extractions and library preparations of the ancient samples (MR) was collected in 2013 in Copenhagen, Denmark using an Oragen Dx collection kit (DNA Genotek Inc., Kanata, Canada). MR and her parents and both sets of grandparents are originally from South India (Tamil Nadu).

SI 3.2.2 Extractions

Tajik, Avar and Mari

DNA extraction from the Tajik sample was performed in Moscow, Russia and those from the Avar and Mari samples were performed in Ufa, Russia. Total genomic DNA was extracted from peripheral blood leucocytes. After incubation with proteinase K, a mixture of phenol and chloroform was used for DNA deproteinisation, followed by ethanol precipitation[6]. The final concentrations of the DNA extracts, measured using the Qubit 2.0 Fluorometer (Invitrogen, Carlsbad, CA), were as follows: 92 ng/µl (Tajik), 900 ng/µl (Avar) and 600 ng/µl (Mari).

Indian

DNA extraction from this sample was performed in Copenhagen, Denmark. The sample was incubated for two hours at 50° C. Total genomic DNA was extracted using a prepIT®•L2P extraction kit (DNAgenotek, Kenata, Canada). DNA extraction was done following the manufacturer's guidelines with the following modifications: extraction was performed on 1 ml of saliva and, DNA was eluted in 100 µl of water instead of TE buffer. The final concentration of the DNA extract, measured using the Qubit 2.0 Fluorometer, was 240 ng/µl.

SI 3.2.3 Library preparation and sequencing

2.5 µg from each of the four extracts were diluted in water to a final concentration of 10 ng/µl and fragmented using a Bioruptor (NGS, Diagenode, cat # UCD600), with 3 cycles of 15 seconds on and 90 seconds off for the Indian sample, and, 3 cycles of 30 seconds on and 90 seconds off for the Tajik, Avar and Mari samples. Fragments ranging between 300-700 bp were generated, as visualized on a 2% agarose gel. The fragmented extracts were concentrated down to 45 µl using a SpeedVac. Two blunt-end libraries were constructed for each of the four DNA extracts using the NEBNext DNA Sample Prep Master Mix Set 2 (New England Biolabs, E6070), and following the protocols outlined in the kit manual and Orlando *et al.* (2013)[4], with the modifications noted for the ancient library as well as the following. Ligation reactions were performed for 15 minutes at 25° C. Libraries were amplified as follows: two reactions per library with 12.5 µl input DNA (entire library volume was 25 µl), 1X Phusion Master Mix (M0531, New England Biolabs), 500 nM each of Illumina inPE1.0 primer and custom-made index primer (5' - CAAGCAGAAGACGGCATACGAGATNNNNNNGTGACTGGAGTTCAGACGT

GTGCTCTTCCG, where N's correspond to a 6 nucleotide index tag), and water to 50 µl. Both PCR reactions for each of the libraries were indexed using the same index primer. Cycling conditions included an initial denaturing at 98° C for 30 seconds; 10 cycles of: 98° C for 10 seconds, 65° C for 30 seconds, 72° C for 2 minutes; and a final extension at 72° C for 5 minutes. The two PCR products originating from the same library were pooled together and were purified through a Qiagen Qiaquick column and eluted in 20 µl of Qiagen EB, following a 5-minute incubation at 37°C. All amplified libraries were run on 2% agarose gels and size selection was performed between approximately 150 bp-400 bp, using E.Z.N.A. Gel Purification Kit (Omega Bio-Tek) with the following modifications. Gel solubilization was performed at 40° C on a heating block, and, final elution was performed in 30 µl after incubation at 37° C for 10 minutes. Size selected libraries were visualized on Agilent 2100 Bioanalyzer High Sensitivity DNA chip.

An equimolar pool of all libraries was sequenced over one test lane on the Illumina MiSeq (100 cycles, paired end mode) and subsequently over 6 and 2 lanes on the llumina HiSeq 2000 and the llumina HiSeq 2500 in rapid run mode (100 cycles, paired end mode), respectively, at the Danish National High-Throughput DNA Sequencing Centre. Sequencing was carried out a period of several months in 2012-2013.

References for SI 3

1. Yang, D.Y., Eng, B., Waye, J.S., Dudar, J.C., Sanders, S.R. Technical Note: Improved DNA Extraction from Ancient Bones Using Silica-Based Spin Columns. *Am. J. Phys. Anthropol.* **105**, 539-543 (1998)

2. Svensson, E.M. *et al*. Tracing genetic change over time using nuclear SNPs in ancient and modern cattle. *Anim. Genet.* **38**, 378-383 (2007)

3. Malmström, H. *et al*. Ancient DNA Reveals Lack of Continuity between Neolithic Hunter-Gatherers and Contemporary Scandinavians. *Curr. Biol.* **19**, 1758-1762 (2009)

4. Orlando, L. *et al.* Recalibrating *Equus* evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* **499**, 74-78 (2013)

5. Meyer, M., Kircher M. Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. *Cold Spring Harbor Protocols* 6 (June): pdb.prot5448. doi:10.1101/pdb.prot5448 (2010)

6. Powell, R., Gannon, F. Purification of DNA by phenol extraction and ethanol precipitation. Oxford Practical Approach Series. Oxford University Press, UK. http://fds.oup.com/www.oup.co.uk/pdf/pas/9v1-7-3.pdf (2002)

SI 4. Processing and mapping of raw sequence data from ancient and modern individuals

SI 4.1 Ancient samples: MA-1 and AG-2

Basecalling was performed using the Illumina software CASAVA 1.8.2. Sequences from various runs originating from the same library amplification were merged, with the requirement of a 100% match to the 6 nucleotide index used during library preparation. Adapter sequences were trimmed and, filtered for N's and reads shorter than 25 bases using AdapterRemoval-1.2[1]. Trimmed reads were mapped to the human reference genome builds hg18 and 37.1 using bwa-0.5.9[2], with seed length disabled to improve mapping efficiency in ancient DNA datasets (-l 1000)[3]. The trimmed reads were also mapped separately to the revised Cambridge Reference Sequence (rCRS, NC_012920.1) using the same parameters as above for the nuclear genome, except that a minimum mapping quality of 25 was required. The nuclear and mtDNA alignments were sorted using samtools[4] and, filtered for PCR duplicates using Picard MarkDuplicates-1.88 (http://picard.sourceforge.net) and for paralogs using the X1 tag (not equal to 0) as defined by BWA. Read depth and coverage were determined using pysam (http://code.google.com/p/pysam/) and BEDtools[5]. Mapping statistics are shown in Table SI 2. Reads are available for download through NCBI SRA accession number SRP029640, and, reads and alignments are available at http://www.cbs.dtu.dk/suppl/malta.

SI 4.2 Sixteen present-day humans, Denisova, Tianyuan

SI 4.2.1 Sample overview

Four individuals from the Volga-Ural region, the Caucasus, and, South and Central Asia (Mari, Avar, Indian and Tajik ancestry, respectively) were sequenced for this study (SI 3). The data for 11 high coverage modern humans and the Denisova genome was obtained from Meyer *et al*. (2012)[6], for one low-coverage Cambodian individual from Reich *et al*. (2010)[7] and, for the ancient Tianyuan individual from Fu et *al*. (2013)[8].

SI 4.2.2 Read processing of the Mari, Avar, Tajik and Indian individuals

Basecalling was performed using the Illumina software CASAVA 1.8.2. The sequences were split requiring an exact match to the 6 nucleotide index used during the library preparation. The raw reads from the four individuals were trimmed using AdapterRemoval-1.1[1] for adapter sequences, and, leading and trailing Ns to a minimum length of 25 bases. The trimmed reads were mapped to the human reference genome build 37.1 using bwa-0.6.2[2] and the alignments were subsequently filtered for a mapping quality of at least 30. Hereafter, the alignments were sorted, merged to the library level, filtered for PCR duplicates using Picard MarkDuplicates-1.56 (http://picard.sourceforge.net), merged to sample/individual level, realigned using GATK[9] and updated for the md-tags using samtools[4]. Read depth and coverage were determined using pysam (http://code.google.com/p/pysam/) and BEDtools[5]. The Avar, Indian, Mari and Tajik individuals were sequenced to 12.8X, 15.9X, 12.0X and

16.4X, respectively. We note that the both libraries from the Mari individual yielded significantly lower read numbers mapping to the human reference genome, most likely originating from low-quality extract. Statistics are shown in Tables SI 3 and SI 4. Reads and alignments are available for demographic research under data access agreement with E.W. (ewillerslev@snm.ku.dk).

SI 4.2.3 Read processing of the present-day human genomes, Denisova and Tianyuan genomes

High coverage data from the 11 modern individuals (SRX103808)[6] were deplexed allowing for one mismatch in the indexes. These, and the reads from the low coverage Cambodian individual (ERR019687)[7], were mapped and processed identical to the Mari, Avar, Tajik and Indian individuals, except that –q 15 was used to trim low quality ends of the reads.

The data from chromosome 21 of the ancient 40,000 year old Tianyuan individual (ERR206777)[8] was also mapped and processed similarly as all the above modern individuals, except that it was aligned with the seed region in bwa disabled (-l 1024) to allow for better sensitivity to ancient data[3].

For the Denisova genome[6] (DenisovaPinky), the alignments to hg19/GRCh37 were downloaded from http://cdna.eva.mpg.de/denisova/alignments/. To ensure consistency with our data, all single end read libraries were extracted from the alignments using Picard-1.87 SamToFastq. Hereafter, the reads were mapped and processed as described for the modern individuals, except that the seed region in bwa was disabled. Statistics on read mapping are shown in Tables SI 3 and SI 4.

Alignments for the above datasets are available at:
http://www.cbs.dtu.dk/suppl/malta.

SI 4.2.4 Genotyping

All modern samples (except the low coverage Cambodian individual) and the Denisova individual were genotyped using samtools-0.1.18 mpileup and bcftools[4]. Each sample was genotyped individually and filtered/masked according to the following criteria to achieve a high confidence single nucleotide polymorphism (SNP) set[10]:

1. Strand and distance bias with a threshold of 0.0001 and SNPs within 5 nts of alignment gap

2. Phred-scaled genotype posterior probability quality > 20

3. Read depths between 10 and 100 for autosomes, 5-50 for Y and X chromosomes (X except in females: 10-100) and 10+ for mtDNA (MT)

4. Not within an annotated repeat

5. Allele fraction of minor allele (allelic balance) greater than 0.2 for heterozygotes

6. Not within 5 nts of another variant call

7. Heterozygote calls on X, Y and MT for males and MT for females were masked

To produce the final call set only bi-allelic sites were included and the individual calls were merged to a final set using GATK CombineVariants-2.5-2[9] and are available at: http://www.cbs.dtu.dk/suppl/malta. The higher rates of missing calls in the Avar, Indian, Mari and Tajik samples are due to the lower sequencing depth of these individuals compared to the minimum threshold of 10X for the high confidence calls. Statistics on the called sites are shown in Table SI 5.

SI 4.3 Biological sex of MA-1 and AG-2

Information on the biological sex of ancient individuals from genetics can provide information on burial structure and social structure in prehistoric groups, and is also necessary before contamination can be estimated using X or Y chromosomal sequences (see SI 5). To investigate biological sex using the low-coverage sequencing data from MA-1 and AG-2, an approach was used wherein the fraction of high-quality Y chromosome alignments is divided by the total number of alignments to the sex chromosomes to obtain the statistic R_y, which is then compared to thresholds obtained from a reference panel of known sex[11]. The datasets used for the analysis were the larger of the 2 MA-1 libraries (MA-1_1st extraction), and a merged AG-2 sequence set comprising of both AG-2_supernatant and AG-2_1st extraction libraries in order to increase the total read number for this sample. The analysis was restricted to sequences with mapping quality 30 or higher. R_y values of 0.0934±0.0003 (±1 SE) and 0.0928±0.0007 for MA-1 and AG-2 were obtained, respectively. For MA-1, this ratio was obtained from 100,925 sequences that aligned to the Y chromosome out of a total of 1,080,281 alignments to both X and Y chromosomes. These R_y values are well over the conservative threshold of 0.075 for males[11], allowing confident assignment as males in both cases. The R_y statistics for the two ancient Siberian individuals are shown in Figure SI 1, together with 30 modern and ancient reference individuals for comparison[7,10,12-17].

Since there is evidence of contamination in Afontova-Gora-2 (SI 5), the analysis was repeated on both the individuals using only sequences with a C → T mismatch in the first 5 bases, where the T base in the sequence read had a phred-scaled base quality of at least 30. Both assignments were found to be robust to this restriction ('MA-1 PMD' and 'AfontovaGora PMD', Figure SI 1).

References for SI 4

1. Lindgreen, S. AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC Res Notes* **5**, 337 (2012)

2. Li, H., Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009)

3. Schubert, M. *et al.* Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics* **13**, 178 (2012)

4. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009)

5. Quinlan, A.R., Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010)

6. Meyer, M. *et al.* A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science* **338**, 222-226 (2012)

7. Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053-1060 (2010)

8. Fu, Q. *et al.* DNA analysis of an early modern human from Tianyuan Cave, China. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 2223-2227 (2013)

9. DePristo, M.A. *et al*. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491-498 (2011)

10. Rasmussen, M. *et al.* An Aboriginal Australian Genome Reveals Separate Human Dispersals in Asia. *Science* **334**, 94-98 (2011)

11. Skoglund P., Storå J., Götherström A., Jakobsson M. Accurate sex identification in ancient human remains using DNA shotgun sequencing. *Journal of Archaeological Science* **40**, 4477-4482 (2013)

12. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012)

13. Green, R.E. *et al*. A Draft Sequence of the Neandertal Genome. *Science* **328**, 710-722 (2010)

14. Sánchez-Quinto, F. *et al*. Genomic Affinities of Two 7,000-Year-Old Iberian Hunter-Gatherers. *Curr. Biol.* 22, 1494-1499 (2012)

15. Skoglund, P. *et al*. Origins and Genetic Legacy of Neolithic Farmers and Hunter-Gatherers in Europe. *Science* **336**, 466-469 (2012)

16. Keller, A. *et al*. New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nat. Commun*. **3**, doi: 10.1038/ncomms1701 (2012)

17. Rasmussen, M. *et al*. Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463**, 757-762 (2010)

Table SI 2 Mapping statistics for libraries from the two ancient samples. Reads were mapped against hg18 and 37.1 human genome reference builds and rCRS/NC_012920.1 for the mtDNA. For the mtDNA, reads were mapped with a minimum mapping quality of 25. For the nuclear genome, reads were mapped without quality filtering; however the BAMs were filtered for downstream analyses for a minimum mapping quality (q) of 30 and a minimum base quality (Q) of 20 and the numbers reported below reflect this filtering. The final numbers of reads in the mtDNA and nuclear BAMs account for filtering of clones (MarkDuplicates) and paralogs (BWA X1 tag). The percentage of reads mapping to the nuclear genome (% Nuclear mapped) is calculated as the final number of reads in the BAM as a percentage of the total number of trimmed reads.

Library	Genome reference build	Total reads	Total trimmed	Avg length (bp)	Mapped mtDNA (q25)	Final in BAM mtDNA	mtDNA depth	mtDNA covered > 1X (%)	Mapped nuclear	Final in BAM nuclear (q30,Q20)	Nuclear depth	Nuclear covered > 1X (%)	% Nuclear mapped
AG-2 _supernatant	hg18	93817640	93325393	87	2143	1610	8.4	99.7	4873864	3336889	0.1	8.8	3.6
	37.1								4877604	3338420	0.1	8.8	3.6
AG-2 _1stextraction	hg18	300206999	299960544	80	5206	1620	7.8	98.4	8352799	2388103	0.1	5.9	0.8
	37.1								8361873	2390400	0.1	5.9	0.8
MA-1 _supernatant	hg18	204287004	200098019	85	5923	4185	20.9	100	12548934	8682483	0.2	20.6	4.3
	37.1								12565893	8685692	0.2	20.5	4.3
MA-1 _1stextraction	hg18	238502007	225519620	84	44045	15406	76.6	100	82117087	38459258	1.0	59.2	17.1
	37.1								82187013	38473754	1.0	59.0	17.1

Table SI 3 Mapping statistics for libraries from four modern individuals sequenced in this study. Summary per library of reads and mapping characteristics of Avar, Indian, Mari and Tajik.

Library	Sample	Raw reads	Mapped (q30)	% endogenous	Reads in final BAM	% duplicates	% in final BAM
Avar_lib1	Avar	156,229,790	117,585,711	75.3	116,729,684	0.73	74.7
Avar_lib2	Avar	444,509,806	316,960,804	71.3	311,609,712	1.69	70.1
Indian_lib1	Indian	200,492,954	151,057,025	75.3	150,291,567	0.51	75.0
Indian_lib2	Indian	574,906,508	429,399,687	74.7	425,959,330	0.80	74.1
Mari_lib1	Mari	695,905,062	253,538,476	36.4	245,915,581	3.01	35.3
Mari_lib2	Mari	448,475,758	167,098,570	37.3	163,766,668	1.99	36.5
Tajik_lib1	Tajik	370,121,948	293,021,804	79.2	291,135,409	0.64	78.7
Tajik_lib2	Tajik	321,918,716	252,353,801	78.4	250,894,236	0.58	77.9

Table SI 4 Mapping statistics for 16 present-day, Denisova and Tianyuan individuals. Summary of reads, average depth and percentage of the genome covered by at least one read. (*: Only chromosome 21)

Sample	Population	Reference	Mapped Reads	Average Depth (X)	Covered > 1X
Avar	Avar	This study	428,339,396	12.8	83.1
Indian	Indian	This study	576,250,898	15.9	89.1
Mari	Mari	This study	409,682,249	12.0	87.3
Tajik	Tajik	This study	542,029,645	16.4	88.3
DNK02	Dinka	*6*	773,846,667	24.3	90.0
HGDP00456	Mbuti	*6*	646,936,869	20.3	90.0
HGDP00521	French	*6*	717,048,651	22.6	90.0
HGDP00542	Papuan	*6*	687,911,064	21.6	90.0
HGDP00665	Sardinian	*6*	631,826,599	19.9	90.0
HGDP00778	Han	*6*	707,741,116	22.3	90.0
HGDP00927	Yoruba	*6*	849,278,661	26.7	90.1
HGDP00998	Karitiana	*6*	678,487,589	21.3	90.0
HGDP01029	San	*6*	857,598,755	26.9	90.1
HGDP01284	Mandenka	*6*	653,599,826	20.6	90.0
HGDP01307	Dai	*6*	757,727,346	23.8	90.1
HGDP00711	Cambodian	*7*	45,543,321	1.5	62.2
Tianyuan	Tianyuan	*8*	670,897	1.0*	43.8*
Denisova	Denisova	*6*	1,113,867,270	24.3	88.5

Table SI 5 Sample-wise summary of the bi-allelic SNP high confidence call set created for the analysis. Missing: missing and masked sites, HomRef: Homozygote reference allele calls, Het: Heterozygote alternative allele calls, HomAlt: Homozygote alternative allele calls, Accessible sites: total number of high confidence sites useable for analysis.

Sample	Missing	HomRef	Het	HomAlt	Accessible sites
Avar	760,808,278	622,402,501	449,160	215,023	623,066,684
Indian	521,503,586	861,479,377	616,395	275,604	862,371,376
Mari	1,148,377,653	235,221,532	189,930	85,847	235,497,309
Tajik	443,001,981	939,908,587	657,660	306,734	940,872,981
DNK02	25,706,656	1,356,543,046	1,148,850	476,410	1,358,168,306
HGDP00456	48,861,222	1,333,291,890	1,160,557	561,293	1,335,013,740
HGDP00521	30,380,858	1,352,176,995	883,649	433,460	1,353,494,104
HGDP00542	39,368,054	1,343,227,988	715,125	563,795	1,344,506,908
HGDP00665	47,594,690	1,334,984,620	864,356	431,296	1,336,280,272
HGDP00778	33,342,225	1,349,206,107	844,203	482,427	1,350,532,737
HGDP00927	20,965,814	1,361,247,984	1,172,877	488,287	1,362,909,148
HGDP00998	38,756,923	1,343,917,891	636,539	563,609	1,345,118,039
HGDP01029	19,481,135	1,362,569,810	1,235,964	588,053	1,364,393,827
HGDP01284	43,424,073	1,338,813,400	1,164,050	473,439	1,340,450,889
HGDP01307	32,040,301	1,350,515,888	843,417	475,356	1,351,834,661
Denisova	37,286,072	1,344,886,396	250,829	1,451,665	1,346,588,890

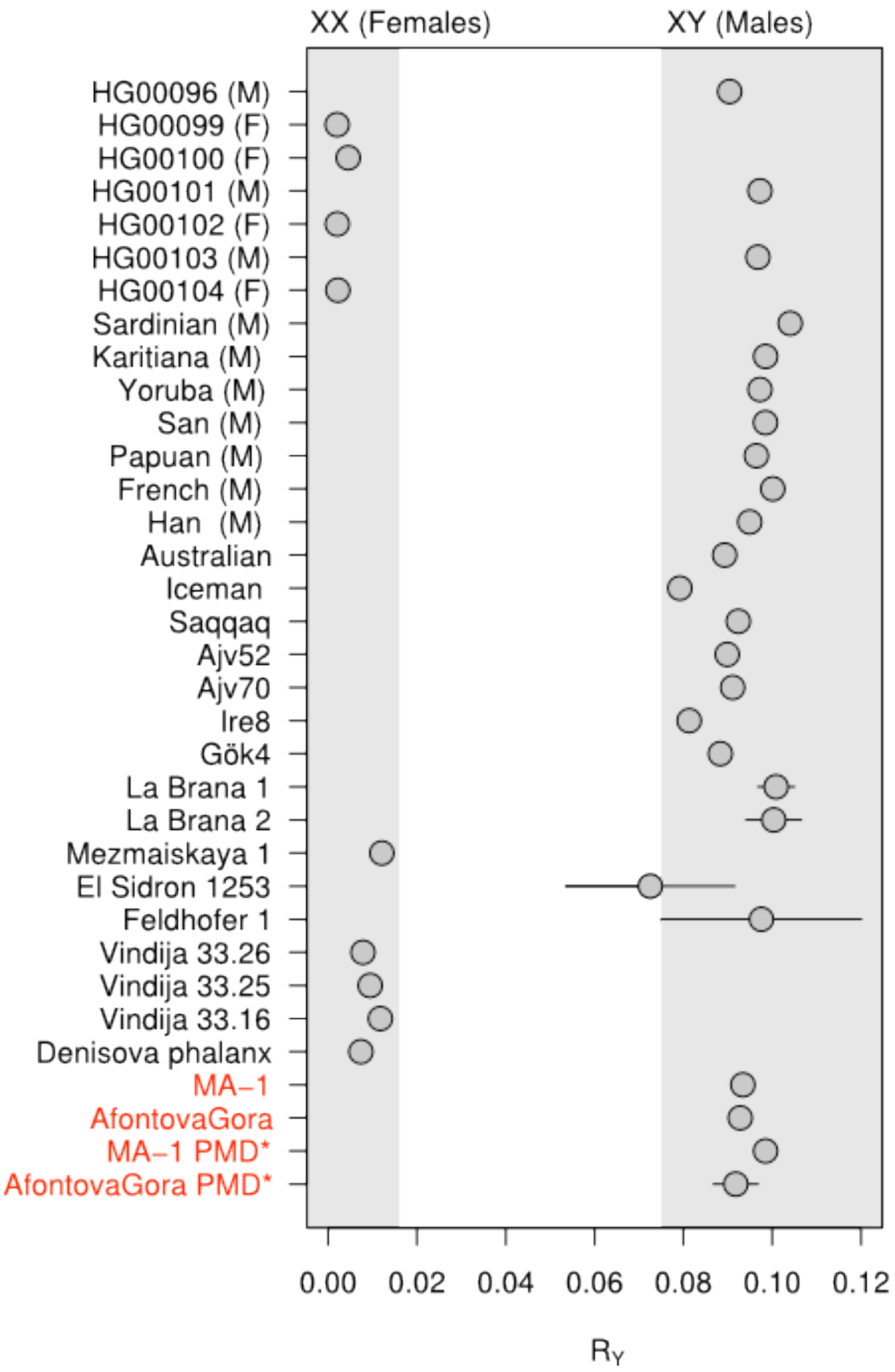


Figure SI 1 Sex assignment of MA-1 and AG-2 (AfontovaGora) compared to a reference panel of modern and ancient individuals[8,11,13-18]. MA-1 and AfontovaGora represent the complete datasets as defined in the text, while MA-1 PMD and AfontovaGora PMD consist of sequence reads with evidence of *post-mortem* damage.

SI 5. DNA contamination estimates

SI 5.1 Mitochondrial DNA contamination estimates

SI 5.1.1 Description of the method

Due to the large copy number of mitochondria in the cell, multifold higher coverage is usually observed for the mtDNA genome compared to nuclear chromosomes in ancient DNA shotgun data sets. Together with the fact that mtDNA is haploid, this makes it suitable for detecting contamination by investigating whether more than a single allele is present at each position[1]. However, contamination at some loci may go undetected if the contaminant and the ancient individual carry the same allele. To circumvent this, we first identified consensus calls in the ancient mtDNA that are near-private to the ancient individual (at an allele frequency of less than 1% in a sample of 311 modern human mtDNA genomes)[1]. Subsequently, we assume that any contaminating DNA has the alternative major allele at these loci, and thus estimates of contamination can be obtained by quantifying the presence of such alternative alleles at the diagnostic positions[1]. Note that this method assumes the absence of heteroplasmic sites. If such sites were present they would lead to an overestimation of the contamination estimates.

The presence of near-private consensus alleles and potential contaminating reads at these positions were counted, and a 95% confidence interval was obtained assuming that the allele observed in each read is a random outcome of drawing one of two alleles (endogenous and contaminant).

SI 5.1.2 Data filtering

Contamination estimates were made for both ancient libraries originating from MA-1 and AG-2. Positions with a depth of less than 10X were excluded, as were positions where the consensus allele was either C or G in a transition polymorphism since these are sensitive to *post-mortem* nucleotide misincorporations. A base quality of 30 was required.

SI 5.1.3 Results

Results are shown in Table SI 6. Two diagnostic sites were identified for the MA-1_1stextraction library where 92 of 93 reads overlapping these sites support the consensus, yielding a point estimate of contamination of 1.1% (95% CI: 0.0-3.2%). MA-1_supernatant yielded a contamination estimate of 10% (95% CI: 0.0-23.1%). For the AG-2 libraries, only AG-2_supernatant had coverage on informative positions and provided a contamination estimate of 40% (95% CI: 9.6-70.4%). The analysis was repeated on a merged dataset with combined reads from AG-2_supernatant and AG-2_1stextraction libraries (labeled AG-2_merged). Only 28 of 43 reads were found to support the consensus allele, yielding a point estimate of contamination of 34.9% (95% CI: 20.7-49.1%). As a comparison, sequence reads from the Tyrolean Iceman[2] were remapped using BWA 0.5.9[3], and 93 of 96 reads were found to support the consensus (3.1% contamination; 95% CI: 0-6.6%). Apart from the MA-1_1stextraction library, the mtDNA contamination estimates show that the other three libraries contain significantly higher contamination.

SI 5.2 X chromosome-based DNA contamination estimates

SI 5.2.1 Description of the method

Contamination estimates were obtained using data from the X chromosome of MA-1 and AG-2. Since both were determined to be males (SI 4.3), the X chromosome is haploid and hence any discordance in observed bases in a single site is either due to sequencing errors or contamination. For this reason discordance rates in the X chromosome contain information about contamination, which is what is exploited here. This was done by using methods published previously[4], both to test for contamination and to estimate the amount. These methods are based on a fixed set of SNPs known to be polymorphic in Europeans. It was assumed that regardless of population, the probability of a site being polymorphic is higher for the set of known polymorphic sites compared to their adjacent sites. It was also assumed that the error rate for the set of known polymorphic sites is the same as the adjacent sites. Under these assumptions, the base discordance rate for the known polymorphic sites should be the same as for the adjacent sites if there is no contamination. In contrast, contamination from any human source will lead to a higher discordance rate for the known polymorphic sites than the adjacent sites (because contamination will only lead to discordance in sites that are polymorphic). Hence, by comparing discordance rate in known polymorphic sites to the discordance rates in their adjacent sites, contamination estimates can be determined.

Two approaches were used: “test 1” in which all reads are used and “test 2” in which only a single sampled read is used. Test 2 is less powerful but does not assume that the errors are independent between reads and sites. Details and validation of the method are presented elsewhere[4].

SI 5.2.2 Data

The set of known polymorphic sites were identified using 60 unrelated CEPH individuals from the HapMap phase II release 27 data[5]. This set was pruned such that no polymorphic sites were less than 10 bases apart. Based on these 60 individuals, we also estimated the allele frequencies in Europeans.

For the MA-1_supernatant, MA-1_1stextraction and the AG-2_merged (there was not enough data in the two individual AG-2 libraries for this analysis) datasets, the following filtering was performed:

- The X chromosome was trimmed to remove the regions that are homologous with the Y chromosome (first and last 5Mb).
- The sites were then filtered based on mappability (100mer), so that no region will map to another region of the genome with an identity above 98%
- Reads with a mapping quality score of less than 30 and bases with a base quality score less than 20 were removed.
- Sites with a read depth of less than 3 or above 40 were removed for MA-1_1stextraction.
- Sites with a read depth of less than 2 or above 40 were removed for MA-1_supernatant and AG-2_merged.

SI 5.2.3 Results

The results for the two MA-1 libraries are shown in Tables SI 7 and SI 8. Autosomal contamination in MA-1_supernantant is estimated to be 22.0% or 22.6% while the contamination in MA-1_1st extraction is estimated to be 1.6% or 2.0% (depending on the test). The results for AG-2 are shown in Tables SI 9 and SI 10. Autosomal contamination in AG-2 is significantly higher and estimated to be almost 30%. Note that due to the low depth and the high contamination rate we cannot accurately determine the contaminant read for both samples.

SI 5.3 Conclusions: Libraries for downstream analyses

Based on both mitochondrial and autosomal contamination estimates, **MA-1_1stextraction dataset** will be used hereafter for all further analyses and will be referred to as MA-1, since it is at a higher depth (1X) and provides low mtDNA and autosomal contamination estimates. Such a sequential manner of DNA extraction from bone, leading to an enrichment in the endogenous content (as in MA-2_1stextraction library) after an initial 'washing out' of surface contaminants (MA-1_supernatant library), was also observed previously and has been attributed to undigested pellets being a rich source of trapped, endogenous DNA molecules[6,7]. Since the two AG-2 libraries are highly contaminated and at a much lower depths than MA-1, the use of **AG-2_merged** will be restricted to estimating library error rate (SI 6), and Principal component analysis using only damaged sequences in order to circumvent the severe contamination in these two libraries (SI 15).

References for SI 5

1. Krause, J. *et al*. A Complete mtDNA Genome of an Early Modern Human from Kostenki, Russia. *Curr. Biol.* **20**, 231-236 (2010)

2. Keller, A. *et al*. New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nat. Commun*. **3**, doi: 10.1038/ncomms1701

3. Li, H., Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009)

4. Rasmussen, M. *et al.* An Aboriginal Australian Genome Reveals Separate Human Dispersals in Asia. *Science* **334**, 94-98 (2011)

5. Frazer, K.A. *et al.* A Second Generation Human Haplotype Map of over 3.1 Million SNPs. *Nature* **449**, 851–861 (2007)

6. Orlando, L. *et al*. True single-molecule DNA sequencing of a Pleistocene horse bone. *Genome Res.* **21**, 1705-1719 (2011)

7. Ginolhac, A. *et al*. Improving the performance of true single molecule sequencing for ancient DNA. *BMC Genomics* **13**, 177 (2012)

Table SI 6 Mitochondrial contamination estimates for MA-1, AG-2 and the Tyrolean Iceman. mtDNA contamination estimates were obtained for all four sequenced ancient libraries (MA-1_supernatant, MA-1_1st extraction, AG-2_supernatant, AG-2_1st extraction) as well as a merged dataset consisting of AG-2_supernatant and AG-2_1st extraction (AG-2_merged), and, the Tyrolean Iceman.

Sample	Average depth	Number of informative sites	Positions of informative sites (rCRS)	Observed majority alleles	Observed minority alleles	Fraction potential contamination (%)	95% CI contamination fraction (%)
MA-1 _supernatant	20.9	1	14365	18	2	10.0	0-23.1
MA-1 _1st extraction	76.6	2	326,14365	92	1	1.1	0-3.2
AG-2_supernatant	8.4	1	11152	6	4	40.0	9.6-70.4
AG-2 _1st extraction	7.8	0	-	0	0	-	-
AG-2_merged	15.6	3	54, 13360, 14240	28	15	34.9	20.7-49.1
Iceman	57.3	2	3514, 8138	93	3	3.1	0-6.6

Table SI 7a Table of discordance for reads for MA-1 _supernatant. The position is relative to known polymorphic sites. Minor reads are the number of bases that do not match the most common base.

Position	-4	-3	-2	-1	0	1	2	3	4
Minor reads	5	1	4	5	85	6	3	3	3
All reads	1104	1107	1101	1105	1029	1106	1105	1102	1105

Table SI 7b Table of discordance for reads for MA-1_1st extraction. The position is relative to known polymorphic sites. Minor reads are the number of bases that do not match the most common base.

Position	-4	-3	-2	-1	0	1	2	3	4
Minor reads	16	9	9	14	41	13	14	13	12
All reads	4209	4234	4237	4229	4209	4234	4226	4224	4231

Table SI 8a Test and estimates of contamination for MA-1_supernatant. *The p-value is obtained using fisher's exact test.

Test	p-value*	Estimate (%)	SE
Test 1	<1e-10	22.0	2.18
Test 2	<1e-10	22.6	3.46

Table SI 8b Test and estimates of contamination for MA-1_1st extraction. *The p-value is obtained using fisher's exact test.

Test	p-value*	Estimate (%)	SE
Test 1	4.8e-9	1.96	0.477
Test 2	0.0050	1.59	0.754

Table SI 9. Table of discordance for reads for AG-2_merged. The position is relative to known polymorphic sites. Minor reads are the number of bases that do not match the most common base.

Position	-4	-3	-2	-1	0	1	2	3	4
Minor reads	1	2	1	1	46	0	2	1	1
All reads	485	484	484	484	439	484	483	482	483

Table SI 10. Test and estimates of contamination for AG-2_merged. *The p-value is obtained using fisher's exact test.

Test	p-value*	Estimate (%)	SE
Test 1	<1e-10	28.3	3.65
Test 2	<1e-10	29.7	5.69

SI 6. Error rate estimation and DNA damage

SI 6.1 Error rate estimation

SI 6.1.1 Description of the method

The error rates (single nucleotide polymorphisms (SNPs), sequencing errors or *post-mortem* DNA damage) were estimated using a method similar to a previously published method[1] that makes use of a high quality genome. The estimation is based on the rationale that any given human sample should have the same expected number of derived alleles compared to some outgroup, in this case the chimpanzee. The numbers of derived alleles are counted from the high quality genome and it is assumed that any excess of derived alleles (compared to the high quality genome) observed in our sample is due to errors. If the high quality genome has no errors then the error rate estimate of the sample is equal to the true error rate. However, if the high quality genome does have errors, the estimated error rate can roughly be understood as the excess error rate relative to the error rate of the high quality genome.

The overall error rates were estimated using a method of moment estimator, while the type specific error rates were estimated based on a maximum likelihood approach. The model and the estimation methods are described in details elsewhere[2].

SI 6.1.2 Data

The chimpanzee, panTro2, from the multiway alignment hg19 multiz46, which also includes human was used as an outgroup. For the high quality genome, sequencing data for an individual (NA06985) from the 1000 Genomes Project Consortium[3] was used, and all reads with a mapping quality score less than 30 and all bases with a base quality score less than 20 were excluded. The same quality filters were used for all the 16 modern genomes (SI 4.2) and, MA-1 and AG-2.

SI 6.1.3 Results

Both the type specific error rates and the overall error rates are shown in Figure SI 2. Note that transitions (C → T, G → A) are observed at much higher rates for the ancient genomes than transversions, the former being caused by *post-mortem* cytosine deamination in ancient DNA templates (SI 6.2). Additionally, the Mari genome has a high error rate, and was initially shown to be a low quality library based on the sequencing statistics (SI 4.2.2).

SI 6.2 DNA damage in MA-1

Figure SI 3, generated using the mapDamage package[4], shows the typical patterns of *post-mortem* ancient DNA damage and degradation that are observed in the MA-1 library. Depurination leading to fragmentation of DNA molecules[5], observed as an increase in guanine (G) and adenine (A) residues upstream of read start, is seen in ~32% and ~40% of all reads, respectively. Also observed at the ends of ~3% of the DNA molecules are cytosine (C) deamination patterns leading to base modification (C

→ T at 5' ends/ G → A at 3' ends)[5].

References for SI 6

1. Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053-1060 (2010)

2. Orlando, L. *et al.* Recalibrating *Equus* evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* **499**, 74-78 (2013)

3. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012)

4. Ginolhac, A., Rasmussen, M., Gilbert, M.T.P., Willerslev, E., Orlando, L. mapDamage: testing for damage patterns in ancient DNA sequences. *Bioinformatics* **27**, 2153-2155 (2011)

5. Briggs, A.W. *et al*. Patterns of damage in genomic DNA sequences from a Neanderthal. *Proc.Natl.Acad.Sci.U.S.A.* **104**, 14616-14621 (2007)

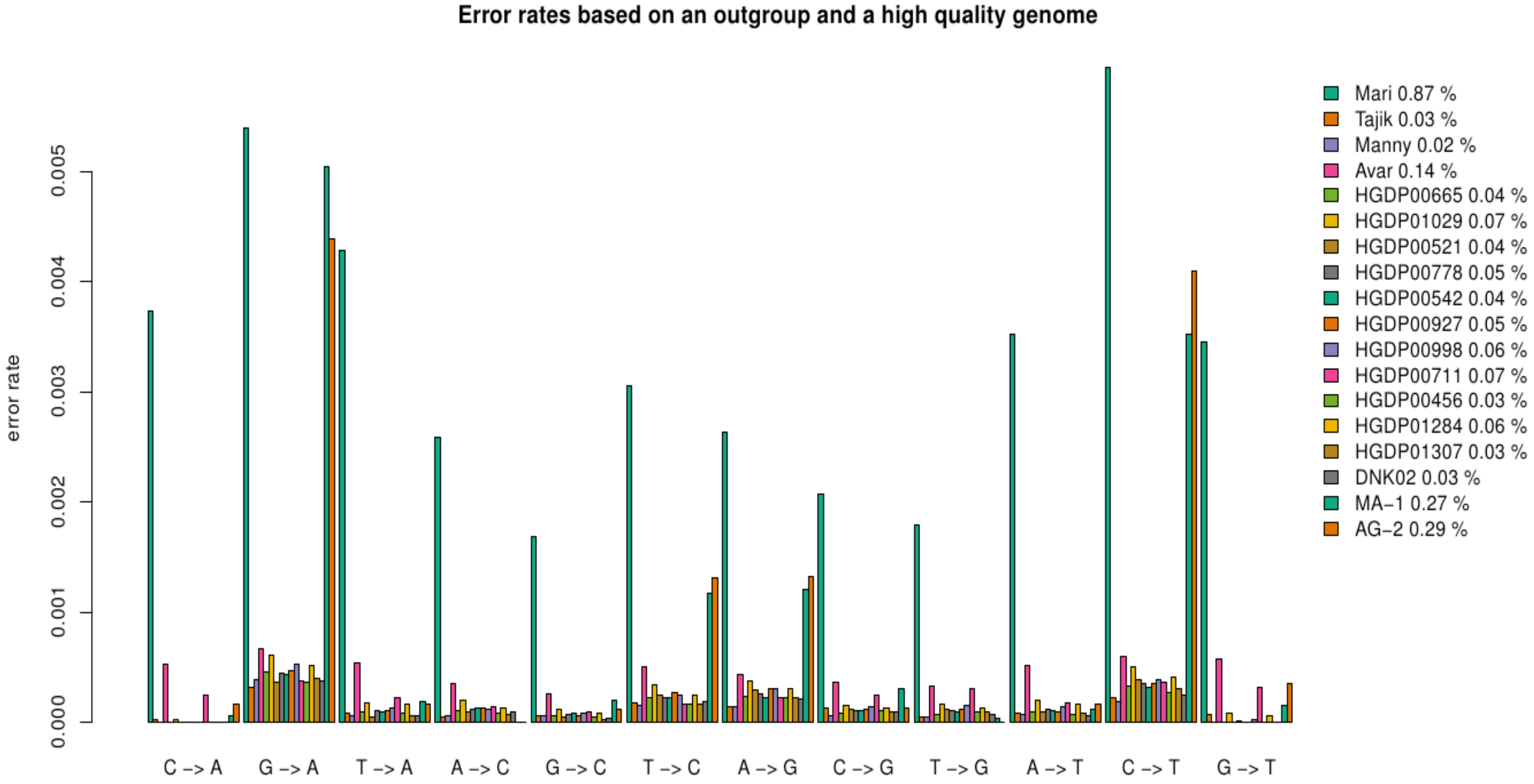


Figure SI 2 Type specific error estimates. Error estimates are presented for the 16 modern genomes and, MA-1 and AG-2. Each bar represents the type specific error of a sample. The overall error rates are shown in the right panel, next to the sample ID. ‘Manny’ refers to the Indian genome.

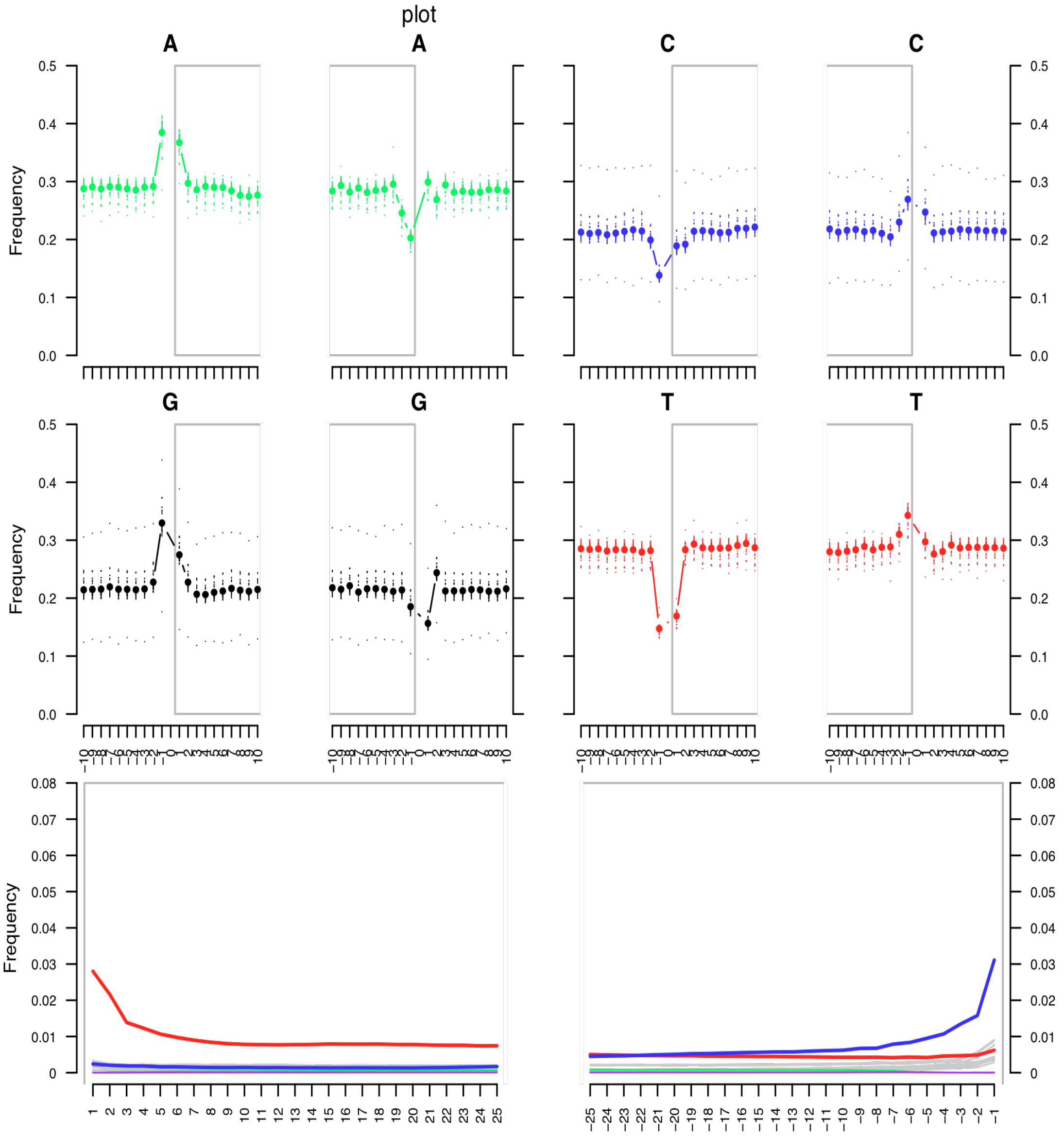


Figure SI 3 DNA damage patterns for MA-1. Depurination leading to DNA fragmentation and cytosine deamination leading to nucleotide misincorporations, observed in the MA-1 library. Depurination patterns are shown using the base composition of the first and last 10 bases that are sequenced (within grey frames), as well as 10 bases located immediately upstream and downstream of the reads. Each dot represents the average base composition at that position. Cytosine deamination patterns are shown in the bottom panel, with the frequency of C → T mismatches in red and G → A mismatches in blue.

SI 7. mtDNA haplogroup of MA-1

Sequence reads from MA-1 were mapped to the revised Cambridge Reference Sequence (rCRS, NC_012920.1) and, filtered for PCR duplicates and paralogs requiring a minimum mapping quality of 25 (SI 4.1). A file of variants, filtered for a minimum depth of 10, was generated. Indels were excluded from the analysis. MA-1 carries all the three SNPs characterizing hg U (A11467G, A12308G and G12372A). For comparison, we included in the analysis the individual Dolni Vestonice 14 (DV-14; GenBank accession number KC521458), which has been shown to be basal to the extant hg U5[1]. Both the MA-1 and DV-14 sequences were analyzed for the presence of diagnostic mutations of the major sub-hgs of extant hg U lineages, using information from mtDNA tree Build 15 (Sept 30, 2012)[2]. A phylogenetic tree was built, with the age estimates (kiloyears, +/- SD) of different sub-hgs of hg U[3], including all major extant branches of mtDNA hg U lineages from its root and also the two ancient hg U lineages (Figure SI 4a). The mtDNA sequences of MA-1 and DV-14 share only the three basal mutations inside hg U with each other, and do not belong to any known modern branch of hg U.

To show the present spread of hg U and its different sub-hgs (Figure SI 4b), the average frequencies, divided into four frequency classes, were calculated in regional groups, using a dataset consisting of *ca*. 30,000 partial mtDNA genomes[4-78]. Today, hg U is a pan western Eurasian haplogroup, with distribution across Europe, the Middle East, South and Central Asia, western Siberia and North Africa (Figure SI 4b). The overall frequency of hg U is low or absent in extant central and south Siberian populations, i.e. the region close to where MA-1 originated, as well as in East Asia (Figure SI 4b).

References for SI 7

1. Fu, Q. *et al.* A revised timescale for human evolution based on ancient mitochondrial genomes. *Curr. Biol.* **23**, 553–559 (2013)

2. Van Oven, M. & Kayser, M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.* **30**, 386–394 (2009)

3. Behar, D. M. *et al.* A “Copernican” reassessment of the human mitochondrial DNA tree from its root. *Am. J. Hum. Genet.* **90**, 675–684 (2012)

4. Rando, J. C. *et al.* Phylogeographic patterns of mtDNA reflecting the colonization of the Canary Islands. *Ann. Hum. Genet.* **63**, 413–428 (1999)

5. Rando, J. C. *et al.* Mitochondrial DNA analysis of northwest African populations reveals genetic exchanges with European, near-eastern, and sub-Saharan populations. *Ann. Hum. Genet.* **62**, 531–550 (1998)

6. Corte-Real, H. B. *et al.* Genetic diversity in the Iberian Peninsula determined from mitochondrial sequence analysis. *Ann. Hum. Genet.* **60**, 331–350 (1996)

7. Behar, D. M. *et al.* The genome-wide structure of the Jewish people. *Nature* **466**, 238–242 (2010)

8. Macaulay, V. A. *et al.* The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. *Am. J. Hum. Genet.* **64**, 232–249 (1999)

9. Richards, M. *et al.* Tracing European founder lineages in the Near Eastern mtDNA pool. *Am. J. Hum. Genet.* **67**, 1251–1276 (2000)

10. Tambets, K. *et al.* The topology of the maternal lineages of the Anatolian and Trans-Caucasus populations and the peopling of the Europe: some preliminary considerations. *Archaeogenetics: DNA and the population prehistory of Europe*. McDonald Institute for Archaeological Research, Cambridge. 219–235 (2000)

11. Bermisheva, M. A. *et al.* Phylogeographic analysis of mitochondrial DNA in the Nogays: A strong mixture of maternal lineages from eastern and western Eurasia. *Mol. Biol. (Mosk)* **38**, 516–523 (2004)

12. Krings, M. *et al.* mtDNA analysis of Nile River Valley populations: A genetic corridor or a barrier to migration? *Am. J. Hum. Genet.* **64**, 1166–1176 (1999)

13. Kivisild, T. *et al.* Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears. *Am. J. Hum. Genet.* **75**, 752–770 (2004)

14. Quintana-Murci, L. *et al.* Where West meets East: The complex mtDNA landscape of the Southwest and Central Asian corridor. *Am. J. Hum. Genet.* **74**, 827–845 (2004)

15. Metspalu, M. *et al.* Most of the extant mtDNA boundaries in South and Southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. *BMC Genet.* **5**, 26 (2004)

16. Calafell, F., Underhill, P., Tolun, A., Angelicheva, D., Kalaydjieva, L. From Asia to Europe: mitochondrial DNA sequence variability in Bulgarians and Turks. *Ann. Hum. Genet.* **60**, 35–49 (1996)

17. Richard, C. *et al.* An mtDNA perspective of French genetic variation. *Ann. Hum. Biol.* **34**, 68–79 (2007)

18. Bertranpetit, J. *et al.* Human mitochondrial DNA variation and the origin of Basques. *Ann. Hum. Genet.* **59**, 63–81 (1995)

19. Helgason, A. *et al.* mtDNA and the islands of the North Atlantic: estimating the proportions of Norse and Gaelic ancestry. *Am. J. Hum. Genet.* **68**, 723–37. (2001)

20. Piercy, R., Sullivan, K. M., Benson, N., Gill, P. The application of mitochondrial DNA typing to the study of white Caucasian genetic identification. *Int J. Legal Med.* **106**, 85–90 (1993)

21. Rousselet, F., Mangin, P. Mitochondrial DNA polymorphisms: a study of 50 French Caucasian individuals and application to forensic casework. *Int. J. Legal Med.* **111**, 292–298 (1998)

22. Cali, F. *et al.* MtDNA control region and RFLP data for Sicily and France. *Int. J. Legal Med.* **114**, 229–231 (2001)

23. Helgason, A., Sigurdadottir, S., Gulcher, J., Ward, R., Stefanson, K. mtDNA and the origins of the Icelanders: deciphering signals of recent population history. *Am. J. Hum. Genet.* **66**, 999–1016 (2000)

24. Sajantila, A. *et al.* Genes and languages in Europe: an analysis of mitochondrial lineages. *Genome Res.* **5**, 42–52 (1995)

25. Pereira, L., Prata, M. J. & Amorim, A. Diversity of mtDNA lineages in Portugal: not a genetic edge of European variation. *Ann. Hum. Genet.* **64**, 491–506. (2000)

26. Larruga, J. M., Diez, F., Pinto, F. M., Flores, C., Gonzalez, A. M. Mitochondrial DNA characterisation of European isolates: the Maragatos from Spain. *Eur. J. Hum. Genet.* **9**, 708–716. (2001)

27. Salas, A., Comas, D., Lareu, M. V, Bertranpetit, J. & Carracedo, A. mtDNA analysis of the Galician population: a genetic edge of European variation. *Eur. J. Hum. Genet.* **6**, 365–375 (1998)

28. Crespillo, M. *et al.* Mitochondrial DNA sequences for 118 individuals from northeastern Spain. *Int. J. Legal Med.* **114**, 130–132 (2000)

29. Kittles, R. A. *et al.* Autosomal, mitochondrial, and Y chromosome DNA variation in Finland: evidence for a male-specific bottleneck. *Am. J. Phys. Anthropol.* **108**, 381–399 (1999)

30. Pult, I. *et al.* Mitochondrial DNA sequences from Switzerland reveal striking homogeneity of European populations. *Biol. Chem. Hoppe Seyler* **375**, 837–840 (1994)

31. Meinilä, M., Finnilä, S., Majamaa, K. Evidence for mtDNA admixture between the Finns and the Saami. *Hum. Hered.* **52**, 160–170 (2001)

32. Dupuy, B. M., Olaisen, B. MtDNA sequences in the Norwegian Saami and main population. *Advances in forensic haemogenetics.* **6,** 23–25 (1996)

33. Passarino, G. *et al.* Different genetic components in the Norwegian population revealed by the analysis of mtDNA and Y chromosome polymorphisms. *Eur. J. Hum. Genet.* **10**, 521–529 (2002)

34. Opdal, S. H. *et al.* Increased number of substitutions in the D-loop of mitochondrial DNA in the sudden infant death syndrome. *Acta Paediatr.* **87**, 1039–1044 (1998)

35. Tambets, K. *et al.* The Western and Eastern Roots of the Saami--the Story of Genetic "Outliers" Told by Mitochondrial DNA and Y Chromosomes. *Am. J. Hum. Genet.* **74**, 661–682 (2004)

36. Delghandi, M., Utsi, E. & Krauss, S. Saami mitochondrial DNA reveals deep maternal lineage clusters. *Hum. Hered.* **48**, 108–114 (1998)

37. Pliss, L. *et al.* Mitochondrial DNA portrait of Latvians: towards the understanding of the genetic structure of Baltic-speaking populations. *Ann. Hum. Genet.* **70**, 439–458 (2006)

38. Bermisheva, M., Tambets, K., Villems, R., Khusnutdinova, E. Diversity of mitochondrial DNA haplotypes in ethnic populations of the Volga-Ural region of Russia. *Mol. Biol. (Mosk)* **36**, 990–1001 (2002)

39. Belyaeva, O. *et al.* Mitochondrial DNA variations in Russian and Belorussian populations. *Human biology* **75**, 647–60 (2003)

40. Malyarchuk, B. A. *et al.* Mitochondrial DNA variability in Poles and Russians. *Ann. Hum. Genet.* **66**, 261–283 (2002)

41. Orekhov, V. *et al.* Mitochondrial DNA sequence diversity in Russians. *FEBS Letters* **445**, 197–201 (1999)

42. Malyarchuk, B. A., Derenko, M. V Mitochondrial DNA variability in Russians and Ukrainians: Implications to the origin of the Eastern Slavs. *Ann. Hum. Genet.* **65**, 63–78 (2001)

43. Belledi, M. *et al.* Maternal and paternal lineages in Albania and the genetic structure of Indo-European populations. *Eur. J. Hum. Genet.* **8**, 480–486 (2000)

44. Cvjetan, S. *et al.* Frequencies of mtDNA haplogroups in southeastern Europe--Croatians, Bosnians and Herzegovinians, Serbians, Macedonians and Macedonian Romani. *Coll. Antropol.* **28**, 193–198 (2004)

45. Varesi, L. *et al.* Mitochondrial control-region sequence variation in the Corsican population, France. *Am. J. Human Biol.* **12**, 339–351. (2000)

46. Tolk, H. V. *et al.* The evidence of mtDNA haplogroup F in a European population and its ethnohistoric implications. *Eur. J. Hum. Genet.* **9**, 717–723 (2001)

47. Tagliabracci, A., Turchi, C., Buscemi, L., Sassaroli, C. Polymorphism of the mitochondrial DNA control region in Italians. *Int. J. Legal. Med.* **114**, 224–228 (2001)

48. Francalacci, P., Bertranpetit, J., Calafell, F., Underhill, P. A. Sequence diversity of the control region of mitochondrial DNA in Tuscany and its implications for the peopling of Europe. *Am. J. Phys. Anthropol.* **100**, 443–460 (1996)

49. Mogentale-Profizi, N. *et al.* Mitochondrial DNA sequence diversity in two groups of Italian Veneto speakers from Veneto. *Ann. Hum. Genet.* **65**, 153–166 (2001)

50. Parson, W., Parsons, T. J., Scheithauer, R., Holland, M. M. Population data for 101 Austrian Caucasian mitochondrial DNA d-loop sequences: application of mtDNA sequence analysis to a forensic case. *Int. J. Legal Med.* **111**, 124–132 (1998)

51. Handt, O. *et al.* Molecular genetic analyses of the Tyrolean Ice Man. *Science* **264**, 1775–1778 (1994)

52. Vanecek, T., Vorel, F., Sip, M. Mitochondrial DNA D-loop hypervariable regions: Czech population data. *Int. J. Legal Med*. **118**, 14–18 (2004)

53. Richards, M. *et al.* Paleolithic and neolithic lineages in the European mitochondrial gene pool. *Am. J. Hum. Genet.* **59**, 185–203 (1996)

54. Pfeiffer, H., Forster, P., Ortmann, C., Brinkmann, B. The results of an mtDNA study of 1,200 inhabitants of a German village in comparison to other Caucasian databases and its relevance for forensic casework. *Int. J. Legal Med.* **114**, 169–172 (2001)

55. Lutz, S., Weisser, H. J., Heizmann, J., Pollak, S. Location and frequency of polymorphic positions in the mtDNA control region of individuals from Germany. *Int. J. Legal Med.* **111**, 67–77 (1998)

56. Baasner, A., Schafer, C., Junge, A., Madea, B. Polymorphic sites in human mitochondrial DNA control region sequences: population data and maternal inheritance. *Forensic Science International* **98**, 169–178 (1998)

57. Hofmann, S. *et al.* Population genetics and disease susceptibility: characterization of central European haplogroups by mtDNA gene mutations, correlation with D loop variants and association with disease. *Hum. Mol. Genet.* **6**, 1835–1846 (1997)

58. Lehocký, I., Baldovic, M., Kádasi, L. & Metspalu, E. A database of mitochondrial DNA hypervariable regions I and II sequences of individuals from Slovakia. *Forensic Sci. Int-Gen.* **2**, e53–9 (2008)

59. Dimo-Simonin, N., Grange, F., Taroni, F., Brandt-Casadevall, C. & Mangin, P. Forensic evaluation of mtDNA in a population from south west Switzerland. *Int. J. Legal Med.* **113**, 89–97 (2000)

60. Derbeneva, O. A., Starikovskaia, E. B., Volod'ko, N. V, Wallace, D. C. & Sukernik, R. I. Mitochondrial DNA variation in Kets and Nganasans and the early peoples of Northern Eurasia. *Genetika* **38**, 1554–1560 (2002)

61. Saillard, J., Evseva, I., Tranebjaerg, L., Norby, S. Mitochondrial DNA diversity among Nenets. *Archaeogenetics: DNA and and the population prehistory of Europe*. McDonald Institute for Archaeological Research, Cambridge. 255–258 (2000)

62. Derbeneva, O. A., Starikovskaya, E. B., Wallace, D. C., Sukernik, R. I. Traces of early Eurasians in the Mansi of northwest Siberia revealed by mitochondrial DNA analysis. *Am. J. Hum. Genet.* **70**, 1009–1114 (2002)

63. Kong, Q.-P. *et al.* Mitochondrial DNA sequence polymorphisms of five ethnic populations from northern China. *Human genetics* **113**, 391–405 (2003)

64. Fedorova, S. A. *et al.* Autosomal and uniparental portraits of the native populations of Sakha (Yakutia): implications for the peopling of Northeast Eurasia. *BMC Evolutionary Biology* **13**, 127 (2013)

65. Derenko, M. V. *et al.* Diversity of mitochondrial DNA lineages in South Siberia. *Ann. Hum. Genet.* **67**, 391–411 (2003)

66. Comas, D. *et al.* Admixture, migrations, and dispersals in Central Asia: evidence from maternal DNA lineages. *Eur. J. Hum. Genet.* **12**, 495–504 (2004)

67. Yao, Y. G., Lu, X. M., Luo, H. R., Li, W. H., Zhang, Y. P. Gene admixture in the silk road region of China: evidence from mtDNA and melanocortin 1 receptor polymorphism. *Genes Genet. Syst.* **75**, 173–178 (2000)

68. Comas, D. *et al.* Trading genes along the silk road: mtDNA sequences and the origin of Central Asian populations. *Am. J. Hum. Genet.* **63**, 1824–1838 (1998)

69. Yao, Y.-G. *et al.* Genetic relationship of Chinese ethnic populations revealed by mtDNA sequence diversity. *Am. J. Phys. Anthropol.* **118**, 63–76 (2002)

70. Kivisild, T. *et al.* The emerging limbs and twigs of the East Asian mtDNA tree. *Mol. Biol. Evol.* **19**, 1737–1751 (erratum 20:162) (2002)

71. Yao, Y.-G., Kong, Q.-P., Bandelt, H.-J., Kivisild, T., Zhang, Y.-P. Phylogeographic differentiation of mitochondrial DNA in Han Chinese. *Am. J. Hum. Genet.* **70**, 635–651 (2002)

72. Nishimaki, Y. *et al.* Sequence polymorphism in the mtDNA HV1 region in Japanese and Chinese. *Legal Med.* **1**, 238–249 (1999)

73. Tsai, L. C. *et al.* Sequence polymorphism of mitochondrial D-loop DNA in the Taiwanese Han population. *Forensic Sci. Int.* **119**, 239–47 (2001)

74. Imaizumi, K., Parsons, T. J., Yoshino, M. & Holland, M. M. A new database of mitochondrial DNA hypervariable regions I and II sequences from 162 Japanese individuals. *International Journal of Legal Medicine* **116**, 68–73 (2002)

75. Nagai, A., Nakamura, I., Shiraki, F., Bunai, Y. & Ohya, I. Sequence polymorphism of mitochondrial DNA in Japanese individuals from Gifu Prefecture. *Legal medicine (Tokyo, Japan)* **5 Suppl 1**, S210–3 (2003)

76. Snäll, N. *et al.* A rare mitochondrial DNA haplotype observed in Koreans. *Human Biology* **74**, 253–62 (2002)

77. Kolman, C., Sambuughin, N., Bermingham, E. Mitochondrial DNA analysis of Mongolian populations and implications for the origin of New World founders. *Genetics* **142**, 1321–1334 (1996)

78. Oota, H., Settheetham-Ishida, W., Tiwawech, D., Ishida, T., Stoneking, M. Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. *Nat. Genet.* **29**, 20–21 (2001)

79. Derbeneva, O. A. *et al.* Analysis of mitochondrial DNA diversity in the Aleuts of the Commander Islands and its implications for the genetic history of Beringia. *Am. J. Hum. Genet.* **71**, 415–421 (2002)

80. Starikovskaya, Y. B., Sukernik, R. I., Schurr, T. G., Kogelnik, A. M., Wallace, D. C. mtDNA diversity in Chukchi and Siberian Eskimos: implications for the genetic history of Ancient Beringia and the peopling of the New World. *Am. J. Hum. Genet.* **63**, 1473–1491 (1998)

81. Schurr, T. G., Sukernik, R. I., Starikovskaya, Y. B., Wallace, D. C. Mitochondrial DNA variation in Koryaks and Itel'men: population replacement in the Okhotsk Sea-Bering Sea region during the Neolithic. *Am. J. Phys. Anthropol.* **108**, 1–39 (1999)

82. Derenko, M. V., Shields, G. F. Diversity of mitochondrial DNA nucleotide sequences in three groups of aboriginal inhabitants of Northern Asia. *Mol. Biol. (Mosk)* **31**, 784–789 (1997)

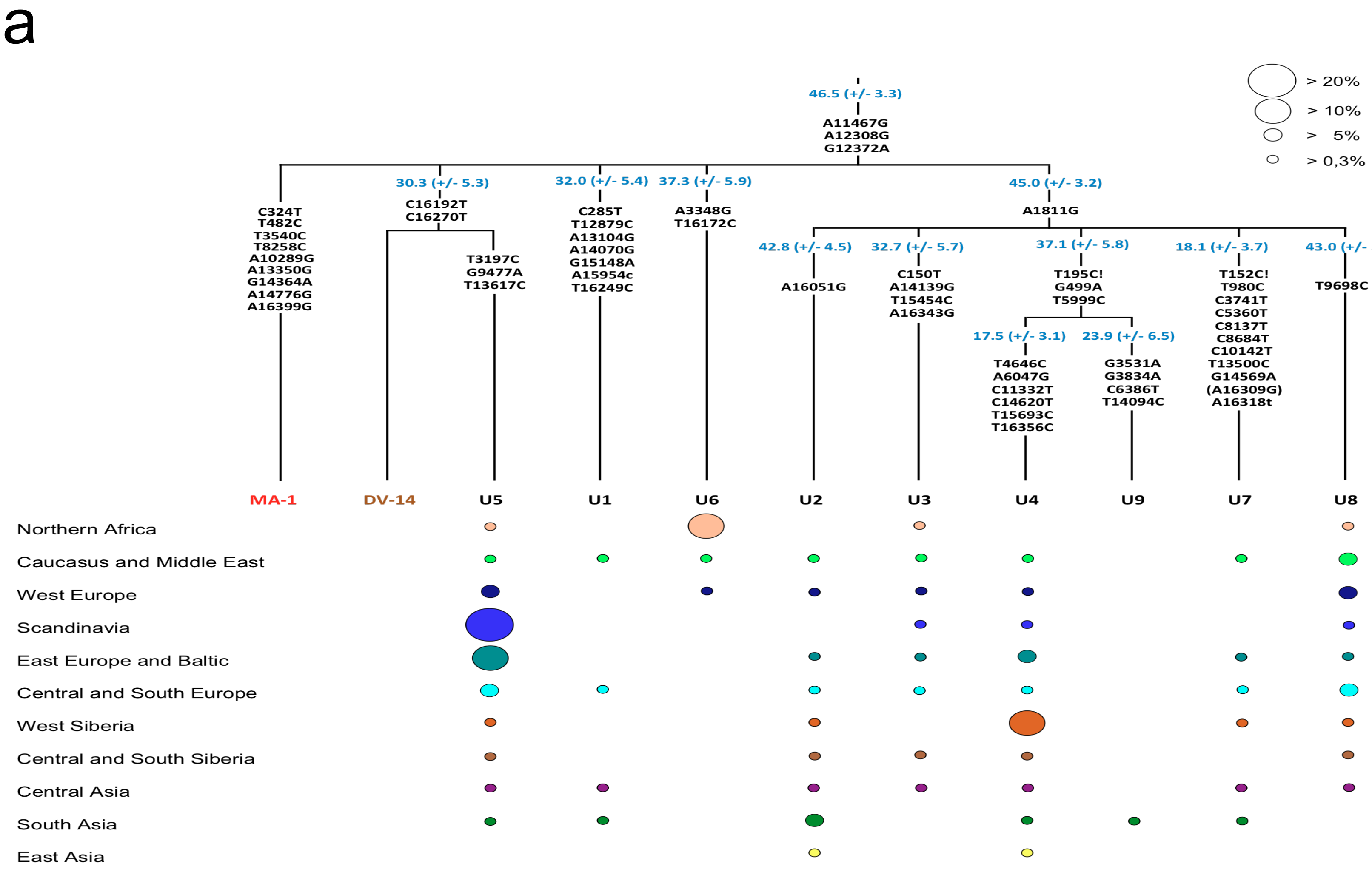

a
46.5 (+/- 3.3)
A11467G
A12308G
G12372A
> 20%
> 10%
> 5%
> 0,3%
C324T
T482C
T3540C
T8258C
A10289G
A13350G
G14364A
A14776G
A16399G
30.3 (+/- 5.3)
C16192T
C16270T
T3197C
G9477A
T13617C
32.0 (+/- 5.4)
C285T
T12879C
A13104G
A14070G
G15148A
A15954c
T16249C
37.3 (+/- 5.9)
A3348G
T16172C
45.0 (+/- 3.2)
A1811G
42.8 (+/- 4.5)
A16051G
32.7 (+/- 5.7)
C150T
A14139G
T15454C
A16343G
37.1 (+/- 5.8)
T195C!
G499A
T5999C
17.5 (+/- 3.1)
T4646C
A6047G
C11332T
C14620T
T15693C
T16356C
23.9 (+/- 6.5)
G3531A
G3834A
C6386T
T14094C
18.1 (+/- 3.7)
T152C!
T980C
C3741T
C5360T
C8137T
C8684T
C10142T
T13500C
G14569A
(A16309G)
A16318t
43.0 (+/- 3.6)
T9698C
MA-1
DV-14
U5
U1
U6
U2
U3
U4
U9
U7
U8
Northern Africa
Caucasus and Middle East
West Europe
Scandinavia
East Europe and Baltic
Central and South Europe
West Siberia
Central and South Siberia
Central Asia
South Asia
East Asia

Figure SI 4 mtDNA haplogroup placement of MA-1 (previous page). **(a)** Placement of the ancient mtDNA lineages of MA-1 and Dolni Vestonice 14 (DV-14)[1], for comparison, on a phylogenetic tree depicting the main extant branches of hg U. The haplogroup nomenclature usage is in accordance with the mtDNA tree Build 15 (Sept 30, 2012)[2]. The age estimates (kiloyears, +/- SD) shown on the branches of different sub-hgs of hg U are from Behar *et al.*[3]. The regional frequency data of hg U sub-hgs is pooled into four frequency classes (shown in the legend) and comprises data from previous studies and unpublished data from Estonian Biocentre (indicated with asterisks): Northern Africa[4-7]; Caucasus and Middle East[7-16,*] ; West Europe[6,9,17-28]; Scandinavia[9,19,24,29-36,*]; East Europe and Baltic[9,24,37-42,*]; Central and South Europe[9,16,22,43-59,*]; West Siberia[60-62,*]; Central and South Siberia[63-65,*]; Central Asia[14,66-68,*]; South Asia[14,15,69]; East Asia[63,67-77].

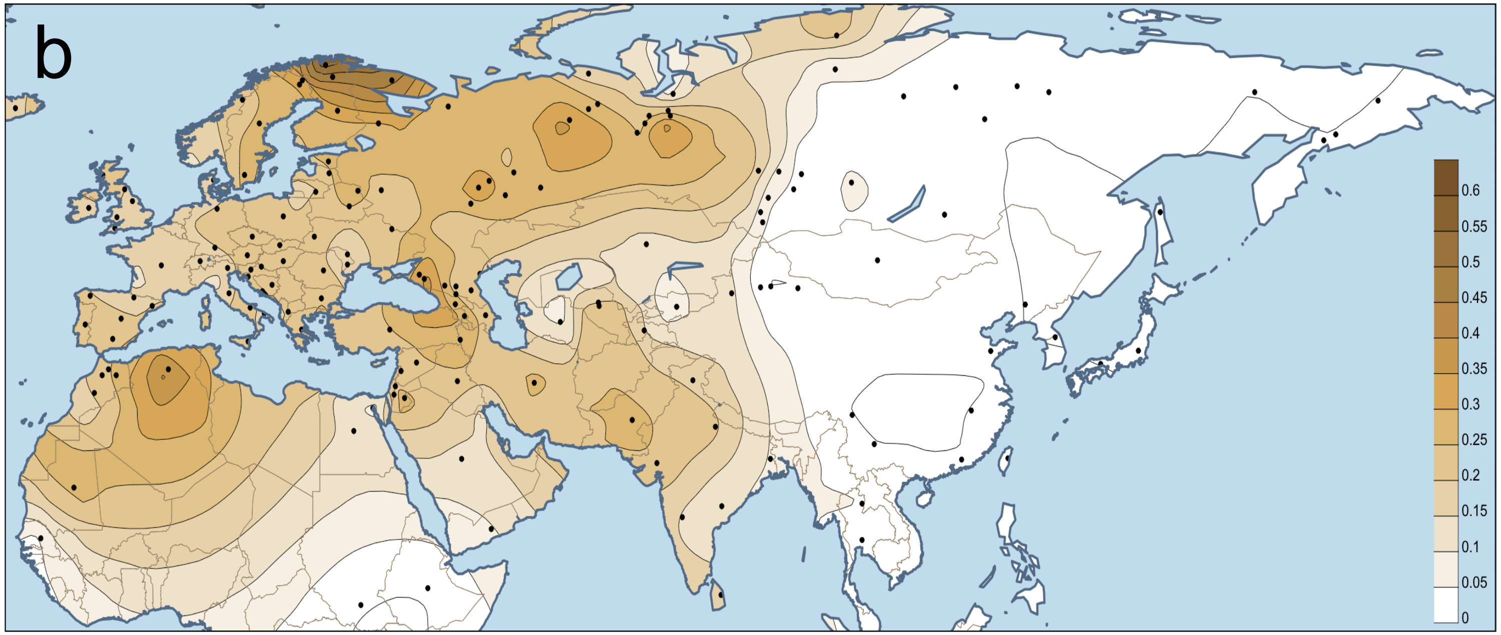


Figure SI 4 mtDNA haplogroup placement of MA-1. (b) Geographical spread of hg U. A map of Eurasia and North Africa showing the frequency distribution of hg U (shown as a fraction of all known mtDNA hgs) was generated using the Kriging method in Surfer 8 program (Golden Software, Inc.). The data points[78-82] (and references from panel a) are shown as black dots, the scale bar indicates the spatial frequency differences of hg U from lowest (white) to highest (dark brown) values.

SI 8. Y chromosome haplogroup of MA-1

Due to low depth-of-coverage of the MA-1 individual (1.5X on 5.8 million bases), genotyping at each site on the Y chromosome was performed by selecting the allele with the highest frequency of bases with a base quality of 13 or higher. Additionally, a multi-fasta file was generated from the variable positions on the Y chromosomes available from 24 Complete Genomics public genomes[1]. SNPs were filtered for quality (using VQHIGH as the threshold, as defined by Complete Genomics), with tri-allelic positions excluded and only those Y chromosome regions determined as being phylogenetically informative being used[2]. This yielded a final dataset of 22492 positions. MA-1 Y chromosome data was then included, and MEGA phylogenetic software[3] was used to construct a Neighbor Joining (NJ) tree with default parameters (Figure SI 5a). MA-1 is placed as a basal lineage to hg R[2,4]. Phylogenetically informative positions and their state in MA-1 were then determined to confirm the placement of MA-1 on the tree. In the course of this analysis, the original dataset was severely pruned. Non-informative positions, including those with more than four Ns in the public dataset, were excluded (633 positions). Moreover, the following positions were also excluded which were 1) in reference state in all individuals including MA-1 (7172 positions); 2) N in MA-1 and either N or reference state among the rest of the individuals (9682 positions); 3) ‘N-ref’ – those with only N or reference state among all individuals (586 positions) and ‘N-alt’ - positions with alternative alleles, but difficult to classify (11 positions); 4) reference specific (79 positions); and, 5) recurrent (28 positions). This resulted in 4301 positions being retained that were classified according to their hg affiliations. Among those phylogenetically informative positions, 1889 non-N positions were retrieved from MA-1. When counting from the split of hg DE on the unrooted phylogenetic tree, MA-1 is determined to be carrying the derived allele in 183 sites and the ancestral allele in 1706 sites. The position of MA-1 on the phylogenetic tree is established by the state of the 313 basal mutations separating hgs DE and R, where MA-1 has 143 informative positions. Of these, 138 are in the derived and 5 in the ancestral state, placing MA-1 as a lineage basal to hg R. With only a few exceptions characterized below, all other informative positions in MA-1 are in the ancestral state, further supporting the phylogenetic positioning of MA-1 on the tree.

Among the derived markers in the final dataset only a few (11) mutations were detected that are likely to be false positives based on the phylogenetic analysis, where it is assumed that recurrent mutation is less likely than a sequencing error. One position among the 35 private to MA-1 is characteristic of a distant hg – namely C3c1[4]. Based on current data, 10 additional phylogenetically non-concordant positions in MA-1 were found – 1 position for hgs E, G, Q, R1b, R1 each, 2 defining positions for hg I and 3 private mutations for R1b individuals (shown in red on Figure SI 5a). Additionally, among the mutations originally excluded (the reference-private mutations), two positions were found where MA-1 is in derived state.

References for SI 8

1. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78-81 (2010)

2. Wei, W. *et al.* A calibrated human Y-chromosomal phylogeny based on resequencing. *Genome Res.* **23**, 388-395 (2013)

3. Tamura, K. *et al.* MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* **28**, 2731-2739 (2011)

4. *International Society of Genetic Genealogy Y-DNA Haplogroup Tree 2013, Version: 8.58*, <http://www.isogg.org/tree/> (2012)

5. Bosch, E. *et al.* High-resolution analysis of human Y-chromosome variation shows a sharp discontinuity and limited gene flow between northwestern Africa and the Iberian Peninsula. *Am. J. Hum. Genet.* **68**, 1019-1029 (2001)

6. Balanovsky, O. *et al.* Two sources of the Russian patrilineal heritage in their Eurasian context. *Am. J. Hum. Genet.* **82**, 236-250 (2008)

7. Battaglia, V. *et al.* Y-chromosomal evidence of the cultural diffusion of agriculture in Southeast Europe. *European Journal of Human Genetics* **17**, 820-830 (2009)

8. Al-Zahery, N. *et al.* Y-chromosome and mtDNA polymorphisms in Iraq, a crossroad of the early human dispersal and of post-Neolithic migrations. *Molecular phylogenetics and evolution* **28**, 458-472. (2003).

9. Cadenas, A. M., Zhivotovsky, L. A., Cavalli-Sforza, L. L., Underhill, P. A., Herrera, R. J. Y-chromosome diversity characterizes the Gulf of Oman. *European Journal of Human Henetics* **16**, 374-386 (2008)

10. Cinnioglu, C. *et al.* Excavating Y-chromosome haplotype strata in Anatolia. *Hum Genet.* **114**, 127-148 (2004)

11. Dulik, M. C., Osipova, L. P., Schurr, T. G. Y-chromosome variation in Altaian Kazakhs reveals a common paternal gene pool for Kazakhs and the influence of Mongolian expansions. *Plos One* **6**, e17548 (2011)

12. Flores, C. *et al.* Isolates in a corridor of migrations: a high-resolution analysis of Y-chromosome variation in Jordan. *Journal of Human Genetics* **50**, 435-441 (2005)

13. Hammer, M. F. *et al.* Dual origins of the Japanese: common ground for hunter-gatherer and farmer Y chromosomes. *Journal of Human Genetics* **51**, 47-58 (2006)

14. Karafet, T. M. *et al.* High levels of Y-chromosome differentiation among native Siberian populations and the genetic signature of a boreal hunter-gatherer way of life. *Human Biology* **74**, 761-789 (2002)

15. Kharkov, V. N. *et al.* Gene Pool Structure of Eastern Ukrainians as Inferred from the Y-Chromosome haploroups. *Russian Journal of Genetics* **40**, 326-331 (2004)

16. Kivisild, T. *et al.* The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. *The American Journal of Human Genetics* **72**, 313-332 (2003)

17. Lappalainen, T. *et al.* Regional differences among the Finns: a Y-chromosomal perspective. *Gene* **376**, 207-215 (2006)

18. Lappalainen, T. *et al.* Migration waves to the Baltic Sea region. *Annals of Human Genetics* **72**, 337-348 (2008)

19. Luis, J. R. *et al.* The Levant versus the Horn of Africa: evidence for bidirectional corridors of human migrations. *Am. J. Hum. Genet.* **74**, 532-544 (2004)

20. Myres, N. M. *et al.* A major Y-chromosome haplogroup R1b Holocene era founder effect in Central and Western Europe. *European journal of human genetics* **19**, 95-101 (2011)

21. Nasidze, I. *et al.* Mitochondrial DNA and y-chromosome variation in the caucasus. *Annals of Human Genetics* **68**, 205-221 (2004)

22. Regueiro, M., Cadenas, A. M., Gayden, T., Underhill, P. A., Herrera, R. J. Iran: tricontinental nexus for Y-chromosome driven migration. *Hum. Hered.* **61**, 132-143 (2006)

23. Semino, O. *et al.* The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. *Science* **290**, 1155-1159 (2000)

24. Sengupta, S. *et al.* Polarity and temporality of high-resolution y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists. *Am. J. Hum. Genet.* **78**, 202-221

(2006)

25. Tambets, K. *et al.* The Western and Eastern Roots of the Saami--the Story of Genetic "Outliers" Told by Mitochondrial DNA and Y Chromosomes. *Am. J. Hum. Genet.* **74**, 661-682 (2004)

26. Underhill, P. A. *et al.* Separating the post-Glacial coancestry of European and Asian Y chromosomes within haplogroup R1a. *European Journal of Human Genetics* **18**, 479-484 (2010)

27. Yunusbayev, B. *et al.* The Caucasus as an Asymmetric Semipermeable Barrier to Ancient Human Migrations. *Molecular Biology and Evolution* **29**, 359-365 (2012)

28. Sahoo, S. *et al.* A prehistory of Indian Y chromosomes: evaluating demic diffusion scenarios. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 843-848 (2006)

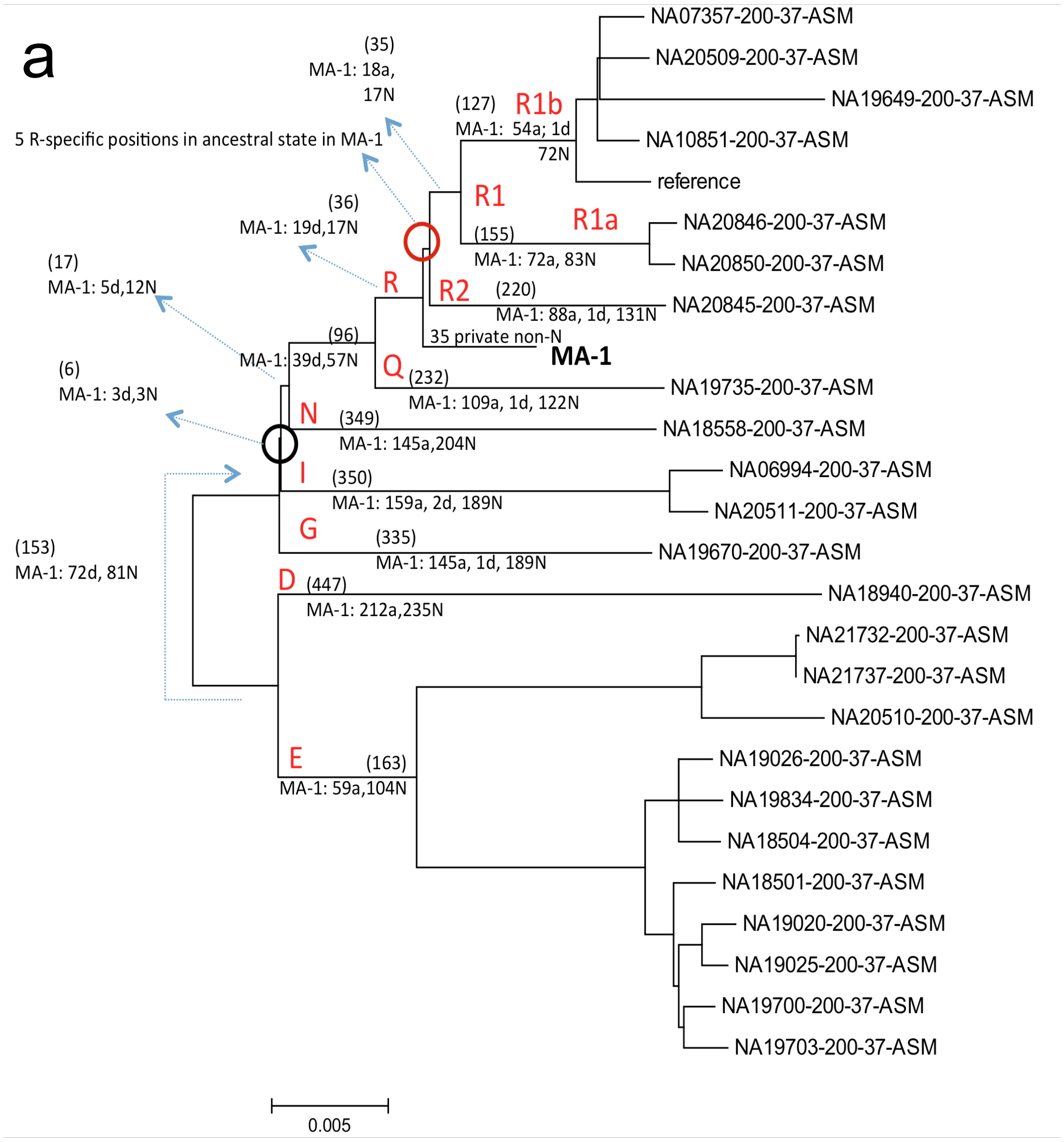


Figure SI 5a An unrooted Neighbor Joining tree of 24 Y chromosomes from the publicly available Complete Genomics dataset[1] and MA-1. The tree is based on 22492 Y chromosome SNPs and constructed with MEGA using default parameters. MA-1 is ancestral to haplogroup R, based on 5 sites that are in ancestral state compared to all other individuals belonging to this haplogroup. The phylogenetic analysis of sites underlying the tree resulted in a pruned dataset of 4301 informative positions. The derived (d) or ancestral (a) states of called (non-N) sites in the MA-1 Y chromosome are noted on the edges of the tree. The overall number of informative sites at those edges is shown in parentheses. For example, "(232) MA-1: 109a, 1d, 122N " on the hg Q edge means that out of the 232 positions in derived state at this edge, 109 were ancestral, one derived and 122 N in MA-1.

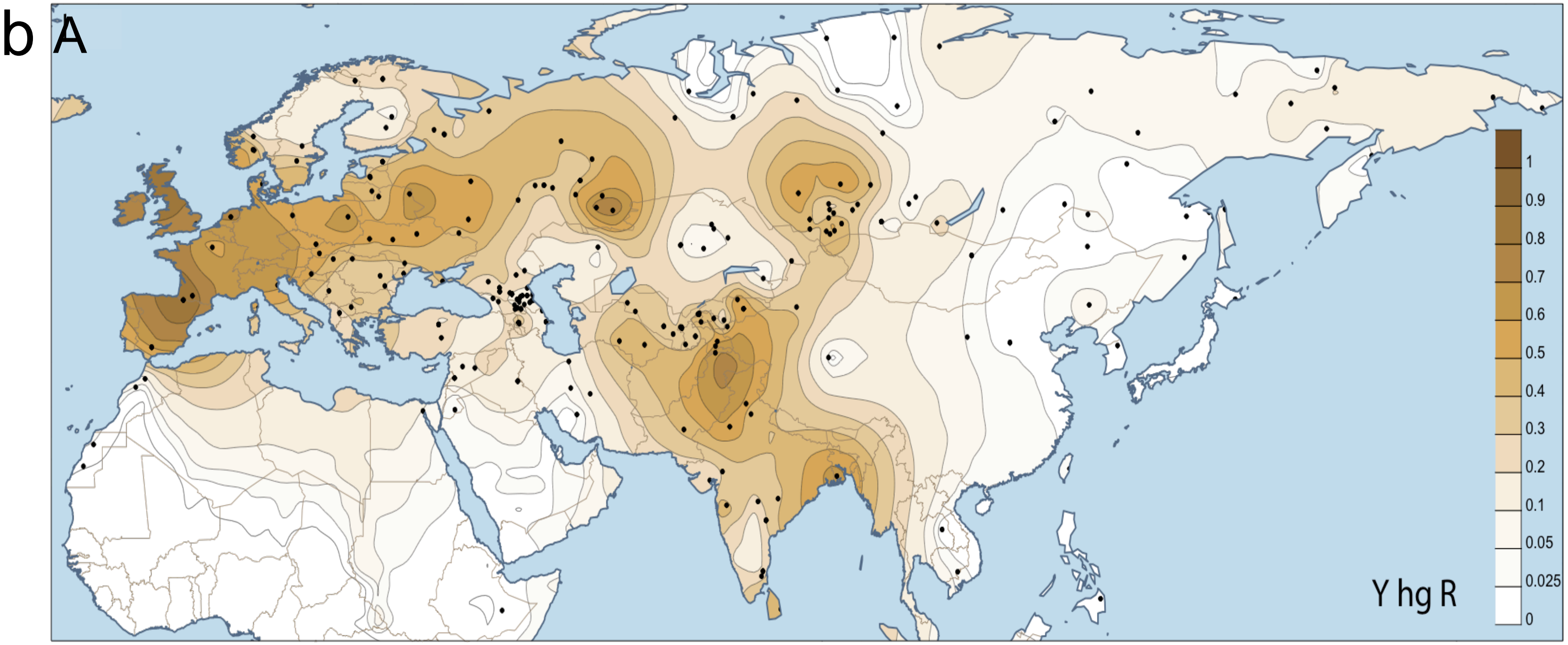


Figure SI 5b The geographical spread of Y-chromosomal haplogroups R and Q in Eurasia - panels A and B, respectively. The frequency patterns (shown as a fraction of all known Y hgs) were generated using the Kriging method in Surfer 8 program (Golden Software, Inc.). The data points are shown in black and the scale bars indicate the haplogroup frequency bins. **Panel A**: Haplogroup R frequencies are aggregates of several sub-clades with partly overlapping, but mostly different, spread patterns. R1b sub-clade is most frequent in western Europe and western Eurasia, having an opposite frequency cline compared to the R1a sub-clade in Europe[20]. R1a is the most widely spread clade with high frequency from East Europe to India, Central Asia and South Siberia[26]. R2 is most common in India with decreasing gradient to other regions[24,28]. The frequency data for surfer plots include data from previous studies[5-27] and unpublished data from the Estonian Biocentre.

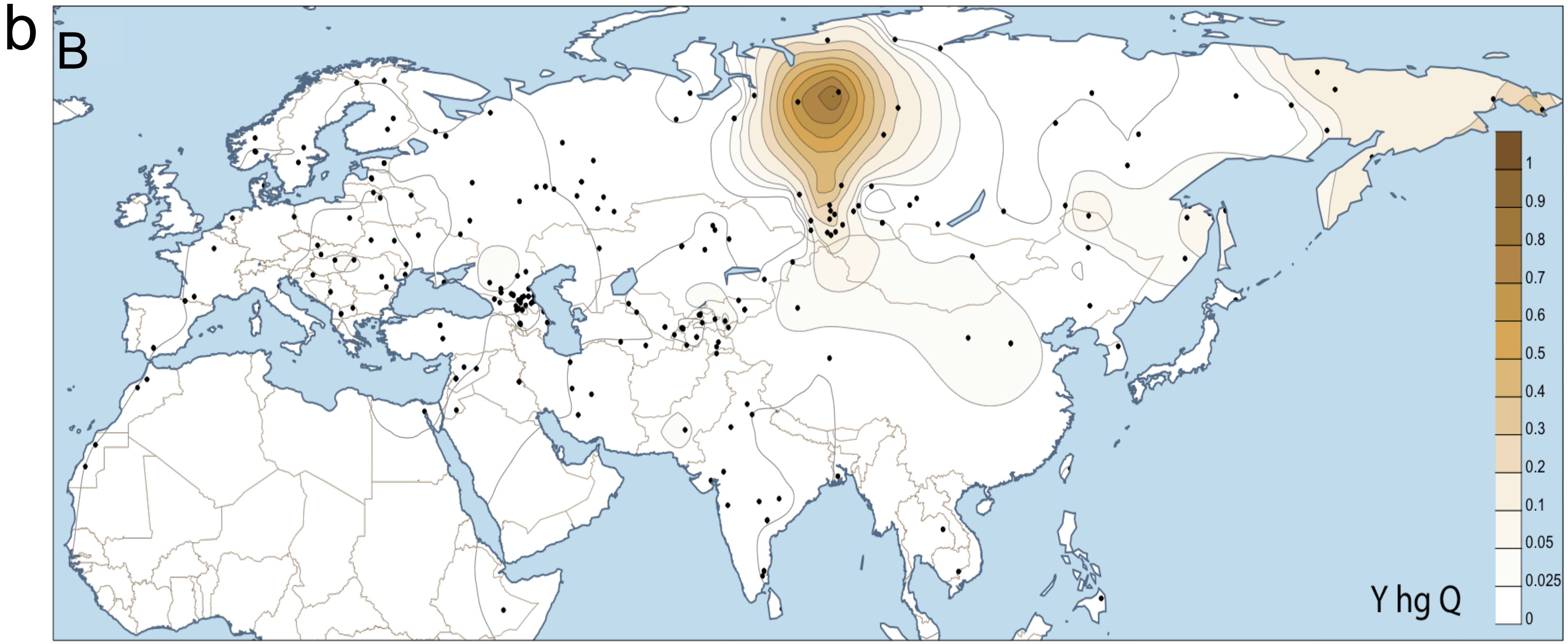


Figure SI 5b The geographical spread of Y-chromosomal haplogroups R and Q in Eurasia - panels A and B, respectively. The frequency patterns (shown as a fraction of all known Y hgs) were generated using the Kriging method in Surfer 8 program (Golden Software, Inc.). The data points are shown in black and the scale bars indicate the haplogroup frequency bins. **Panel B**: Hg Q is present in high frequencies in Siberian populations and in much lower frequencies in other Eurasian regions, like western Eurasia and East Asia. The frequency data for surfer plots include data from previous studies[5-27] and unpublished data from the Estonian Biocentre.

SI 9. Ancestry proportions in MA-1

SI 9.1 ADMIXTURE analysis

SI 9.1.1 Method

A STRUCTURE-like[1], but maximum likelihood-based approach implemented in ADMIXTURE[2] was used to characterize the genomic signatures of MA-1 on the global canvas of genetic structure. ADMIXTURE was run assuming 3 to 15 "ancestral" populations (K=3 to K=15) in 100 replicates. Convergence of individual ADMIXTURE runs were monitored at each K by assessing the maximum difference in Log Likelihood (LL) scores in fractions of runs with the highest LL scores at each K. It was assumed that a global LL maximum was reached at a given K if, say, 10% of the runs with the highest LL score show minimal (<~1 LL unit) variation in LL scores. According to this reasoning, the global LL maximum was reached in runs at K=3 to K=10. ADMIXTURE includes a cross-validation (CV) procedure to help choose the "best" K, which is defined as the K for which the model has the best predictive accuracy. Judging from the cross-validation error distribution, the genetic structure in our sample set is best described at K=9. Figure SI 6 presents plots representing the ADMIXTURE runs that converged close to the maximum LL at each K, from K=3 to K=10, highlighting the model K=9.

SI 9.1.2 Data

A worldwide reference panel, consisting of 1301 samples, was used. It includes 85 previously unpublished samples and 1215 samples from published studies[3-8] (Table SI 11). The samples in the reference panel have been genotyped using three different Illumina genotyping arrays (Human610-Quad, HumanHap650Y, Human660W-Quad). The genotype data from the published and unpublished samples were first combined by array version and then lifted using Liftover tool at UCSC Genome Browser[9] to reflect physical positions of the human genome build 37 (GRCh37) and rs numbers in dbSNP hg19 build 135 using SNAP[10]. Strand was set using the 1000 Genomes Project reference files. AT and GC markers were removed in order to minimize potential strand errors during merger of the data from different Illumina platforms. The dataset used consists of 66,285 SNPs, which were found to be in common between MA-1 and the reference panel of worldwide populations. Sequence read mapping quality of 30 and base quality of 30 was required for MA-1, and a single read was sampled at positions that were covered by multiple reads (we note here that the results using the haploid version of the data, as in this case, looks similar to those using diploid genotypes). Furthermore, all positions where MA-1 displayed a third allele were removed. To allow valid comparisons between the haploid MA-1 data and worldwide modern-day individuals, a single gene copy was sampled from each modern-day individual at all loci[11]. The combined dataset was filtered using PLINK[12] to include only i) SNPs with genotyping success rate of >95% and minor allele frequency >0.5% and, ii) individuals with genotyping success rate of >95%. Despite the low number of SNPs left in the dataset, the data was pruned for LD as ADMIXTURE generally assumes unlinked loci. PLINK was used to calculate an LD (r^2) score for each pair of SNPs in a window of 200 SNPs, and one SNP from the pair was excluded if $r^2 > 0.4$. The window was advanced by 25 SNPs at a time.

Expectedly, only very few markers were removed, and thus the final dataset included 66,260 SNPs in 1302 individuals (including MA-1).

SI 9.1.3 Results

At K = 9, MA-1 is composed of five genetic components of which the two major ones make up *ca*. 70% of the total. The most prominent component is shown in green and is otherwise prevalent in South Asia but does also appear in the Caucasus, Near East or even Europe. The other major genetic component (dark blue) in MA-1 is the one dominant in contemporary European populations, especially among northern and northeastern Europeans. The co-presence of the European-blue and South Asian-green in MA-1 can be interpreted as admixture of the two in MA-1 or, alternatively, MA-1 could represent a proto-western Eurasian prior to the split of Europeans and South Asians. This analysis cannot differentiate between these two scenarios. Most of the remaining nearly one third of the MA-1 genome is comprised of the two genetic components that make up the Native American gene pool (orange and light pink). Importantly, MA-1 completely lacks the genetic components prevalent in extant East Asians and Siberians (shown in dark and light yellow, respectively). Based on this result, it is likely that the current Siberian genetic landscape, dominated by the genetic components depicted in light and dark yellow (Figure SI 6), was formed by secondary wave(s) of immigrants from East Asia.

SI 9.2 NGSadmix analysis

SI 9.2.1 Method

Since the MA-1 genome has an average sequencing depth of 1X, most of its genotypes can only be called with very high uncertainty. Therefore, in addition to the ADMIXTURE analysis described above, a new method called NGSadmix[13] for performing admixture analyses was employed. NGSadmix is a maximum likelihood method that is based on a model very similar to other maximum likelihood-based admixture methods such as Frappe and ADMIXTURE[2,14]. However, whereas all other admixture methods base their inference on called genotypes and implicitly assume that the genotypes are called without error, NGSadmix bases its inference on genotype likelihoods (GLs) and in doing so it takes into account the uncertainty of the genotypes that is inherently present in sequencing data, especially in low depth data.

SI 9.2.2 Data

The analyzed data consisted of raw read data from four sequenced genomes: MA-1, a Karitiana (HGDP00998), a French (HGDP00521) and a Han Chinese (HGDP00778). Additionally, it consisted of a subset of the SNP data from SI 9.1, consisting of individuals from the following 10 populations: Athabascans, Balochi, Chuvash, French, Han, Karitiana, Koryaks, Lezgins, Lithuanians and Tajiks. We used a subset of the SNP dataset because this allows for each of the populations to get their own component.

SI 9.2.3 Data filtering and preparation

Prior to performing the analyses, the SNP chip data was filtered by removing:

- eight individuals that were closely related to other individuals in the dataset (they were estimated to share 1 or 2 alleles Identical-By-Descent in more than 0.25 of their genomes).
- the SNP chip data for the three individuals for whom sequencing data was obtained.

Furthermore, the following SNPs were removed:

- all non-autosomal SNPs
- all SNPs without alleles on the plus strand of hg19
- all SNPs with more than 0.01 missing data
- all SNPs with a minor allele frequency lower than 0.05
- all SNP in high LD (r-squared>0.8)

In total, 338,414 filtered SNPs were retained for the analysis.

The input for NGSadmix is GLs. The likelihoods for the four sequenced genomes were called using the software package ANGSD[15], which implements the methods from samtools[16]. Only the likelihoods for the three genotypes observed in the SNP dataset were used. For the SNP chip data the GLs were set to have a value of 1 for the observed genotype and 0 otherwise, which is equivalent to assuming an error rate of 0. Note that most of the SNPs in the dataset are transitions, which might affect the GLs for the ancient sample.

SI 9.2.4 Results

NGSadmix was run with the number of population components, K, set to 2-8. For each of these K values, NGSadmix was re-run multiple times with different starting points in order to ensure proper convergence. The top likelihood solutions for all K can be seen in Figure SI 7. Results from analyzing only the SNP chip dataset, using ADMIXTURE, can be seen in Figure SI 8. The NGSadmix results correspond well with the ADMIXTURE results in SI 9.1.3.

References for SI 9

1. Pritchard, J. K., Stephens, M., Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959 (2000)

2. Alexander, D. H., Novembre, J., Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655-1664 (2009)

3. Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100-1104 (2008)

4. Behar, D. M. *et al.* The genome-wide structure of the Jewish people. *Nature* **466**, 238-242 (2010)

5. Metspalu, M. *et al.* Shared and Unique Components of Human Population Structure and Genome-Wide Signals of Positive Selection in South Asia. *Am. J. Hum. Genet.* **89**, 731-744 (2011)

6. Rasmussen, M. *et al.* Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463**, 757-762 (2010)

7. Yunusbayev, B. *et al.* The Caucasus as an Asymmetric Semipermeable Barrier to Ancient Human Migrations. *Molecular Biology and Evolution* **29**, 359-365 (2012)

8. Fedorova, S. A. *et al.* Autosomal and uniparental portraits of the native populations of Sakha (Yakutia): implications for the peopling of Northeast Eurasia. *BMC Evol. Biol.* **13**, 127 (2013)

9. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996-1006 (2002)

10. Johnson, A. D. *et al.* SNAP: A web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* **24**, 2938-2939 (2008)

11. Skoglund, P., Jakobsson, M. Archaic human ancestry in East Asia. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 18301-18306 (2011)

12. Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559-575 (2007)

13. Skotte, L., Korneliussen, T.S., Albrechtsen, A. Estimating individual admixture proportions from next generation sequencing data. *Genetics*. Doi: 10.1534/genetics.113.154138 (2013)

14. Tang, H., Peng, J., Wang P, Risch, N. Estimation of Individual Admixture: Analytical and Study Design Considerations. *Genet Epidemiol.* **28**:289-301 (2005)

15. Korneliussen, T.S., Albrechtsen, A, Nielsen, R. ANGSD: Analysis of next generation Sequencing Data. http://www.popgen.dk/angsd (2012)

16. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009)

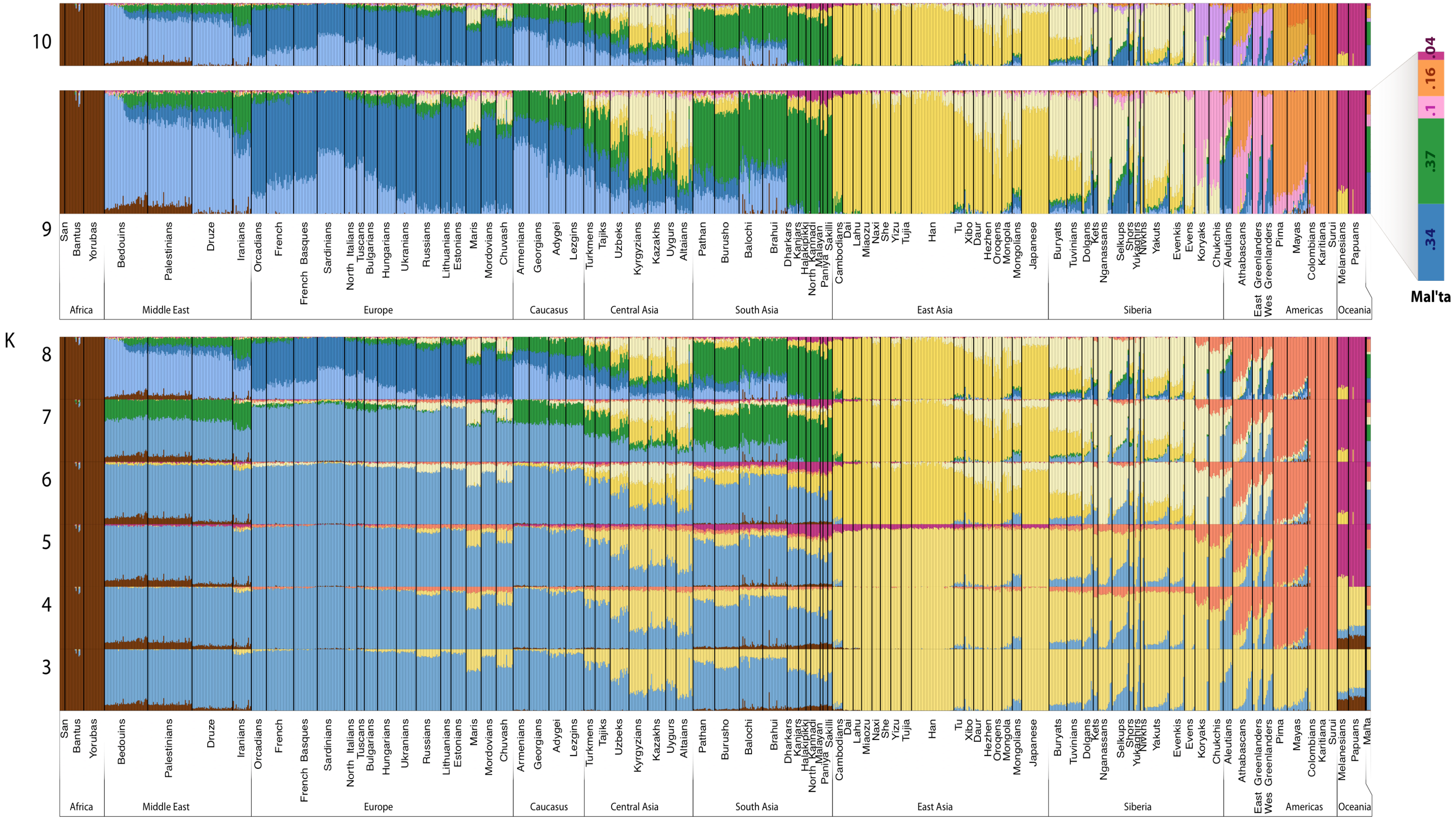

10
9
K
8
7
6
5
4
3
San
Bantus
Yorubas
Bedouins
Palestinians
Druze
Iranians
Orcadians
French
French Basques
Sardinians
North Italians
Tuscans
Bulgarians
Hungarians
Ukranians
Russians
Lithuanians
Estonians
Maris
Mordovians
Chuvash
Armenians
Georgians
Adygei
Lezgins
Turkmens
Tajiks
Uzbeks
Kyrgyzians
Kazakhs
Uygurs
Altaians
Pathan
Burusho
Balochi
Brahui
Dharkars
Kanjars
Halakipikki
North Kannadi
Malayan
Paniya
Sakilli
Cambodians
Dai
Lahu
Miaozu
Naxi
She
Yizu
Tujia
Han
Tu
Xibo
Daur
Hezhen
Oroqens
Mongola
Mongolians
Japanese
Buryats
Tuvinians
Dolgans
Kets
Nganassans
Selkups
Shors
Yukaghirs
Nivkhs
Yakuts
Evenkis
Evens
Koryaks
Chukchis
Aleutians
Athabascans
East Greenlanders
Wes Greenlanders
Pima
Mayas
Colombians
Karitiana
Surui
Melanesians
Papuans
Mal'ta
Africa
Middle East
Europe
Caucasus
Central Asia
South Asia
East Asia
Siberia
Americas
Oceania
.04
.16
.1
.37
.34
Mal'ta

Figure SI 6 Ancestry proportions of MA-1: ADMIXTURE (previous page). Plots representing ancestry components of MA-1 and a worldwide panel of present-day populations, showing converged runs from K=3 to K=15 in 100 replicates. K=9 is highlighted as this model was found to have the best predictive accuracy.

Table SI 11 List of 1,301 individuals used in the ADMIXTURE analysis. Populations and numbers of individuals (count) per population used in the analysis are shown. These individuals were genotyped on three different Illumina genotyping arrays (Human610-Quad, HumanHap650Y, Human660W-Quad). Populations are grouped according to the overall geographical categories in bold (Africa, America, Caucasus, etc.) and the references, including populations that were genotyped for this study (shaded).

Population	Count
Africa	**45**
Li et al. 2008	45
Bantus	19
San	5
Yorubas	21
America	**112**
Li et al. 2008	63
Colombians	7
Karitiana	13
Mayas	21
Pima	14
Surui	8
Rasmussen et al. 2010	49
Aleutians	9
Athabascans	20
East Greenlanders	10
West Greenlanders	10
Caucasus	**71**
Behar et al. 2010	38
Georgians	20
Lezgins	18
Li et al 2008	17
Adygei	17
Yunusbayev et al. 2011	16
Armenians	16
Central Asia	**101**
Behar et al 2010	34
Iranians	19
Uzbeks	15
Yunusbayev et al. 2011	26
Tajiks	15
Turkmens	11
This study	41
Kazakhs	18
Kyrgyzians	19
Uzbeks	4

Population	Count
East Asia	**214**
Li et al. 2008	214
Cambodians	10
Dai	10
Daur	9
Han	44
Hezhen	9
Japanese	28
Lahu	8
Miaozu	10
Mongola	10
Naxi	8
Oroqens	9
She	10
Tu	10
Tujia	10
Uygurs	10
Xibo	9
Yizu	10
Europe	**263**
Behar et al. 2010	46
Chuvash	17
Hungarians	19
Lithuanians	10
Li et al 2008	139
French	28
French Basques	24
North Italians	12
Orcadians	15
Russians	25
Sardinians	28
Tuscans	7
Yunusbayev et al. 2011	48
Bulgarians	13
Mordovians	15
Ukranians	20
This study	30
Estonians	15
Maris	15

Population	Count
Near East	**133**
Li et al. 2008	133
Bedouins	45
Druze	42
Palestinians	46
Oceania	**28**
Li et al. 2008	28
Melanesians	11
Papuans	17
Siberia	**197**
Li et al. 2008	25
Yakuts	25
Rasmussen et al. 2010	137
Altaians	13
Buryats	18
Chukchis	14
Dolgans	7
Evenkis	15
Kets	2
Koryaks	14
Mongolians	9
Nganassans	14
Selkups	10
Tuvinians	15
Yukaghirs	6
Dolgans	3
Evens	8
Kets	2
Nivkhs	3
Shors	4
Yakuts	1
This study	14
Altaians	3
Dolgans	1
Evens	2
Russians	1
Selkups	7
South Asia	**137**
Behar et al. 2010	19
Malayan	2
North Kannadi	9
Paniya	4
Sakilli	4
Li et al. 2008	96
Balochi	24
Brahui	25
Burusho	25
Pathan	22
Metspalu et al. 2011	22
Dharkars	11
Halakipikki	4
Kanjars	7
GRAND TOTAL	**1301**

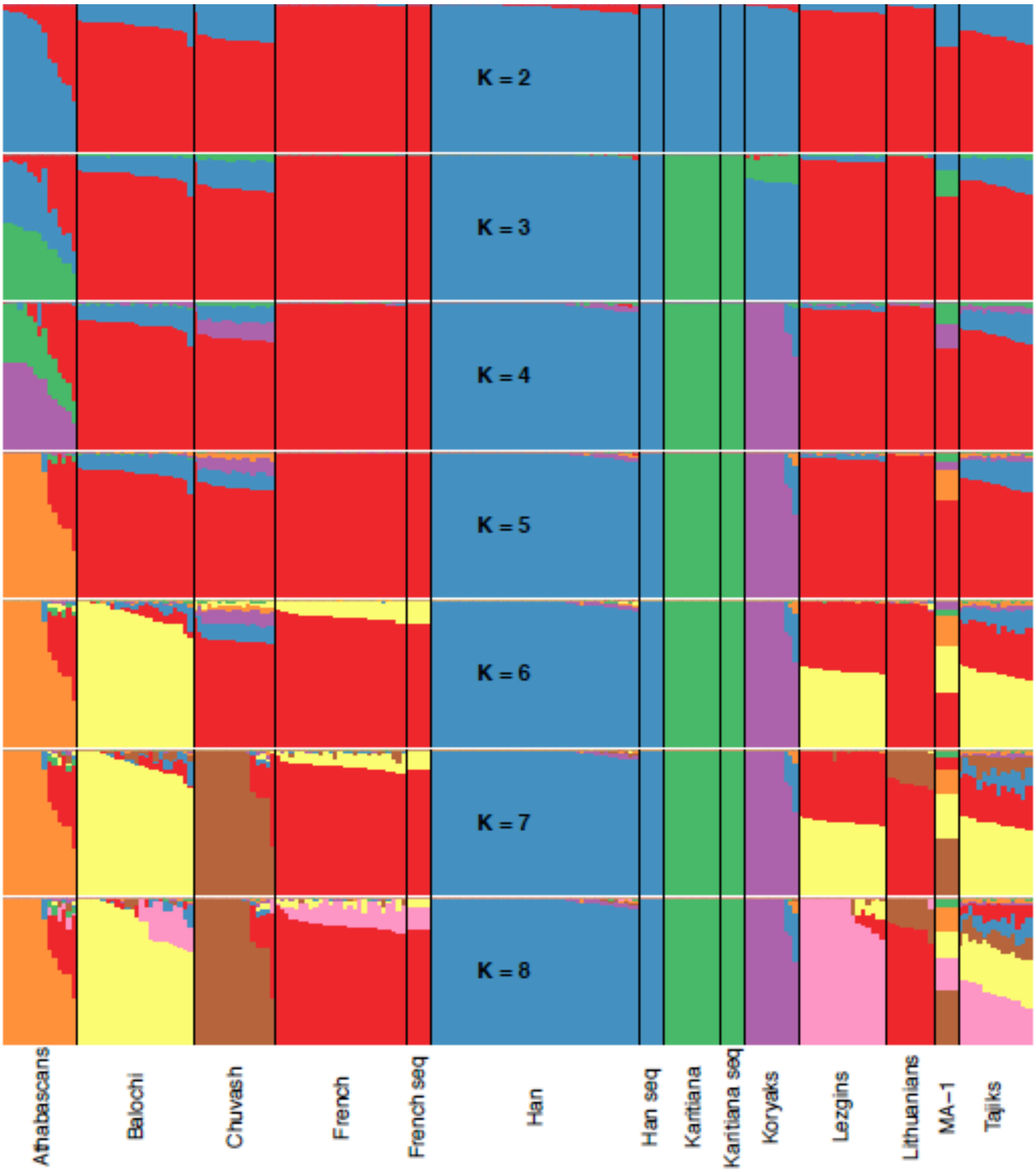


Figure SI 7 Ancestry proportions of MA-1: NGSadmix. Plots representing ancestry components of MA-1 and 10 present-day populations, showing converged runs from K=3 to K=8.

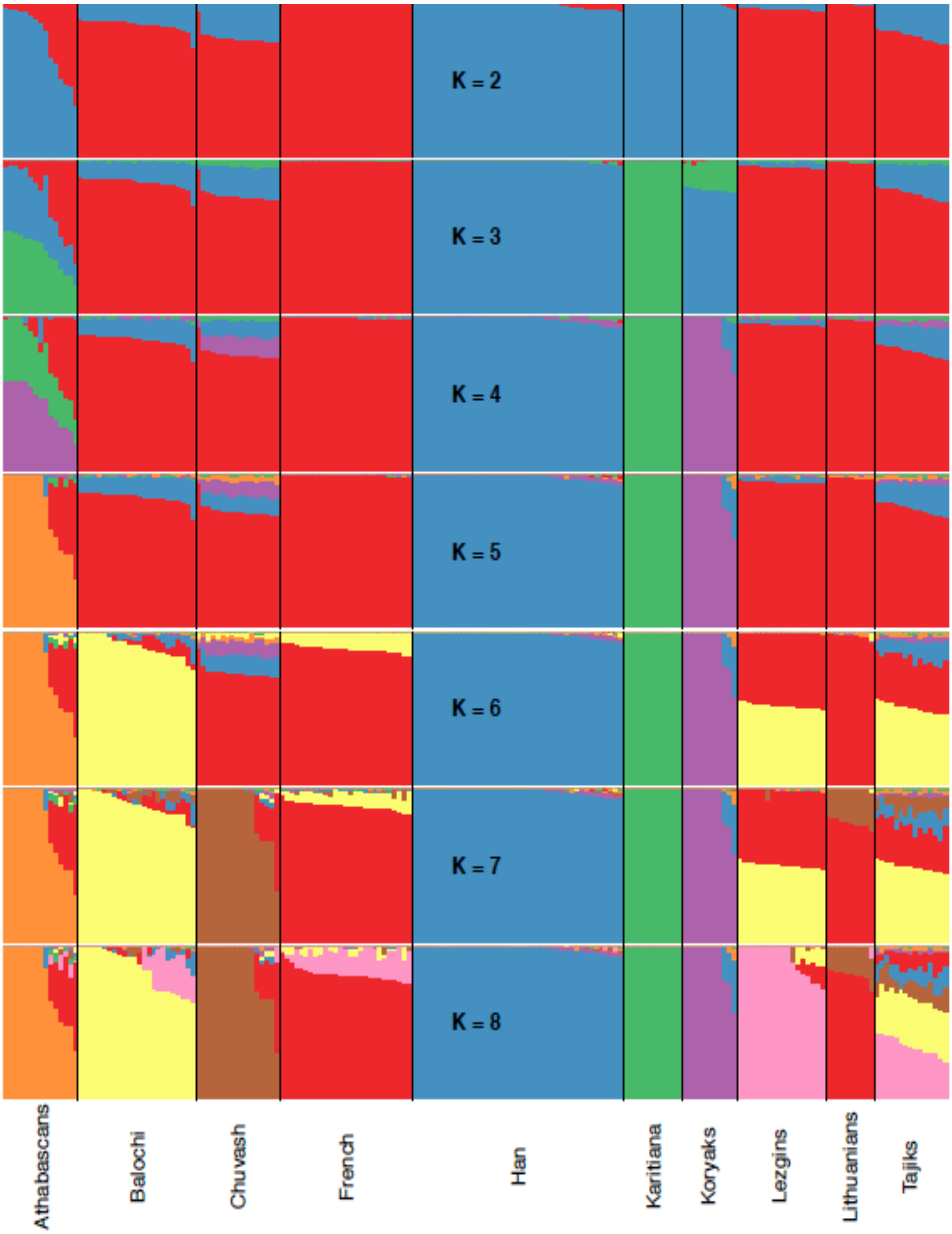


Figure SI 8 SNP chip dataset: ADMIXTURE. Results from the SNP chip dataset only (excluding the three modern genomes and MA-1), showing converged runs from K=2 to K=8.

SI 10. Principal Component Analysis

A single read was sampled from each position in the MA-1 dataset, which overlapped with SNPs in a dataset compiled from Reich *et al*. (2012)[1] where the authors had used local ancestry inference to mask segments of European and African ancestry in Siberian and Native American populations[2-5]. A phred-scaled mapping quality of 30 and base quality score of 30 was required in the sequence data for a haploid genotype to be called, and sequences with indels in their alignment to the reference genome were not considered. SNPs with minor allele frequency of <1% were removed. To reduce the effect of nucleotide misincorporations, the first and last three bases of each sequence read in the MA-1 data were excluded. SNPs where there was no information from MA-1 were excluded, and a single haploid genotype was randomly sampled from each modern individual to match the single-pass nature of the shotgun data[6]. Principal component analysis (PCA) was performed on various population subsets separately using EIGENSOFT 4.0[7], removing one SNP from each pair for which linkage disequilibrium exceeded a low arbitrary threshold ($r^2 > 0.2$). Transition SNPs where the ancient individual displayed a T or an A[8], as well as triallelic SNPs, were excluded. This strategy is valid when only a single individual with potential *post-mortem* damage is included in the PCA, whereas if multiple such individuals are included all transition SNPs should be excluded. Projection of Eurasian and Native American populations on the first two PCs can be seen in Figure 1b. Projection on PC3 and PC4 can be seen in Figure SI 9, in which an affinity between MA-1 and Central and South Asians and Oceanians can be observed, possibly reflecting shared ancestral variation.

To investigate the Native American affinity of MA-1 in the PCA, we performed a two-step analysis[9] where the PCA was first computed on MA-1, a Sardinian (HGDP00667) (haploidized), and a Han individual (HGDP00973) (haploidized). In the second step, all other individuals were projected on the PCs defined by the first three. We find that the configuration of modern populations recapitulate the major patterns in the complete PCA, with Native American and Siberian populations showing clear affinity to MA-1. This suggests that the 24,000-year-old Siberian is representative of key genetic components of Native American ancestry today (Figure SI 10).

References for SI 10

1. Reich, D. *et al.* Reconstructing Native American population history. *Nature* **488**, 370-374 (2012)

2. Hancock, A.M. *et al.* Adaptations to climate-mediated selective pressures in humans. *PLoS Genet.* **7**, e1001375 (2011)

3. Rasmussen, M. *et al*. Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463**, 757-762 (2010)

4. International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-58 (2010)

5. Li, J.Z. *et al*. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100-1104 (2008)

6. Skoglund, P., Jakobsson, M. Archaic human ancestry in East Asia. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 18301-18306 (2011)

7. Patterson, N., Price, A.L., Reich, D. Population Structure and Eigenanalysis. *PLoS Genet.* **2**, e190 (2006)

8. Skoglund, P. *et al*. Origins and Genetic Legacy of Neolithic Farmers and Hunter-Gatherers in Europe. *Science* **336**, 466-469 (2012)

9. Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053-1060 (2010)

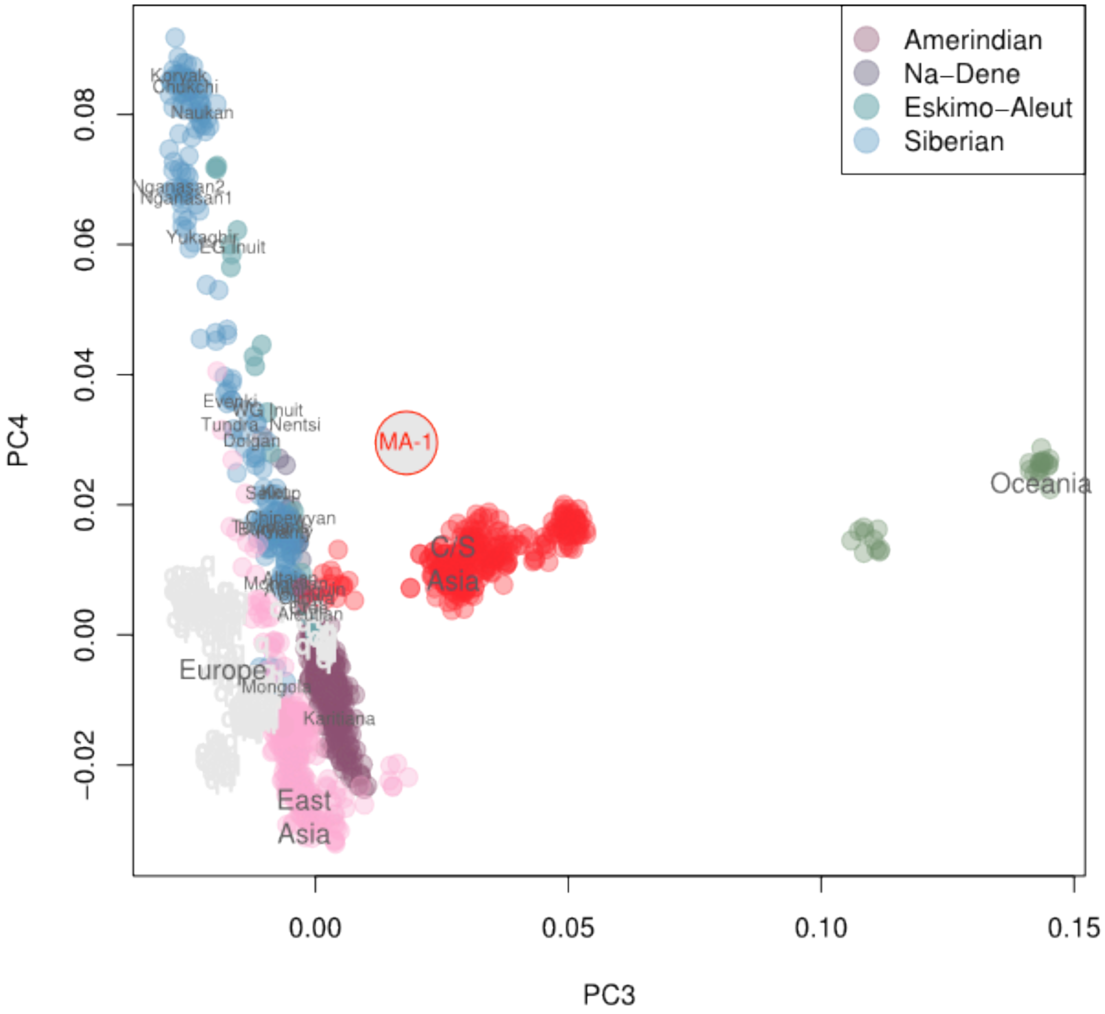


Figure SI 9 PCA on MA-1 using a worldwide SNP panel masked for European and African ancestries[1]. The third and fourth principal components in Eurasian and American genetic variation reveals shared ancestral variation between MA-1 and, Central and South Asians and Oceanians.

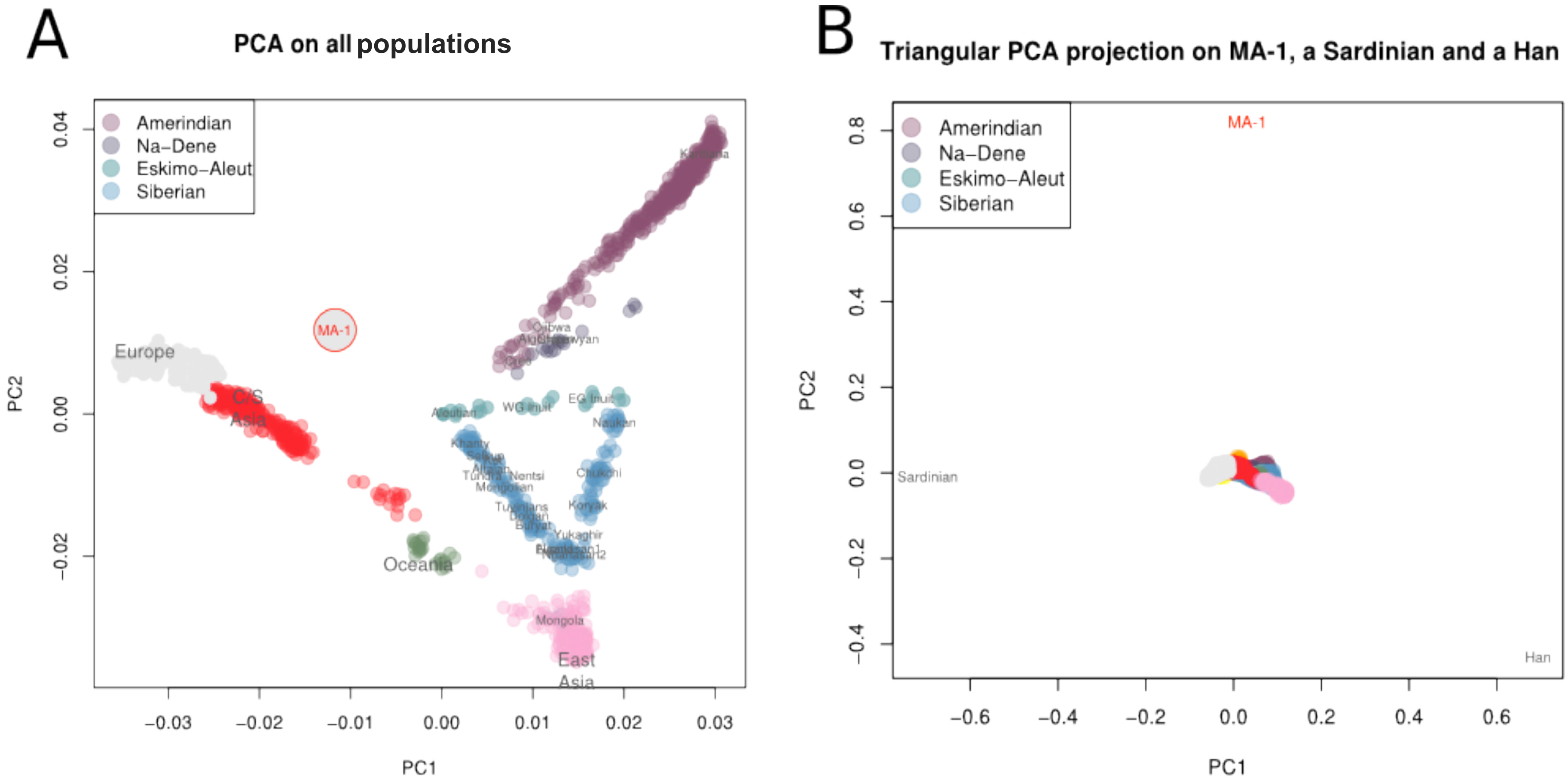


Figure SI 10 PCA projection. **A)** PC1 and PC2 for Eurasian populations (reproduces Figure 1b in the main text). **B)** PCA projection of Eurasian populations on MA-1 and one individual from each longitudinal extreme of Eurasia, closely resembles a full PCA analysis on the same populations.

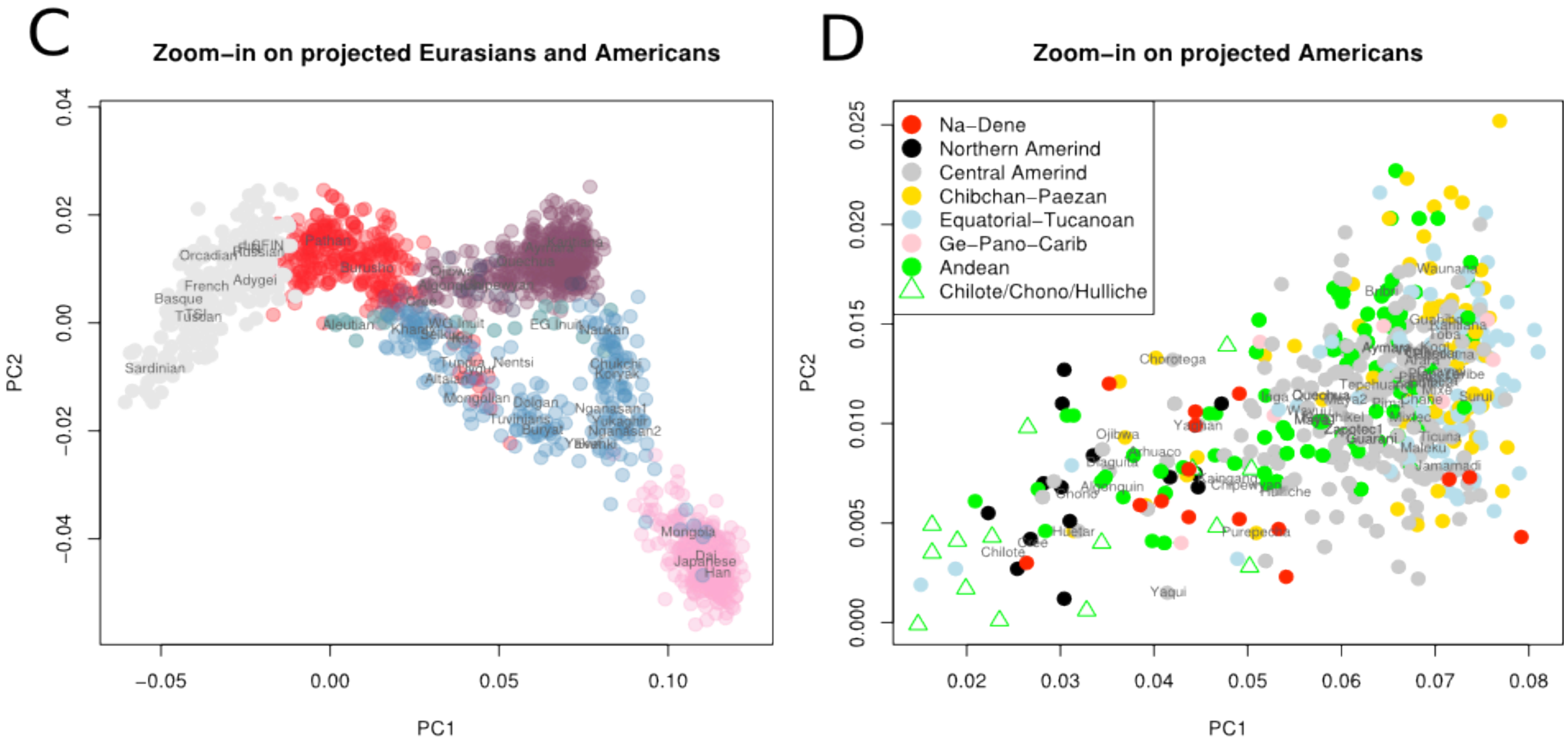


Figure SI 10 PCA projection. PCA projection of Eurasian populations on MA-1 and one individual from each longitudinal extreme of Eurasia closely resembles a full PCA analysis on the same populations. **C)** Zoom in on the PCA projection on Eurasian and American populations. **D)** Zoom-in on the PCA projection of Native American individuals

SI 11. *TreeMix* analysis

SI 11.1 Data filtering

For inferring admixture graphs a total of 17 individuals were used, consisting of the archaic Denisova genome[1], 11 present-day individuals[1], the 4 novel genomes from this study (SI 4.2), and the MA-1 genome. Haploid genotypes from MA-1 were added to variants identified in the other individuals as in SI 10 to alleviate the increased rate of errors in low-coverage ancient DNA sequence data. If multiple sequence reads overlapped a position, one read was randomly sampled[2]. This avoids biasing for, or against, heterozygotes and renders the MA-1 data haploid. All transition SNPs were excluded and, MA-1 sequence reads with a mapping quality less than 30 and bases with base quality less than 30 were discarded. Positions where there was no data from one of the individuals in the analysis were also excluded. This resulted in a final count of 156,250 SNPs for the main analysis. In the data set with the Karitiana individual excluded, a total of 158,539 SNPs were obtained.

SI 11.2 Admixture graph fitting

TreeMix[3] (version 1.12) was used to build ancestry graphs assuming 0 to 10 migration edges, the placement and weight of each being optimized by the algorithm. *TreeMix* was run using the –global option which corresponds to performing a round of global rearrangements of the graph after initial fitting. The sample size correction was also disabled, since all the populations consisted of single individuals (-noss). Standard errors were estimated in blocks with 500 SNPs in each. For those analyses that included one or more *a priori* specified events, a round of optimization was performed on the original migration edge (option -climb).

SI 11.3 Results

Sequentially allowing for between 0 and 10 gene flow events, the increase in likelihood beyond four such events was found to be marginal (Figure SI 11), and at this point no residual exceeded 4 standard errors (Figure SI 12, Figure SI 13). The second of the sequentially inferred events modeled the Karitiana individual as deriving 36.0±17.3% of its ancestry from the lineage leading to MA-1, which would correspond to the shared ancestry between western Eurasians and Native Americans that is detected in other analyses. Several previously established admixture events were also obtained, such as from Denisovans into the Papuan[4], and from western Eurasians into the Indian genome[5]. Additionally, 14.9±2.6% Native American-related ancestry (possibly Northeast Asian) was estimated in Mari and Tajik individuals in this optimized graph.

While the optimized admixture graphs above modeled Karitiana as having MA-1-related ancestry, it is possible that the affinity between the two could also be explained with gene flow in the other direction. To test the statistical support for the direction of gene flow, a bootstrap analysis was performed where an *a priori* migration edge from the Denisova lineage to the ancestry of the Papuan was included, and *TreeMix* was then used to optimize a second migration edge. Over 100 bootstrap pseudoreplicates, a migration edge from MA-1 to Karitiana was observed in 99 cases, and a migration edge from Karitiana to MA-1 in 1 case. This analysis was repeated including all three migration edges inferred for *m*=4 in the first analysis, except the one including MA-1 and Karitiana. Here, all 100 bootstrap pseudoreplicates supported the gene flow direction from MA-1 into Karitiana.

SI 11.4 Admixture graphs for subsets of the 17 genomes

As an additional test to assess whether the signal of gene flow between eastern Eurasians and the MA-1 individual was also detectable with no Native American individual included, the Karitiana genome was excluded and the above analysis was repeated. If the gene flow was mainly from Native American ancestors to the MA-1 lineage, an affinity would also be expected between East Asians and MA-1. In the maximum-likelihood tree (Figure SI 14), there is no strong residual indicating greater covariance between MA-1 and East Asians than predicted by the model. However, a migration edge from Dai to MA-1 was observed when a total of 4 migration edges are inferred in the graph (Figure SI 15), in line with other indications of basal Asian gene flow into MA-1 (SI 14).

Finally, the admixture graph fitting was also repeated using only the complete genomes sequenced by Meyer *et al*. (2012)[1] and MA-1. Here, ~10 million SNPs were obtained and so blocks of 2000 SNPs were used to compute standard errors. The results largely concur with the main analysis, except for in the direction of gene flow between MA-1 and Karitiana (Figure SI 16). However, this analysis lacks representation from northern Europe and Central Asia (such as the 4 genomes sequenced in this study), which affects the general affinity between Karitiana and western Eurasians that we document in SI 14.

SI 11.5 Admixture graphs for simulated data

We wanted to address the curious finding that *TreeMix* identifies an admixture event between Denisova and Papuans before it identifies one between MA-1 and Karitiana, even though the inferred level of shared genetic ancestry is much greater for MA-1 and Karitiana than it is for Denisova and Papuans. Data was generated for five diploid individuals (i.e. 10 chromosomes) and one was sampled (i.e., two chromosomes) from each of five populations, using the coalescent simulator *ms*[6]. The demographic history of these five populations is indicated in Figure SI 17A, with parameters set at *N*=10,000 diploid individuals and a generation time of 25 years. At 30 kya, population 2 receives 7.8% of its genetic material from population 1 and at 5 kya, population 4 receives 41% of its genetic material from population 3. The more ancient admixture event (7.8% mixture) represents a mixture between a pair of populations that are distant evolutionarily, whereas the more recent admixture event (41% mixture) represents a mixture between a pair of populations that are close evolutionarily. Under this demographic history, one million coalescent simulations were performed to obtain one million independent SNPs for input into *TreeMix*. *TreeMix* was applied to the simulated dataset, assuming population 1 as the root and either zero, one, or two migration events.

Figures SI 17B-D display the admixture graph relating the five genomes (populations), with zero, one, or two migration events, respectively. Figure SI 17C indicates that the first admixture event to be identified is that between highly diverged populations 1 and 2. This result is interesting as the identified admixture event was the more ancient of the two events, and it was also substantially smaller in magnitude than the other event. This result is akin to our empirical situation for which we identify Denisova admixture into Papuans before identifying MA-1 mixture with Karitiana, even though the inferred Denisova admixture with Papuans is substantially smaller in magnitude to that inferred between MA-1 and Karitiana (Figure 2, Figures SI 13, 15). Therefore, the first admixture event identified by *TreeMix* may not be the one with the greatest magnitude, but instead may be an event with low magnitude between evolutionarily distant populations.

References for SI 11

1. Meyer, M. *et al.* A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science* **338**, 222-226 (2012)

2. Green, R.E. *et al*. A Draft Sequence of the Neandertal Genome. *Science* **328**, 710-722 (2010)

3. Pickrell, J.K., Pritchard, J.K. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genet*. **8**, e1002967 (2012)

4. Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053-1060 (2010)

5. Reich, D. *et al.* Reconstructing Indian population history. *Nature* **461**, 489-494 (2009)

6. Hudson, RR. Generating samples under a Wright-Fisher neutral model. *Bioinformatics* **18**, 337-338 (2002)

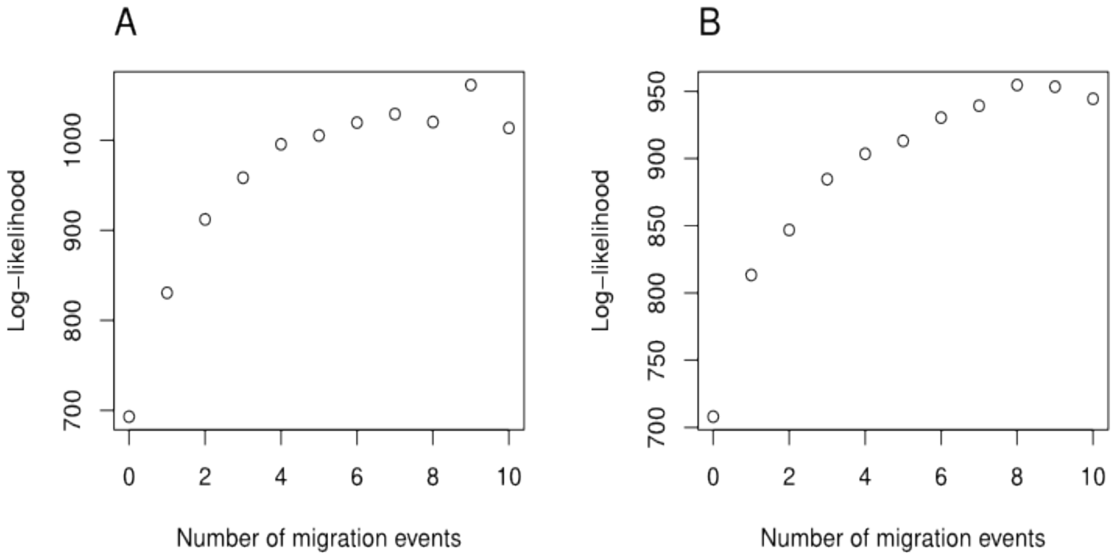


Figure SI 11 Increase in log-likelihood as a function of the number of migration events in *TreeMix*. A) Full analysis using 17 genomes. **B)** Karitiana individual excluded.

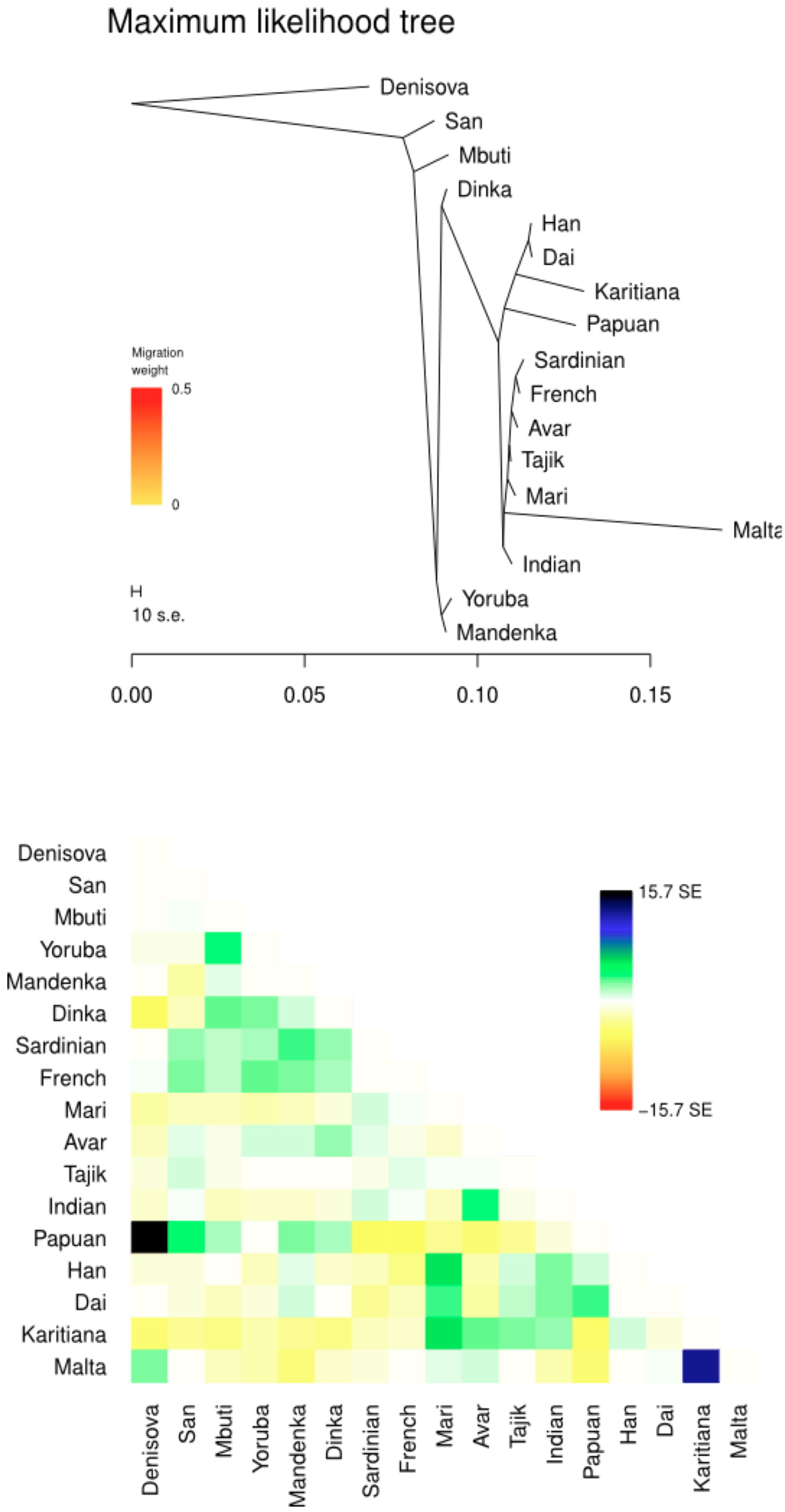
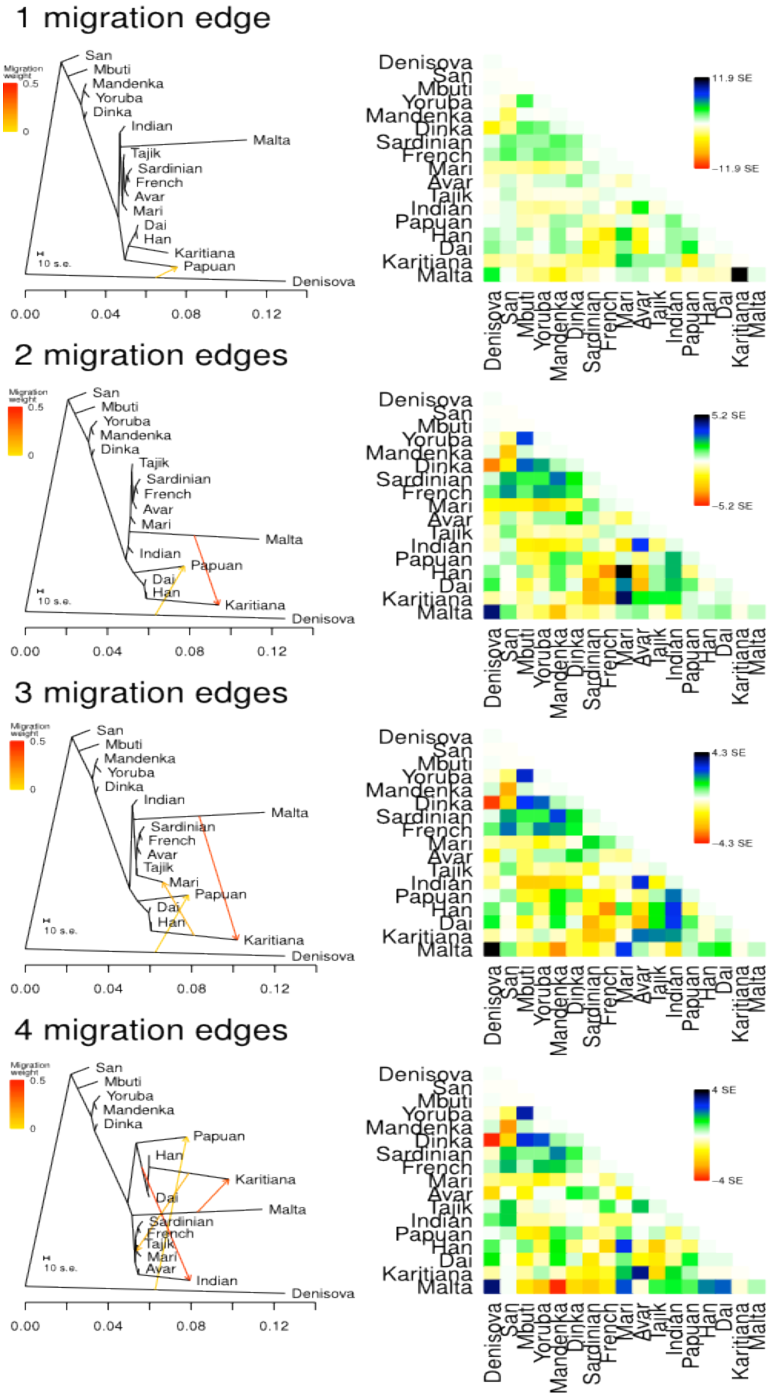


Figure SI 12 Maximum likelihood tree for 17 genomes. Population tree reconstructed using *TreeMix* with the MA-1 genome (Mal'ta), Denisova genome, 11 present-day individuals from Reich *et al.* (2012)[1] and the four genomes produced in this study.



1 migration edge

2 migration edges

3 migration edges

4 migration edges

Figure SI 13 Admixture graphs reconstructed using *TreeMix* and 17 genomes.
Results are presented for 1 to 4 migration events for 16 genomes and MA-1 (Mal'ta).

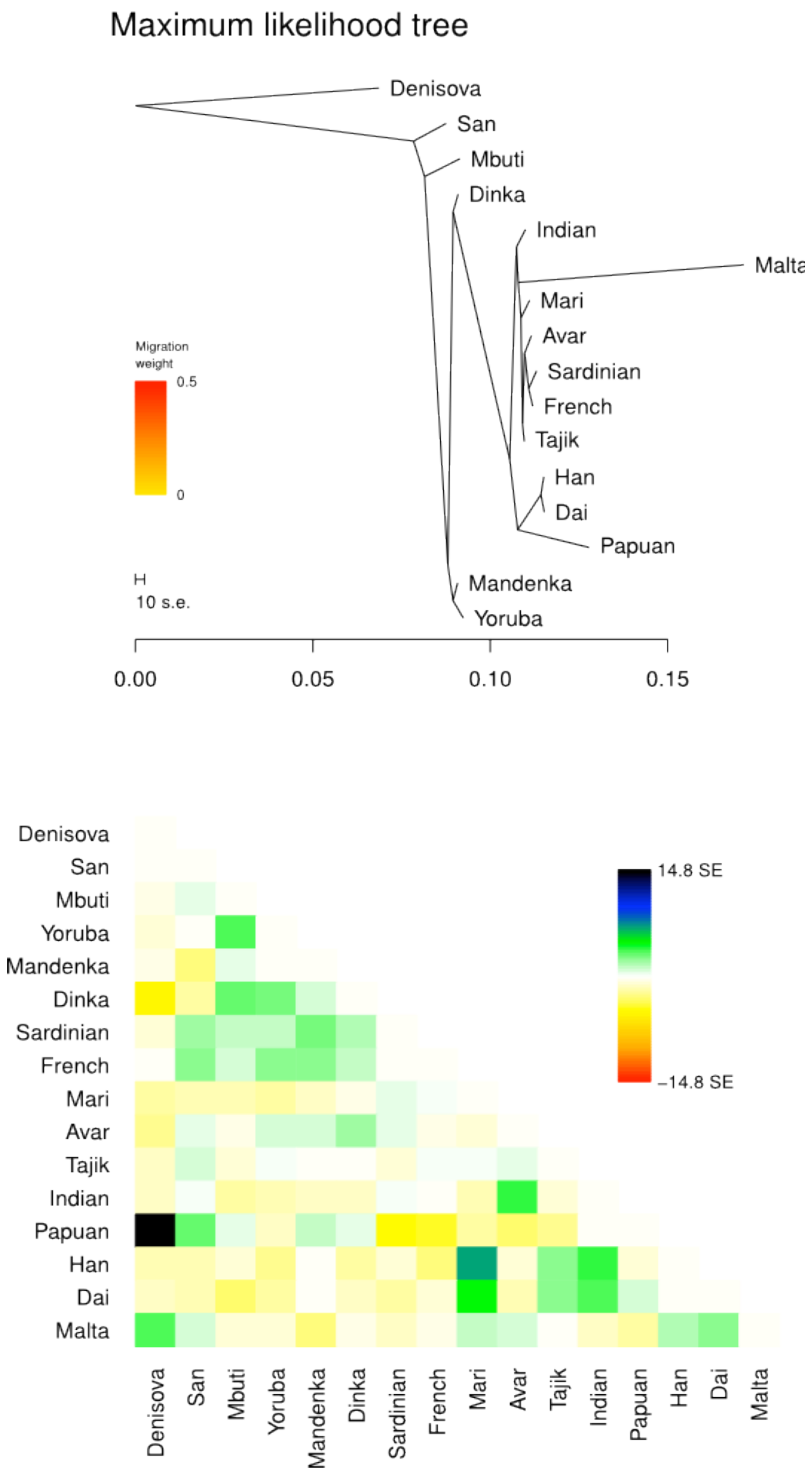


Figure SI 14 Maximum likelihood tree with no Native American. Population tree reconstructed using *TreeMix* with 15 genomes and MA-1 (Mal'ta), the Karitiana individual being excluded from this analysis.

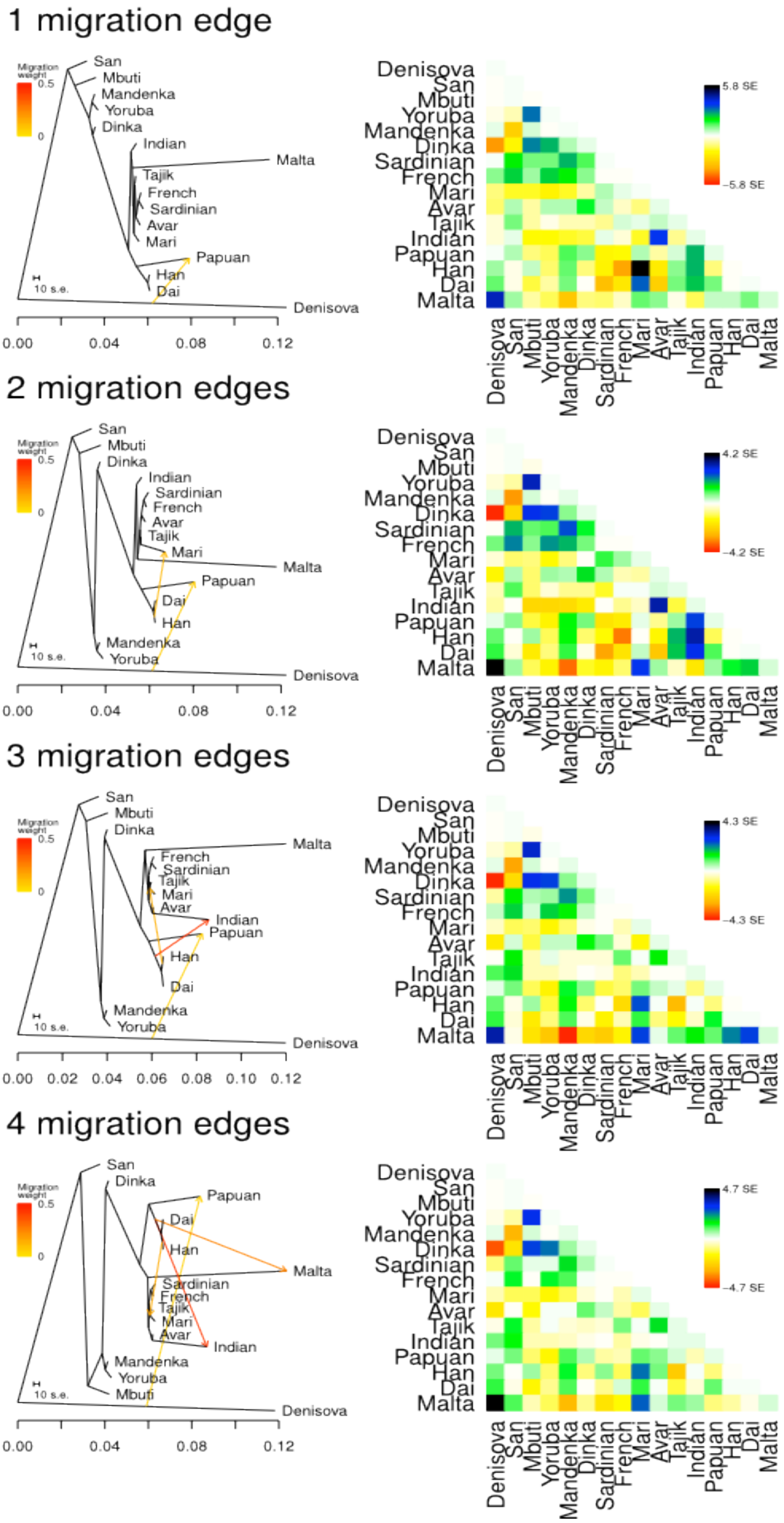


Figure SI 15 Admixture graphs with no Native American. Admixture graphs reconstructed for 1 to 4 migration events for 15 genomes and MA-1 (Mal'ta), the Karitiana individual being excluded from this analysis.

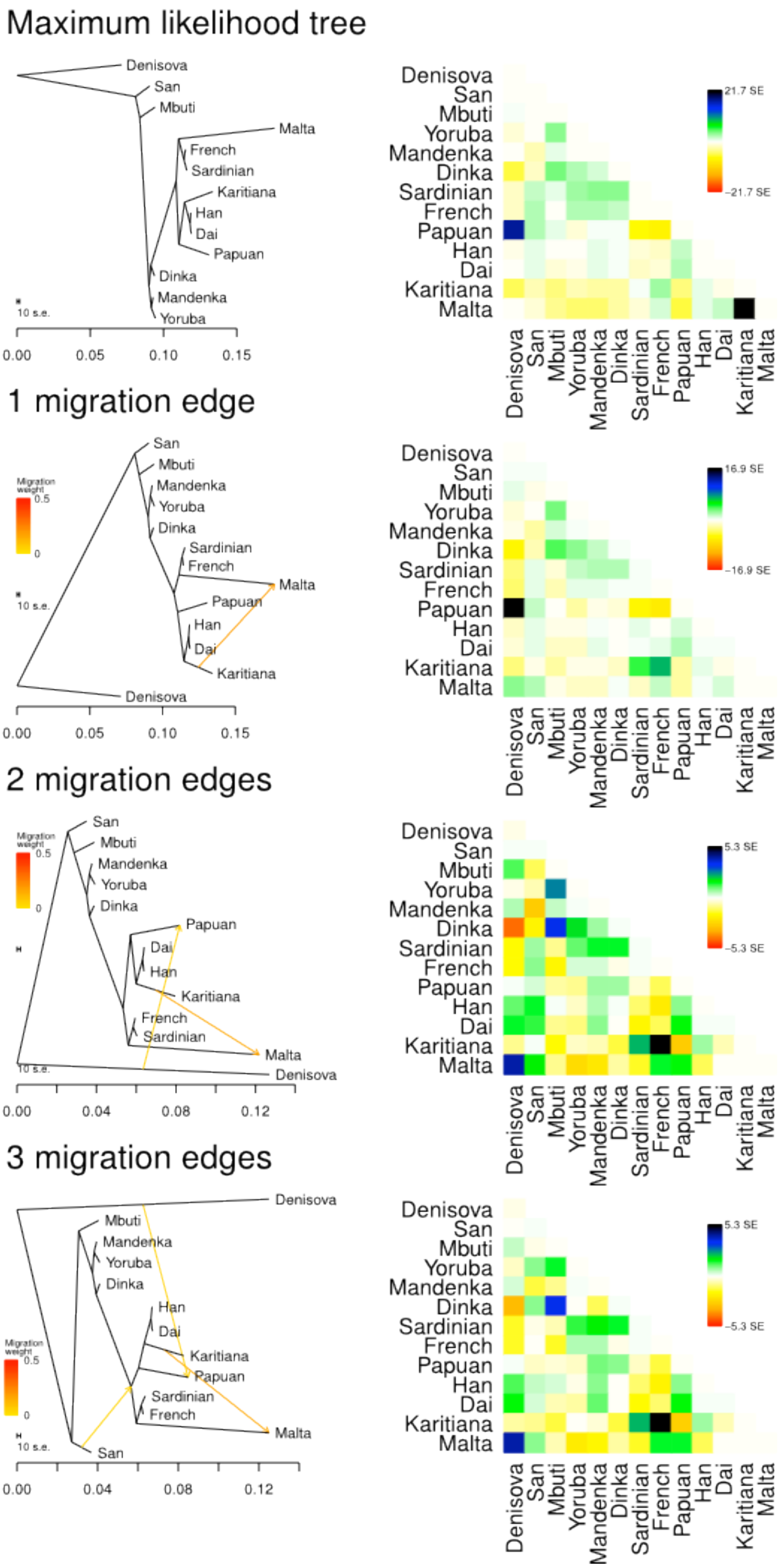


Figure SI 16 *TreeMix* analysis of MA-1 (Mal'ta) and the 12 genomes published by Meyer *et al*. (2012)[1]. Graphs for 0 to 3 migration events are shown.

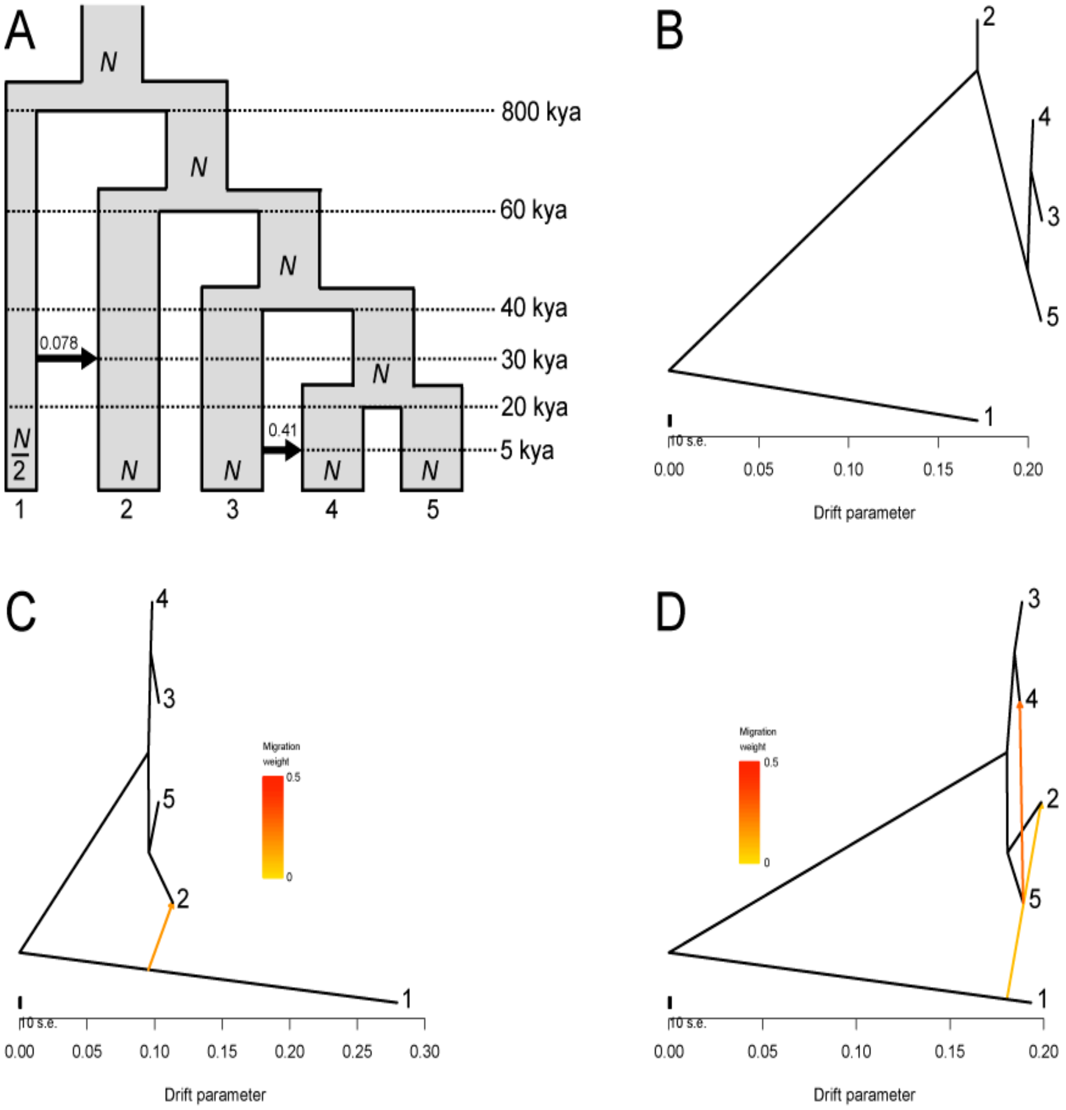


Figure SI 17 ***TreeMix*** **analysis of five simulated genomes sampled from five populations**. (A) Demographic model indicating 7.8% admixture of population 1 into population 2 and 41% admixture of population 3 into population 4. (B-D) Zero, one, and two migration events, respectively.

SI 12. *MixMapper* analysis

Admixture graphs were constructed relating MA-1 to modern groups using *MixMapper* v1.0[1], a recently developed approach that utilizes f_2 distances between populations to fit admixture graphs of population history. *MixMapper* is based on an approach where a scaffold tree is constructed from populations that are assumed to be unadmixed, upon which other populations can then be fitted as either simple branches off the scaffold, two-way admixtures between populations in the scaffold tree, or alternatively three-way admixtures between an unadmixed population, and a two-way admixed population.

In order to place the MA-1 genome in the context of modern day variation, a scaffold tree was constructed using 4 African genomes (San, Yoruba, Mandenka, Dinka), Sardinian and Han[3]. The Mbuti genome was excluded for the main analyses since the Mbuti are known to have substantial amounts of recent Bantu-speaker-related ancestry, and are difficult to fit using admixture graphs[1,2], but similar results were obtained by us when the Mbuti was also included (see below). Sardinian and Han were chosen as representatives of the western and eastern extremes of modern-day Eurasian genetic differentiation. All transitions were excluded, and standard errors of the *f*-statistics were estimated using 500 bootstrap replicates over 50 blocks of the autosomal genome.

When adding the Karitiana genome to this scaffold tree, 422 of 500 bootstrap pseudoreplicates were found to fit Karitiana as a mix between the Sardinian and Han lineages, with 73.7% (95% CI: 61.8-85.7%) of its ancestry being derived from the Han lineage and the remainder 26.3% (14.3-38.2%) from the Sardinian lineage (Figure SI 18; Table SI 12). When the Mbuti was included, the bootstrap support was 432 of 500 and the Sardinian-related ancestry in MA-1 was 13.4-35.4%. Fitting MA-1 to the scaffold with Sardinian, the best fit (496 of 500 bootstraps) suggested that MA-1 is a mixture of the Sardinian and Han lineages, with 22.1-75.0% of its ancestry being derived from the Han lineage (Figure SI 18; Table SI 12).

Subsequently, Karitiana was fitted as a three-way mixture between MA-1 (modeled as a mixture as above, see SI 14) and Han, since it is likely that populations related to MA-1 mediated the mixture event between western Eurasians and Native Americans. Here, 496 bootstrap replicates supported Karitiana as having 26.1% (7.7-44.4%) ancestry from the MA-1 lineage and the remainder from Han, consistent with the previous analysis.

A previous study[1] used *MixMapper* and data ascertained in a San individual, and subsequently genotyped in a large number of human populations, to fit Sardinians as a mixture of a western Eurasian lineage and Karitiana, which would suggest that it is inappropriate to use Sardinians in the scaffold tree. If the Sardinian is replaced with Karitiana in the scaffold tree, the highest number of bootstrap pseudoreplicates for any one combination (231 of 500) fitted Sardinians as a mix between Dinka and Karitiana (with 27.7-33.4% ancestry from the Dinka lineage). Lipson *et al*. (2013)[1] chose the set of populations to be included in the scaffold based on observed f_3 tests and additivity of f_2-distances in the scaffold tree. However, as discussed later on in SI 14, the reason for Native Americans not displaying negative f_3-statstics when tested for having dual ancestry from eastern and western Eurasians could be the extensive amount of genetic drift that has occurred since their divergence. We found that the scaffold tree with Sardinian was more additive than the scaffold tree with Karitiana, suggesting that Karitiana are better fitted as admixed than Sardinians are (Figure SI 19). This was also consistent when Han was replaced by Dai in the scaffold tree and when both East Asians were included (Figure SI 19). It was also consistent when Mbuti was included in the scaffold tree for either (or both) East Asian genomes.

Sardinians also do not display any significant f_3-tests, or evidence of admixture linkage disequilibrium with Native Americans[4,5]. However, we note that if Sardinians themselves have admixture from a Native American source as suggested by Lipson *et al*. (2013)[1], or some other type of complex ancestry (they have evidence of 0.6±0.2% African ancestry[5,6] but that is likely negligible for our estimates), the estimates of western Eurasian ancestry in Karitiana are likely to be affected. Current admixture graph models are not suitable for dealing with possible bidirectional gene flow, and so a future line of study would be to simultaneously infer gene flow between genomic samples from Native American and western Eurasian populations.

References for SI 12

1. Lipson, M. *et al*. Efficient moment-based inference of admixture parameters and sources of gene flow. *Mol. Biol. Evol.* Doi: 10.1093/molbev/mst099 (2013)

2. Pickrell, J.K., *et al*. The genetic prehistory of southern Africa. *Nat. Commun*. **3**, 1143. doi: 10.1038/ncomms2140 (2013)

3. Meyer, M. *et al.* A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science* **338**, 222-226 (2012)

4. Patterson, N. *et al*. Ancient Admixture in Human History. *Genetics* **192**, 1065-1093 (2012)

5. Loh *et al*. Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* **193**, 1233-1254 (2013)

6. Moorjani *et al*. The History of African Gene Flow into Southern European, Levantines, and Jews. *PLoS Genet.* **7**, e1001373 (2011)

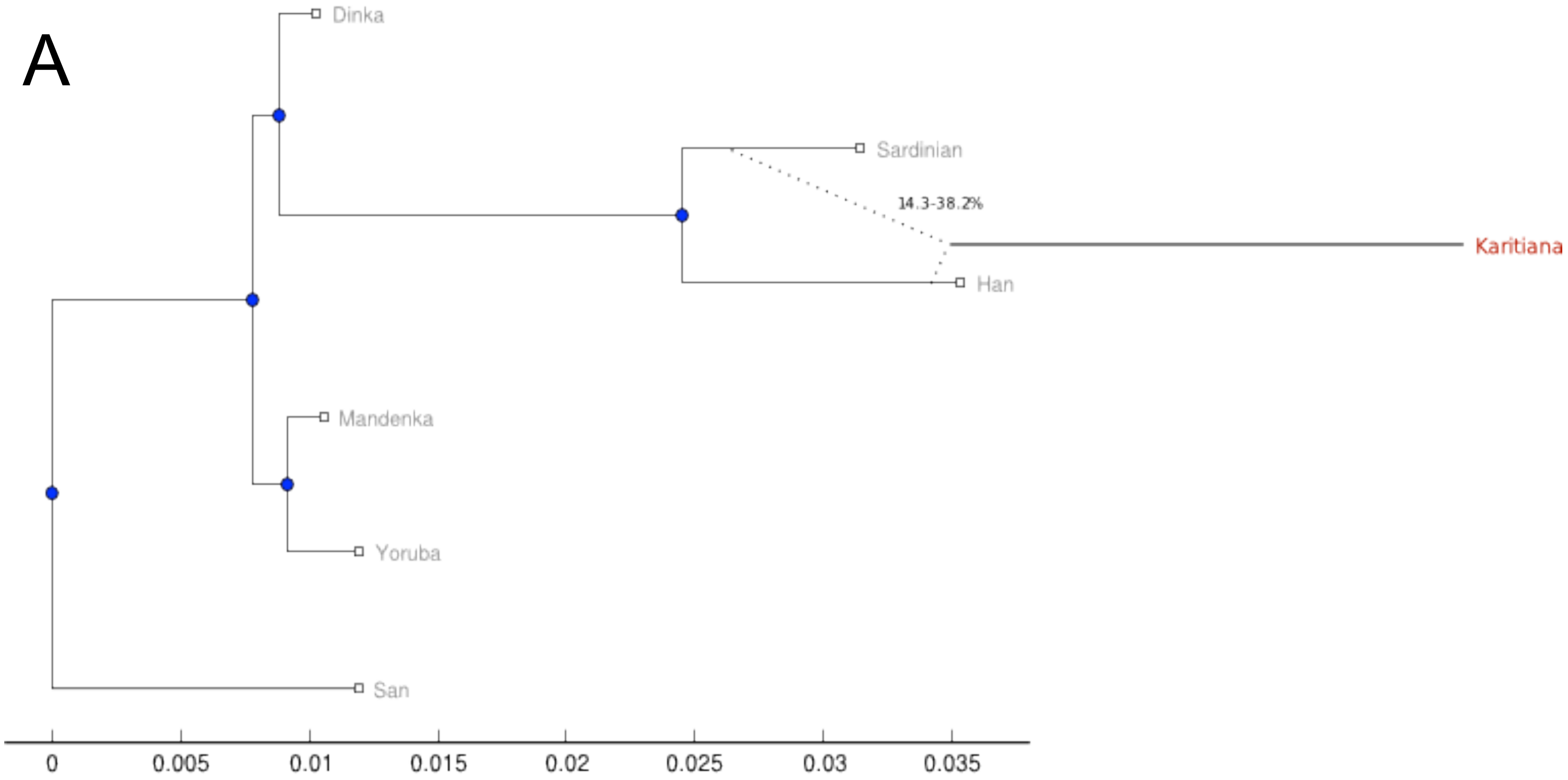


Figure SI 18 Admixture graphs fitted using *MixMapper* by first optimizing a scaffold tree consisting of San, Yoruba, Mandenka, Dinka, Sardinian and Han. In **A)**, Karitiana was fitted onto the scaffold allowing for dual ancestry.

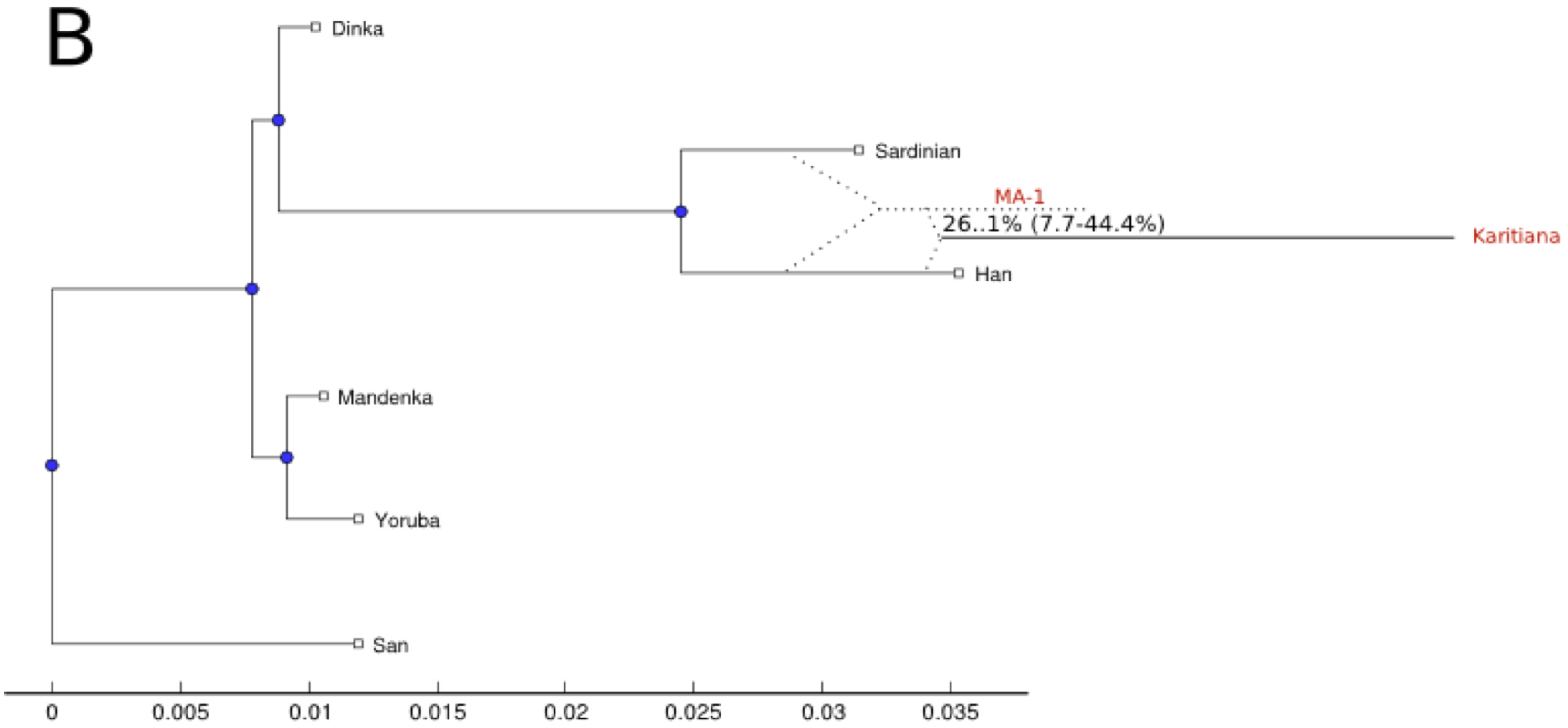


Figure SI 18 Admixture graphs fitted using *MixMapper* by first optimizing a scaffold tree consisting of San, Yoruba, Mandenka, Dinka, Sardinian and Han. In **B)**, MA-1 was first fitted onto the scaffold allowing for dual ancestry, and subsequently Karitiana was fitted to the resulting graph allowing for gene flow from MA-1 and an optional lineage.

Table SI 12 Results of fitting genomes as putative two-way admixtures on a scaffold tree consisting of San, Yoruba, Mandenka, Dinka, Sardinian and Han. The scaffold tree can be seen in Figure SI 16. # Denotes the number of bootstrap replicates of the total 500 in the combination of branches that was best supported. Only the best supported combination is shown. 'Resnorm' is the residual between the distances in the fitted graph compared to the empirically observed f_2-distances. The α-parameter is the mixture proportion from Branch 1 and we show the 95% confidence interval. Branch1Loc and Branch2Loc are the points in the respective branches where the admixing lineage is inferred to branch off. MixedDrift is the amount of genetic drift inferred for the admixed lineage. Note that mixtures where only a minor fraction is from one source such as the Dai and the French should likely be interpreted as false positives for mixture.

AdmixedPop	Branch1 + Branch2	#	Resnorm	α	Branch1Loc	Branch2Loc	MixedDrift
Karitiana	Han+Sardinian	422	9.31e-07	0.618-0.857	0.008-0.011 / 0.011	0.000-0.006 / 0.007	0.020-0.024
MA-1	Han+Sardinian	496	3.98e-06	0.221-0.750	0.001-0.009 / 0.011	0.002-0.007 / 0.007	0.064-0.067
Dai	Han+San	202	6.78e-07	0.964-0.998	0.009-0.011 / 0.011	0.000-0.012 / 0.012	0.000-0.002
French	San+Sardinian	248	3.2e-07	0.005-0.058	0.001-0.012 / 0.012	0.004-0.007 / 0.007	0.001-0.003
Mari	Han+Sardinian	500	1.26e-06	0.236-0.479	0.006-0.011 / 0.011	0.004-0.007 / 0.007	0.002--0.005
Indian	Han+Sardinian	500	2.32e-06	0.223-0.730	0.002-0.010 / 0.011	0.002-0.006 / 0.007	0.002-0.004
Avar	Han+Sardinian	311	1.16e-06	0.146-0.419	0.000-0.004 / 0.011	0.004-0.007 / 0.007	0.001-0.004
Papuan	Han+San	405	1.94e-06	0.885-0.953	0.004-0.005 / 0.011	0.000-0.012 / 0.012	0.020-0.023

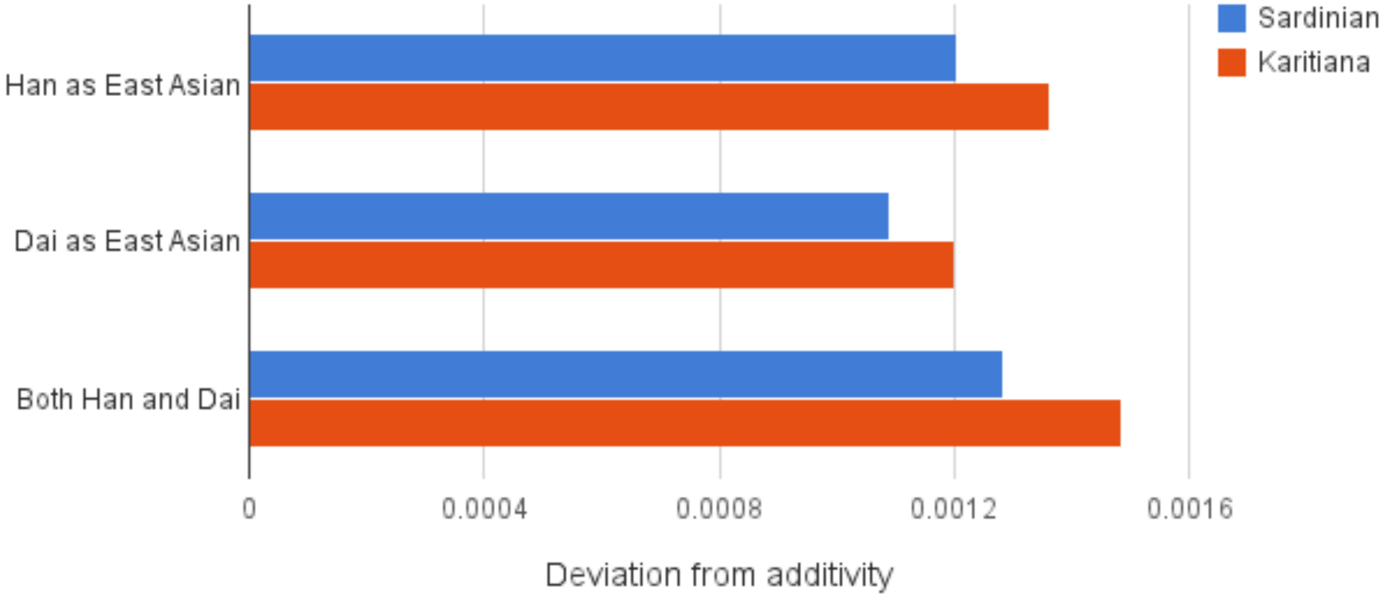


Figure SI 19 Residuals between the empirical f_2-distances and the leaf-to-leaf distances in the inferred scaffold tree when Sardinian or Karitiana are included. Scaffold trees inferred with MixMapper on complete genomes are more additive when Sardinian is included rather than when Karitiana is included. The other populations in the scaffold tree were San, Yoruba, Dinka, and one or both of Han and Dai. Qualitatively similar results were obtained when Mbuti was also included.

13. *D*-statistic tests based on sequencing data (D_{seq})

SI 13.1 Notation and description of the test

To investigate the relationship between MA-1 and a number of modern populations, a *D*-statistic test, equivalent to previously published tests[1,2], was applied to sequencing data from a single genome from each of the populations of interest. This test is also referred to as the ABBA-BABA test in other studies. If H1, H2 and H3 are taken to denote 3 human populations including MA-1, the test can be used to evaluate if the data is inconsistent with the null hypothesis that the tree (((H1, H2), H3), chimpanzee) is correct and that there has been no gene flow between H3 and either H1 or H2 or any populations related to them. There are several different definitions of the *D*-statistic in the literature. The definition used here is from Durand *et al.* (2011)[3]:

D = (nABBA-nBABA)/ (nABBA+nBABA)

where nABBA is the number of sites where H1 has the same allele as the chimpanzee, and H2 and H3 have a different allele (ABBA sites); and, nBABA is the number of sites where H2 has the same allele as the chimpanzee, and H1 and H3 have a different allele (BABA sites). Under the null hypothesis D=0. Hence a test statistic that differs significantly from 0 provides evidence either of the tree being incorrect or of gene flow (assuming no contamination or differential error rates). Following Green *et al.* (2010)[1] and Reich *et al.* (2010)[2], the significance of the deviation from 0 was assessed using a Z-score based on jack knife estimates of the standard deviation of the *D*-statistics. This Z score is based on the assumption that the *D*-statistic (under the null hypothesis) is normally distributed with mean 0 and a standard deviation equal to a standard deviation estimate achieved using the "delete-m Jackknife for unequal m" procedure described in Busing *et al.* (1999)[4].

SI 13.2 Data

MA-1 and 11 high-coverage present-day genomes from SI 4.2, except Dinka, Mbuti, San and Mandenka[5], were included in this test. For the chimpanzee outgroup, the multiway alignment, which includes both chimpanzee and human (pantro2 from the hg19 multiz46), was used.

SI 13.3 Data filtering

The data was filtered as follows before calculating the *D*-statistic[6]. First, all reads with mapping quality below 30 were removed. Subsequently, bases of low quality were removed by dividing all bases into 8 base categories: A, C, G, T on the plus strand and A, C, G, T on the minus strand and then discarding the 50% of the bases with the lowest quality score from each of the 8 categories. More specifically, within each base category, we:

1. found the highest base quality score, Q, for which less than half of the bases in the base category had a quality score smaller than Q.

2. removed all bases with quality score smaller than Q.

3. randomly sampled and removed bases with quality score equal to Q until 50% of the bases from the base category had been removed in total.

The data was filtered separately in each of the 8 base categories to avoid bias in the test in cases where there is a significant difference in the base quality between the categories. After filtering, a single base was sampled at each site for each individual in order to avoid introducing bias due to differences in sequencing depth. Finally, all sites containing transitions were removed, since the ancient MA-1 genome has *post mortem* damage that strongly increases errors in sites that contain transitions (SI 6).

SI 13.4 Test results

The *D*-statistics can be seen in Table SI 13. The table shows both the uncorrected *D*-statistic, D_{seq}, the jackknife bias corrected estimate, and a Z score. A Z score close to 0 indicates that the test is not significant. Following Green *et al*. (2010)[1] and Reich *et al*. (2010)[2], absolute *Z*-values higher than 3.0 were considered to be significant deviations from the null hypothesis. Additionally, only tests where MA-1 was at the H3 position were considered due to previous observation of the *D* test being sensitive to datasets with high error rates at the ingroup (H1,H2) positions[7].

SI 13.5 Contamination correction

In order to determine if the small amount of contamination in MA-1 would significantly affect the results of some of the more important tests, corrections for different amounts of contamination were attempted. To do so, another sequenced Central European (CEU) individual from the 1000 genomes project (NA11881) was included and CEU was assumed to be the contamination source. It was also assumed that the observed *D*-statistic ($\mathrm{D_{obs}}$) is a function of the uncontaminated *D*-statistic ($\mathrm{D_{cor}}$) and the *D*-statistic obtained from the CEU individual ($\mathrm{D_{CEU}}$):

$$\mathrm{D_{obs}} = (1-c)\,\mathrm{D_{cor}} + c\,\mathrm{D_{CEU}}$$

where c is the proportion of contamination from CEU. Two c values were used: one reflecting of the X-based contamination rate determined for MA-1 (1.6%, SI 5.2) and a second, higher rate (10%). For each of these two c values, $\mathrm{D_{cor}}$ was estimated for each 5Mb block of the genome by:

$$D_{cor} = \frac{D_{obs} - cD_{CEU}}{1-c}$$

Block jackknife was carried out in order to get an overall estimate of the *D*-statistic corrected for contamination.

Results are shown in Figure SI 20 for both 1.6% and 10% contamination rates and the tested topology ((population X, French),MA-1), where population X represents one of the 15 other modern genomes than French (SI 4.2). We see that the corrected *D*-statistics are similar to the original results. A contamination rate of 1.6% has a very minimal effect on the results. Furthermore, even assuming a 10% contamination rates the MA-1 show a stronger affinity to western Eurasians (CEU) than East Asians, demonstrating that contamination did not cause the observed affinity between MA-1 and modern Europeans.

References for SI 13

1. Green, R.E. *et al*. A Draft Sequence of the Neandertal Genome. *Science* **328**, 710-722 (2010)

2. Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053-1060 (2010)

3. Durand, E.Y., Patterson, N., Reich, D., Slatkin, M. Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28**, 2239-2252 (2011)

4. Busing, F.M.T.A., Meijer, E., Leeden, R.V.D.L. Delete-m Jackknife for Unequal m. *Statistics and Computing*. **9**, 3-8 (1999)

5. Meyer, M. *et al.* A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science* **338**, 222-226 (2012)

6. Orlando, L. *et al.* Recalibrating *Equus* evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* **499**, 74-78 (2013)

7. Rasmussen, M. *et al.* An Aboriginal Australian Genome Reveals Separate Human Dispersals in Asia. *Science* **334**, 94-98 (2011)

Table SI 13 Summary results for sequence data-based *D*-statistic tests. H1, H2 and H3 represent three human populations (one genome per population), where the test aims to confirm or reject the tree of the topology (((H1, H2), H3), chimpanzee). Tests are presented here for MA-1 at the H3 position and 11 modern-day individuals at the H1/H2 positions. The 'Difference' column represents (nABBA-nBABA); the 'Total' column represents (nABBA+nBABA) i.e. total number of ABBA and BABA sites; the 'D_{seq}' column represents the test statistic: (nABBA-nBABA)/(nABBA+nBABA) where a negative value means that H1 is closer to H3 than H2 is and a positive value implies that H2 is closer to H3 than H1 is; the 'Jackknife' column represents the bias corrected D_{seq}; the 'SD' column represents the estimated Jackknife standard error of the estimate used to obtain the *Z* value; and, the '*Z*' column represents the significance value of the test where a *Z* score of 3 and above is used as a critical value.

H1	H2	H3	Difference	Total	D_{seq}	Jackknife	SD	*Z*
Avar	French	MA-1	-1233	88911	-0.0138678	-0.01387275	0.006996071	-1.982227
Avar	Papuan	MA-1	-8203	96343	-0.08514371	-0.0851478	0.007273142	-11.70659
French	Papuan	MA-1	-7595	109125	-0.06959908	-0.06960216	0.007268381	-9.575597
Avar	Sardinian	MA-1	-1974	89302	-0.02210477	-0.02210356	0.006544603	-3.377557
French	Sardinian	MA-1	-928	99020	-0.009371844	-0.00936692	0.006495788	-1.442757
Papuan	Sardinian	MA-1	6713	108293	0.06198923	0.0619908	0.007125347	8.69982
Avar	Han	MA-1	-4829	94999	-0.05083211	-0.0508405	0.006783595	-7.493389
French	Han	MA-1	-4150	107064	-0.03876186	-0.0387636	0.006726005	-5.762984
Papuan	Han	MA-1	3496	103580	0.03375169	0.03374918	0.006887931	4.90012
Sardinian	Han	MA-1	-3150	106704	-0.02952092	-0.02952764	0.006547938	-4.50843
Avar	YRI	MA-1	-37776	110192	-0.3428198	-0.3428207	0.005540708	-61.87292
French	YRI	MA-1	-41274	123160	-0.335125	-0.335124	0.005553805	-60.34152
Papuan	YRI	MA-1	-33608	123276	-0.272624	-0.2726267	0.006059276	-44.99284
Sardinian	YRI	MA-1	-40311	122731	-0.32845	-0.3284517	0.005486565	-59.86442
Han	YRI	MA-1	-37114	123232	-0.3011718	-0.3011717	0.005699081	-52.84568
Avar	Karitiana	MA-1	2281	93705	0.02434235	0.02434043	0.007045153	3.455191
French	Karitiana	MA-1	3786	106666	0.03549397	0.03549525	0.007307643	4.857103
Papuan	Karitiana	MA-1	11407	105533	0.1080894	0.1080921	0.007184378	15.04506
Sardinian	Karitiana	MA-1	4707	107339	0.04385172	0.04384823	0.007070015	6.202494
Han	Karitiana	MA-1	7856	98904	0.07943056	0.07943509	0.007364753	10.78523
YRI	Karitiana	MA-1	45048	125930	0.3577225	0.3577215	0.005784439	61.84222

H1	H2	H3	Difference	Total	D_{seq}	Jackknife	SD	Z
Avar	Dai	MA-1	-3772	94462	-0.0399314	-0.03993808	0.006491504	-6.151333
French	Dai	MA-1	-2898	106856	-0.02712061	-0.02712309	0.006669438	-4.066402
Papuan	Dai	MA-1	4717	103253	0.0456839	0.04567717	0.007152974	6.3867
Sardinian	Dai	MA-1	-1967	105955	-0.01856448	-0.01857277	0.006632633	-2.798962
Han	Dai	MA-1	1230	92742	0.0132626	0.01326166	0.006486339	2.044697
YRI	Dai	MA-1	38344	122610	0.3127314	0.3127287	0.005652629	55.32495
Karitiana	Dai	MA-1	-6674	98898	-0.06748367	-0.06748738	0.007007639	-9.630015
Avar	Indian	MA-1	-1788	91566	-0.0195269	-0.019529	0.006357489	-3.07148
French	Indian	MA-1	-718	103212	-0.006956555	-0.006955259	0.006329264	-1.09911
Papuan	Indian	MA-1	6820	106076	0.06429353	0.06429486	0.007289902	8.819532
Sardinian	Indian	MA-1	272	103168	0.002636476	0.002633115	0.006227381	0.4233684
Han	Indian	MA-1	3354	103634	0.0323639	0.03236748	0.006238332	5.187909
YRI	Indian	MA-1	40244	122572	0.3283295	0.3283306	0.005235556	62.71148
Karitiana	Indian	MA-1	-4470	104556	-0.04275221	-0.0427505	0.007090573	-6.029443
Dai	Indian	MA-1	2195	103991	0.0211076	0.02111285	0.006555651	3.219756
Avar	Mari	MA-1	-1308	80866	-0.01617491	-0.01618044	0.006397109	-2.528471
French	Mari	MA-1	-44	87220	-0.000504472	-0.000505505	0.006413457	-0.0786582
Papuan	Mari	MA-1	6470	91712	0.07054693	0.07055002	0.006947302	10.15458
Sardinian	Mari	MA-1	717	87301	0.008212964	0.008207313	0.006695953	1.226556
Han	Mari	MA-1	3549	88741	0.03999279	0.03999633	0.006407265	6.241788
YRI	Mari	MA-1	34571	105487	0.3277276	0.327725	0.00547069	59.90608
Karitiana	Mari	MA-1	-3178	88754	-0.03580684	-0.03580566	0.006872115	-5.210454
Dai	Mari	MA-1	2382	88212	0.02700313	0.02700662	0.006476576	4.169352
Indian	Mari	MA-1	670	88904	0.007536219	0.007535854	0.006411522	1.175418
Avar	Tajik	MA-1	-527	89653	-0.005878219	-0.005876752	0.006617611	-0.8882691
French	Tajik	MA-1	958	99556	0.009622725	0.009626758	0.006626812	1.45209
Papuan	Tajik	MA-1	8488	107274	0.07912448	0.0791296	0.007134947	11.08971
Sardinian	Tajik	MA-1	1856	99910	0.01857672	0.0185793	0.006503995	2.856201

H1	H2	H3	Difference	Total	D_{seq}	Jackknife	SD	Z
Han	Tajik	MA-1	4934	104476	0.04722616	0.04723262	0.006585429	7.171311
YRI	Tajik	MA-1	41442	121892	0.3399895	0.339991	0.005564733	61.09718
Karitiana	Tajik	MA-1	-2768	104124	-0.02658369	-0.02658299	0.007253963	-3.664712
Dai	Tajik	MA-1	3757	104151	0.03607263	0.03608075	0.00659694	5.468084
Indian	Tajik	MA-1	1592	101258	0.01572221	0.01572405	0.00641966	2.449073
Mari	Tajik	MA-1	798	87070	0.00916504	0.009166897	0.006475934	1.415246

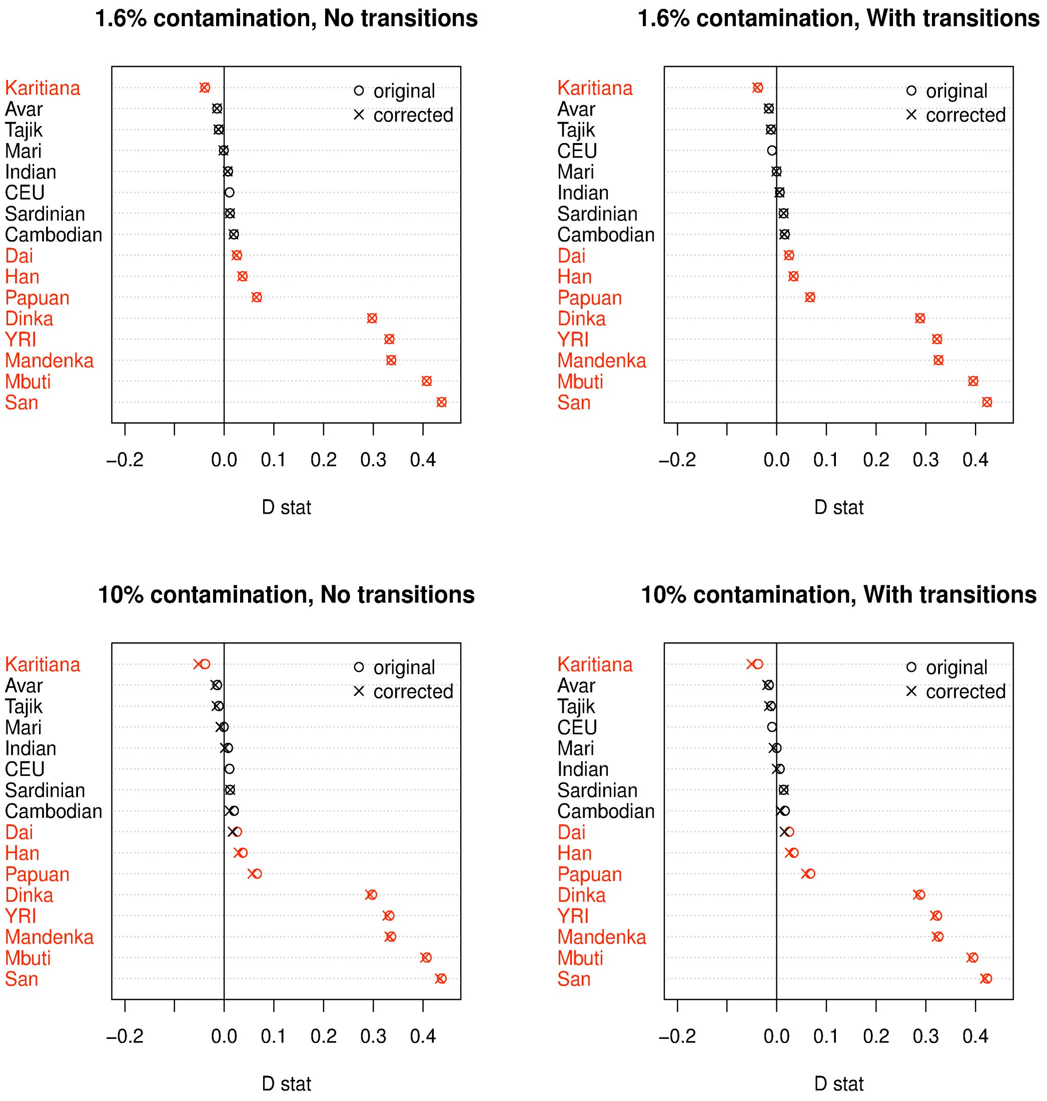


Figure SI 20 *D*-statistics corrected for contamination from a CEU source. Corrected sequence data-based *D*-statistics for the topology ((population X, French),MA-1), where population X represents one of 15 modern genomes (SI 4.2, excluding the French) and the CEU genome. Red color indicates a *Z* score higher than 3.

SI 14. Analyses using *f*-statistics and *D*-statistics based on allele frequency data (D_{freq})

SI 14.1 Methods and Data

To explore the relationship between MA-1 and modern-day human populations as well as explicitly test models of Eurasian population history, the *f*-statistic framework was used[1,2]. f_3-statistics (3-population tests) and *D*-statistic tests were computed using the estimators described in Patterson *et al.* (2012)[2], obtaining standard errors using a block jackknife procedure over 5 megabase blocks in the genome, except when explicitly noted otherwise. For the f_3-statistic, the normalization described in the main text and appendix of Patterson *et al.* (2012)[2] was used (see below).

Except when otherwise noted, a published 364,470 SNP dataset masked for European and African ancestries in Siberian and Native American populations was used[3], which was merged with additional data from Finnish populations[4]. Transition SNPs were included since the data contained very few transversion polymorphisms, but instead the first and last 3 bases of each sequence read were excluded since the majority of nucleotide misincorporations occur at the ends of ancient DNA templates (SI 6.2). For some analyses SNPs that were cleanly ascertained in a San individual[2], and in some cases also a Yoruba individual[2], was used, with the same filtering approach as in the tests where MA-1 was included. Since the San and Yoruba are approximate outgroups to non-African populations[2,5,6], this data is unbiased for all comparisons between non-Africans[2]. For these tests, the 12 individuals that were used for ascertainment by Patterson *et al.* (2012)[2] were always excluded. For other tests (where explicitly noted), SNP data described in Table SI 11 was used.

SI 14.2 Estimates of shared drift using f_3-statistics

When a single ancient genome is analyzed using multivariate approaches for studying population structure, such as PCA and model-based clustering, it is in the population genetic context of the much larger sample set of modern individuals, which may obscure patterns of interest to single-sample ancient data. However, it is difficult to estimate classical measures of pairwise genetic distance, such as F_{ST}, using low-to-medium coverage draft genomes where information is usually available only from a single gene copy from the ancient population. Furthermore, these distance measures are sensitive to genetic drift that has occurred in the population of interest since divergence from the ancient population lineage. If such lineage-specific genetic drift differs between two populations that share an equal amount of their genetic history with the ancient individual, the ancient individual will be observed as being closer to the modern population with the lower degree of historical genetic drift in distance-based methods such at F_{ST}.

To circumvent these issues and obtain a statistic that is informative of the genetic relatedness between a particular sample and each modern population in a reference set, an ‘outgroup f_3-statistic’ was computed of the form:

$$f_3(A,B;O) = \frac{\Sigma_{i=1}^{n}[(p_{iO} - p_{iA})(p_{iO} - p_{iB}) - \left(\left(\frac{h_{iO}(k_{iO} - h_{iO})}{k_{iO}(k_{iO} - 1)}\right)/k_{iO}\right)]}{\Sigma_{i=1}^{n}[2p_{iO}(1 - p_{iO})]}$$

Where h_{iO} is the count of the reference allele and k_{iO} is the number of gene copies in population *O* (the outgroup) at locus *i*. Correspondingly, p_{iN} is the allele frequency at locus *i* in population *N.* Assuming no mutation, selection or gene-flow, and in the absence of ascertainment bias, the expected value of the f_3-statistic[2] equals the sum of expected squared change in allele frequency (normalized for heterozygosity in the outgroup) due to genetic drift on the path in the population tree from the outgroup to the root and from the root to the ancestor of populations *A* and *B*. Since genetic drift in the lineage specific to the outgroup is expected to be constant regardless of which populations A and B are used (in the absence of gene flow), the remaining variation between statistics will depend on how much genetic history is shared between populations A and B. This statistic f_3(outgroup;MA-1,population B) was quantified for a set of 147 worldwide populations (including the novel genotype data from this study) using ~300k SNPs where data was available for MA-1. The highest f_3 -statistics were observed for Native Americans, followed by Siberian and northern European populations (Figure SI 21, Figure SI 22).

SI 14.3 *D*-statistic tests (D_{freq}) on data with various ascertainments document gene flow between the MA-1 lineage and Native American ancestors

The allele frequency based *D*-statistic test (D_{freq}) is a generalization of the sequence data based *D*-statistic test, and can be thought of as a test of treeness in the unrooted population tree (H4,H3),(H2,H1), under the same assumptions as in the previous test.

$$D_{freq}(H1,H2;H3,H4) = \frac{\sum_{i=1}^{n}[(p_{i1}-p_{i2})(p_{i3}-p_{i4})]}{\sum_{i=1}^{n}[(p_{i1}+p_{i2}-2p_{i1}p_{i2})(p_{i3}+p_{i4}-2p_{i3}p_{i4})]}$$

Where p_{i1},p_{i2},p_{i3}, and p_{i4} are the frequencies of an arbitrarily chosen reference allele at SNP *i* of H1, H2, H3 and H4, respectively. The statistic is summed over all *n* SNPs.

Tests assessing whether MA-1, Han and Native Americans can be described as having a tree-like population history were rejected using D_{freq} tests just as in the D_{seq} tests in Table SI 13. For instance, we obtain D_{freq}(Yoruba, MA-1; Han, Karitiana) = 0.069±0.0046 ($Z = 14.98$), using data from Reich *et al.* 2012[3] (Figure SI 23). For the SNPs ascertained in a San, D_{freq}(Yoruba, MA-1; Han, Karitiana) = 0.076±0.0056 ($Z =$ 13.6) was obtained. To ensure that our inferences were unaffected by ascertainment bias, D_{freq}(chimpanzee, MA-1; Han, Karitiana) was also computed for the ~13,000 transition SNPs ascertained in a San and a Yoruba, and $D_{\text{freq}} = 0.067\pm0.011$ ($Z = 6.4$) was obtained. These results are compatible with the results for the sequence data (Table SI 13). Furthermore, the deviation from treeness for D_{freq}(Yoruba, MA-1; Han, Karitiana) was found to be consistent when Karitiana was replaced with one of 51 other Native American populations (Figure SI 23), with $Z > 5$ for all tests.

SI 14.4 No difference between Native American populations in their affinity to MA-1

To further investigate the evidence of gene flow between MA-1 and different Native American populations, f_3(Yoruba; MA-1, Native American) was obtained using 52 Native American populations. Eskimo-Aleut speakers were found to have a lower degree of shared drift with MA-1, but no statistically significant differences existed between populations designated as having 100% 'First American' ancestry by Reich *et al.* (2012)[3]. This is consistent with admixture between the MA-1 lineage and Native Americans occurring prior to the diversification of Native American populations. To test this explicitly, D_{freq}(Han, MA-1; X, Karitiana) tests were performed where X was one of 51 Native American populations other than Karitiana. In agreement with the outgroup f_3 statistics, all populations of entirely 'First American' ancestry[3] were consistent with forming a clade with Karitiana to the exclusion of MA-1 and Han (Figure SI 24). However, in tests involving Eskimo-Aleut populations a significant skew was observed where Karitiana was closer to MA-1 (Figure SI 24).

SI 14.5 Genetic affinities between Native Americans and all western Eurasians

If there was gene flow from a western Eurasian group such as the MA-1 lineage into Native Americans, it should also be detectable with modern populations. In contrast, if the genetic affinity between MA-1 and Native Americans can be explained entirely by gene flow into the MA-1 lineage, then no affinity between Native Americans and modern-day western Eurasians should be seen. Patterson *et al*. (2012)[2] recently showed that 3-population tests in Northern Europeans showed significant evidence of admixture when Sardinians and Native Americans or Northeast Asians were used as source populations, suggesting that this could be due to admixture from Sardinian-related Neolithic farmers during the spread of agriculture in Europe[8,9]. Importantly, this could also be compatible with a model where pre-agricultural populations in Europe were close to Native Americans because they were part of a larger northern Eurasian population that contributed to some of Native American ancestry.

Given a population history where Native Americans can be described as having diverged from East Asian (Han Chinese) ancestors without subsequent gene flow from western or Central Asian-related sources, D_{freq}(Yoruba, X; Han, Karitiana) would be expected to be consistent with 0 also for Middle Eastern, South Asian and Central Asian populations. However, this topology is rejected for all populations from the above-mentioned regions in the Human Genome Diversity Panel (HGDP) (using San and Yoruba-ascertained SNPs), as shown in Figure 3b. To test whether these results could be due to events in the more recent history of East Asians, the analysis was repeated replacing Han with sequence data from chromosome 21 of the Tianyuan individual[10]. Data from the Tianyuan individual that had been ascertained to be informative of archaic ancestry by the authors of the original study was excluded. To increase power, transition SNPs were included and block jackknife resampling was performed over 100 kb blocks. Data was also pooled, including both Yoruba and San for use as outgroup, and using multiple Native American groups instead of using only Karitiana (a similar pattern was seen using only Karitiana): Maya1, Maya2, Aymara, Karitiana, Cree, Ojibwa, Algonquin, and Surui. D_{freq}(Africans, X; Tianyuan, Americans) was tested and several Z scores > 2 were observed. By and large, a clear overall difference can be seen between these tests and when the Native Americans were replaced with Asians D_{freq}(Africans, X; Tianyuan, Asians), in which case the topology is not rejected (Han,Yizu, Dai, Lahu, and Miaozu were used for Asians) (Figure SI 25). This suggests that the skews in the tests D_{freq}(Yoruba, western Eurasian; Han, Karitiana) are not due to recent events in East Asian population history.

As a complementary analysis to the direct tests for asymmetry between Native Americans and East Asians in their relationship to western Eurasians, two outgroup f_3-statistics were compared: f_3 (Yoruba; Han, X) and f_3(Yoruba; Karitiana, X) for worldwide populations, using both the SNP data with broad geographic coverage and the data with outgroup (San) ascertainment. The ratio of these in the absence of gene flow since after the divergence of Han and Karitiana and ascertainment in a true outgroup is expected to be 1.0, but results show that western, southern and Central Eurasians are shifted towards Karitiana, consistent with above results (Figure SI 26). In this analysis, much of the variation between populations in these statistics is seen regardless of whether Han or Karitiana are in the test. A similar pattern is seen using 3-population tests of the form used to document admixture in Northern Europe by Patterson *et al*. (2012)[2]. For San ascertained data, except in the case of Orcadians, when $f_3(X$; Sardinian, Karitiana) was significantly negative there was also a significantly negative score for $f_3(X$; Sardinian, Han). Mozabite, Bedouin, Palestinian, Balochi, Brahui, Burusho, Kalash, Makrani, Pathan and Sindhi populations were investigated in this analysis, in addition to the European populations. This suggests that the signal for admixture in Europe is not specific to Native Americans as a candidate source population, in agreement with the study by Patterson *et al*. (2012)[2].

SI 14.6 Equal affinity of MA-1 to Han and Papuans

If the direction of gene flow was into the ancestors of MA-1 from the Native American population lineage, MA-1 would be expected to be closer to East Asians than to Papuans, much like modern-day Native Americans are. However, as described in the main text, MA-1 is consistent with being equally close to Han and Papuans, in contrast to modern-day Native Americans as seen in the test D_{freq}(Papuan, Han; Sardinian, MA-1) = -0.002±0.005 (Z = -0.36). It is unlikely that MA-1 could be equally close to Han and Papuan but at the same time closer to Karitiana without gene flow into Native Americans after their divergence from East Asian ancestors. Such gene flow is also more consistent with the temporal and geographical placement of MA-1 (carrying unprecedented affinity to both western Eurasians and Native Americans at 24 ka in southern Siberia), rather than extensive gene flow from the Native American lineage into MA-1 ancestors.

However, to investigate whether the result of D_{freq}(Papuan, Han; Sardinian, MA-1) $\approx$ 0 could be due to sequencing errors in MA-1 that would attract it to Papuans (who have more archaic ancestry than Han and Sardinians), D_{freq}(Sardinian, MA-1; Han, Karitiana) was compared to D_{freq}(Sardinian, MA-1; Papuan, Karitiana). If sequencing errors are responsible for the consistency with 0 of D_{freq}(Sardinian, MA-1; Papuan, Han), causing an affinity between MA-1 and Han to go undetected, a strong difference in these two statistics is expected.

However, we observed statistically consistent results for the two tests:

With transitions:

D_{freq}(Sardinian, MA-1; Han, Karitiana) = 0.070±0.0057 (Z=12.4)

D_{freq}(Sardinian, MA-1; Papuan, Karitiana) = 0.064±0.0068 (Z=9.4)

Without transitions:

D_{freq}(Sardinian, MA-1; Han, Karitiana) = 0.071±0.0096 (Z=7.4)

D_{freq}(Sardinian, MA-1; Papuan, Karitiana) = 0.066±0.012 (Z=5.5)

SI 14.7 Evidence for basal East Asian gene flow into MA-1

Putative western Eurasian ancestry in Native Americans does not exclude the possibility of some gene flow into the MA-1 lineage from eastern Eurasian populations. To test this, we compared f_3(Yoruba; MA-1, X) and f_3(Yoruba; Sardinian, X), where Population X represented one population from a set of worldwide populations. Under a model where MA-1 is from the same lineage as the Sardinians, the ratio of these two statistics for unrelated populations is expected to be 1.0, but East Asians and Oceanians were both observed as being closer to MA-1 than to the Sardinian (Figure SI 27). Since the ratio is ~1.09 for both Oceanians and East Asians in the outgroup-ascertained data, the MA-1 lineage may have absorbed some ancestry from populations ancestral to these groups. Indeed, this is also consistent with the admixture graphs inferred using MixMapper (SI 12). However, a note of caution here is that the MA-1 data is of lower quality than the Sardinian data.

SI 14. 8 No European population is as close to Native Americans as is MA-1

Previous studies have found that recent admixture of European and African origin is widespread in Native American populations[3]. While our inferences are always based either on populations that are not recently admixed (Karitiana) or data for which genomic tracts derived from recent admixture had been excluded[3], we tested whether our results could be explained by recent European admixture by comparing results for MA-1 with those for 15 European populations (SI Table 11) in a 525,269 SNP data set. D_{freq}(Yoruba, European; Han, Karitiana) ranged from 0.020±0.007 (Sardinians) to 0.036±0.007 (Lithuanians) in these populations (Figure SI 28). This is in contrast to D_{freq}(Yoruba, MA-1; Han, Karitiana) = 0.089±0.015 in the same data. This shows that the evidence of gene flow presented in this study cannot be explained by historical, post-Columbian admixture between Europeans and Native Americans.

References for SI 14

1. Reich, D. *et al.* Reconstructing Indian population history. *Nature* **461**, 489-494 (2009)

2. Patterson, N. *et al*. Ancient Admixture in Human History. *Genetics* **192**, 1065-1093 (2012)

3. Reich, D. *et al.* Reconstructing Native American population history. *Nature* **488**, 370-374 (2012)

4. Surakka, I. *et al*. Founder population-specific HapMap panel increases power in GWA studies through improved imputation accuracy and CNV tagging. *Genome. Res.* **20**, 1344-1351 (2010)

5. Schlebusch, C.M. *et al*. Genomic Variation in Seven Khoe-San Groups Reveals Adaptation and Complex African History. *Science* **338**, 374-379 (2012)

6. Pickrell, J.K., Pritchard, J.K. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genet*. **8**, e1002967 (2012)

7. Ramachandran, S. *et al*. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 15942-15947 (2005)

8. Keller, A. *et al*. New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nat. Commun*. **3**, doi: 10.1038/ncomms1701 (2012)

9. Skoglund, P. *et al*. Origins and Genetic Legacy of Neolithic Farmers and Hunter-Gatherers in Europe. *Science* **336**, 466-469 (2012)

10. Fu, Q. *et al.* DNA analysis of an early modern human from Tianyuan Cave, China. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 2223-2227 (2013)

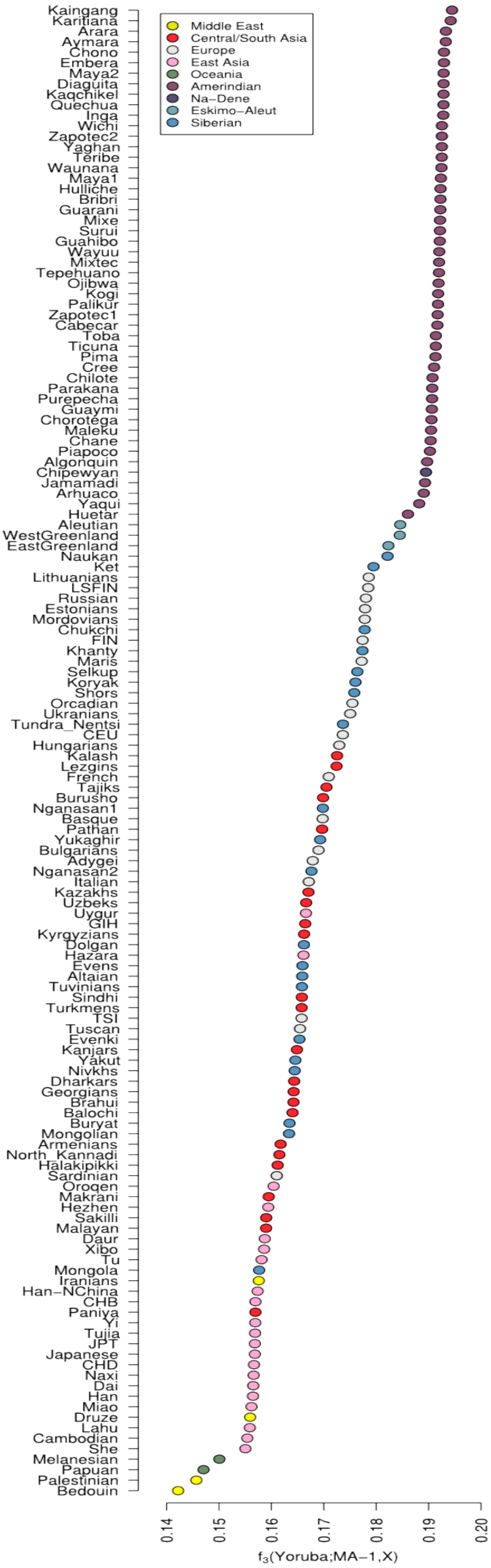


Figure SI 21 Shared drift with MA-1 in 147 non-African populations estimated using *f*$_3$(Yoruba; MA-1, *X*). Error bars are 1 standard error. These results correspond to those in Figure 1c.

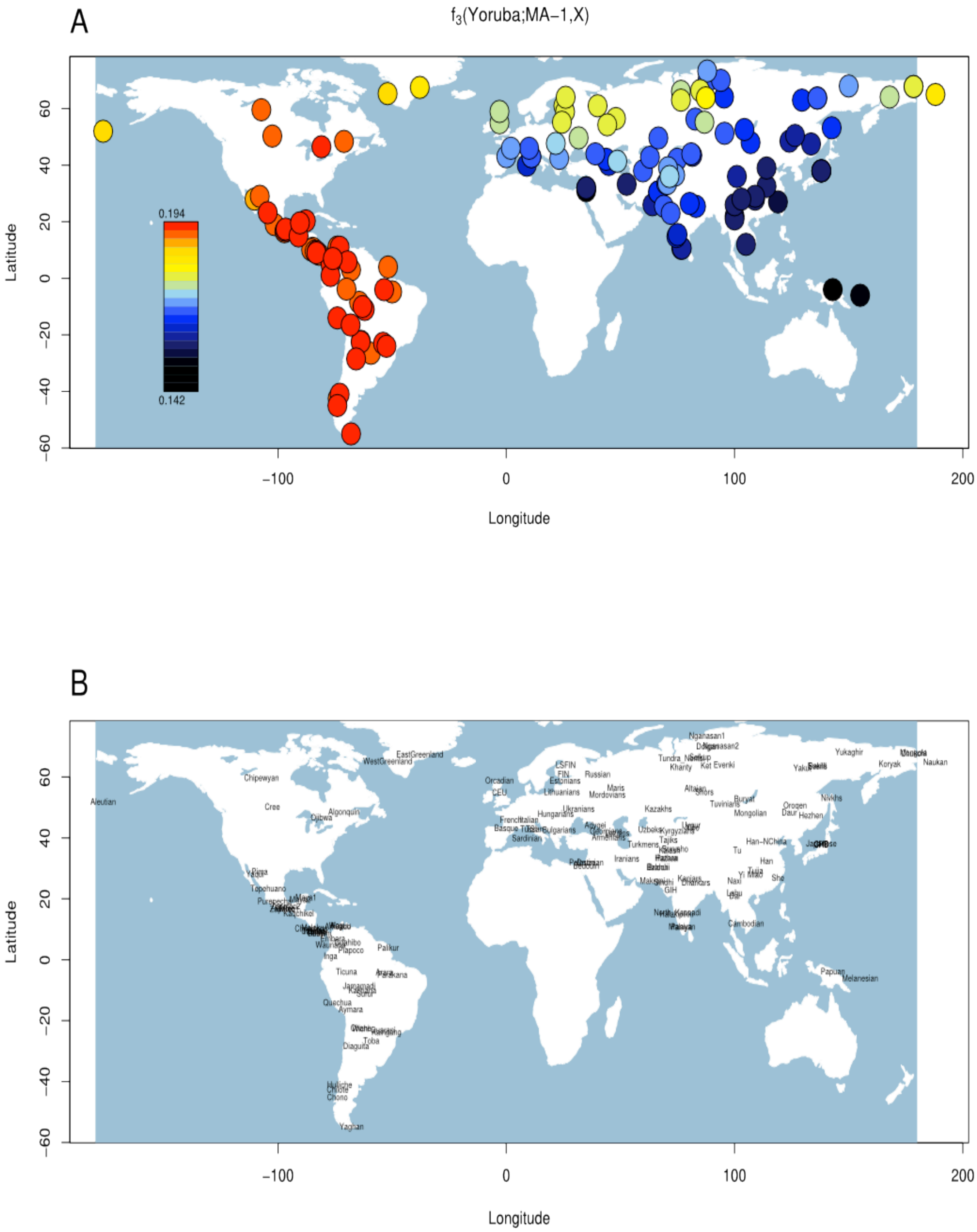


Figure SI 22 Global distribution of estimated shared drift with MA-1 in 147 non-African populations estimated using *f*$_3$(Yoruba; MA-1, *X*). A) Heat map. **B)** Details of sampling locations for different populations. These results correspond to those in Figure 1c.

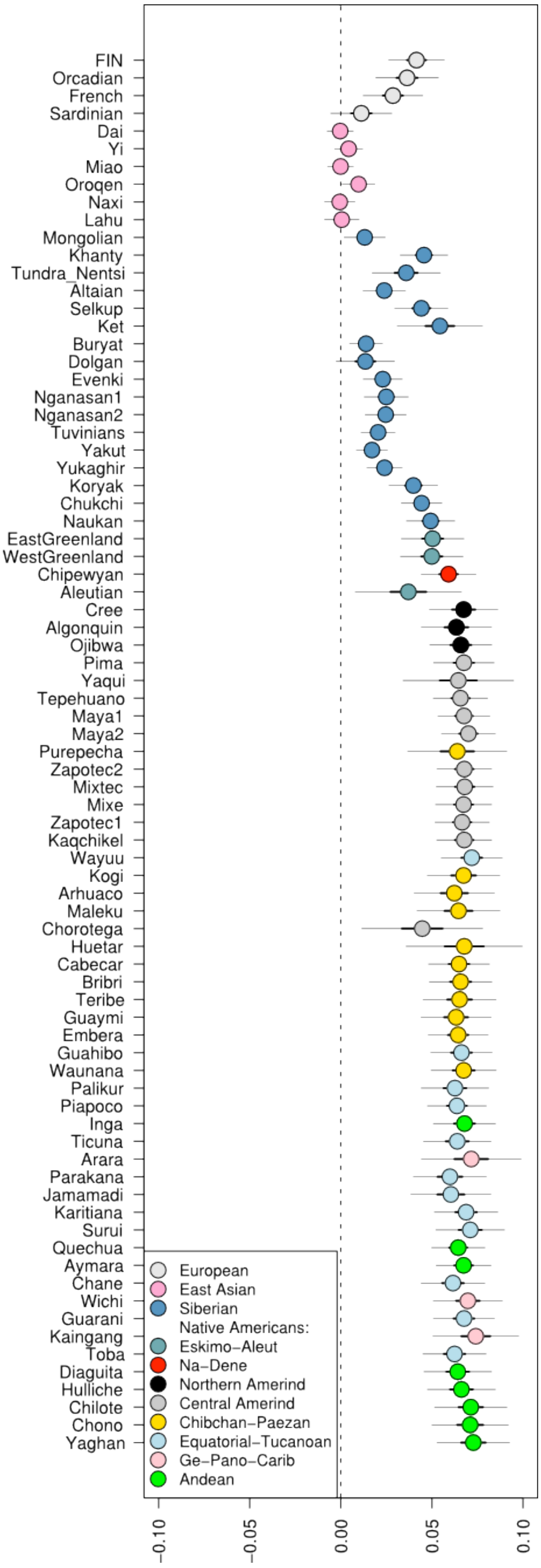


Figure SI 23 *D*-statistic tests of the form D_{freq}(Yoruba, MA-1; Han, *X*) where *X* is a non-African population. Tests using Native Americans, Siberians and Europeans are inconsistent with X and Han being monophyletic to the exclusion of MA-1.

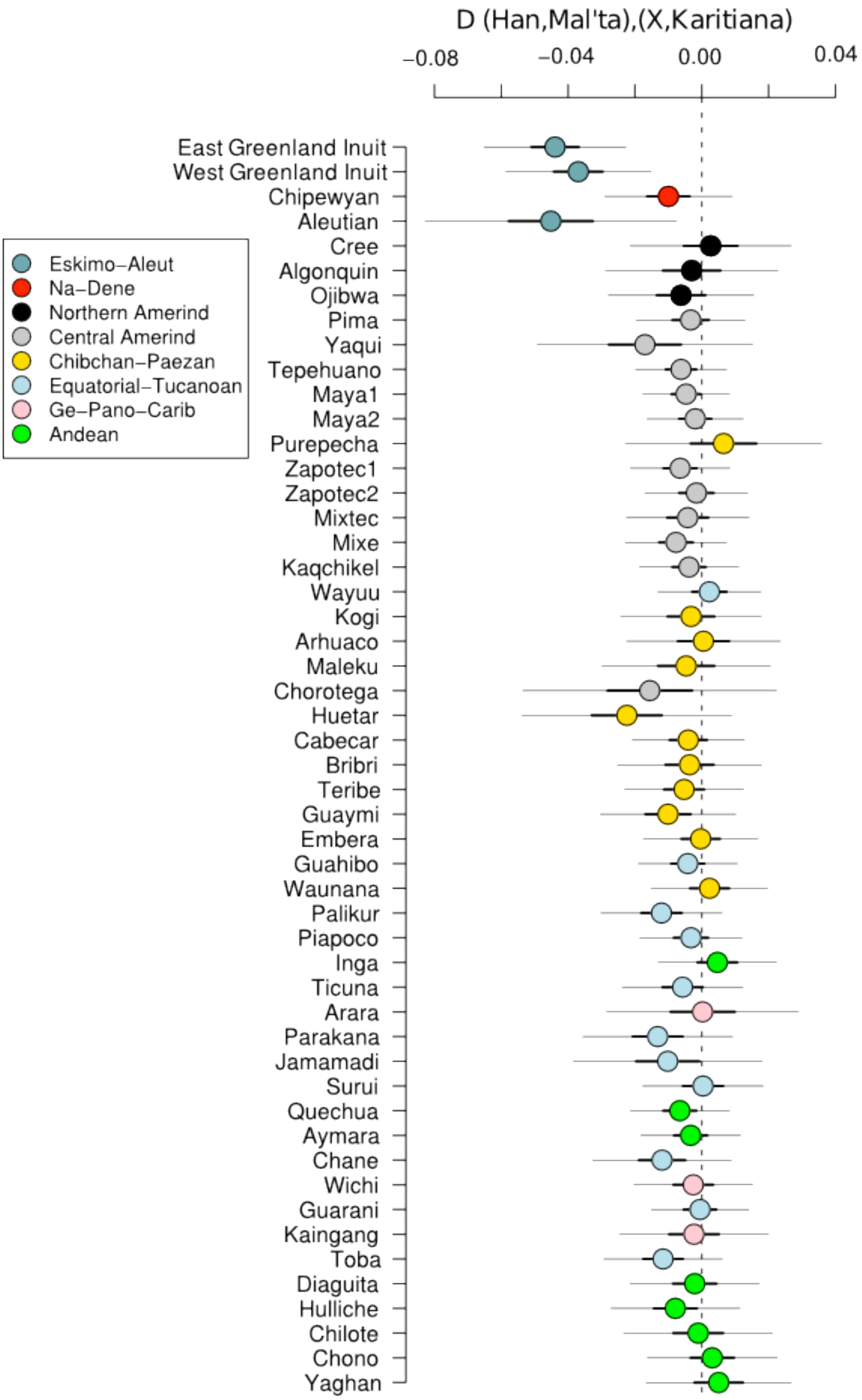


Figure SI 24 Test for differences between 51 Native American populations in their affinity to MA-1 compared to the Karitiana. Only Eskimo-Aleut speakers are significantly closer to Han than to MA-1 (Mal'ta), consistent with previous evidence for them having ancestry additional to the first migration into the Americas (Reich *et al*. 2012)[3].

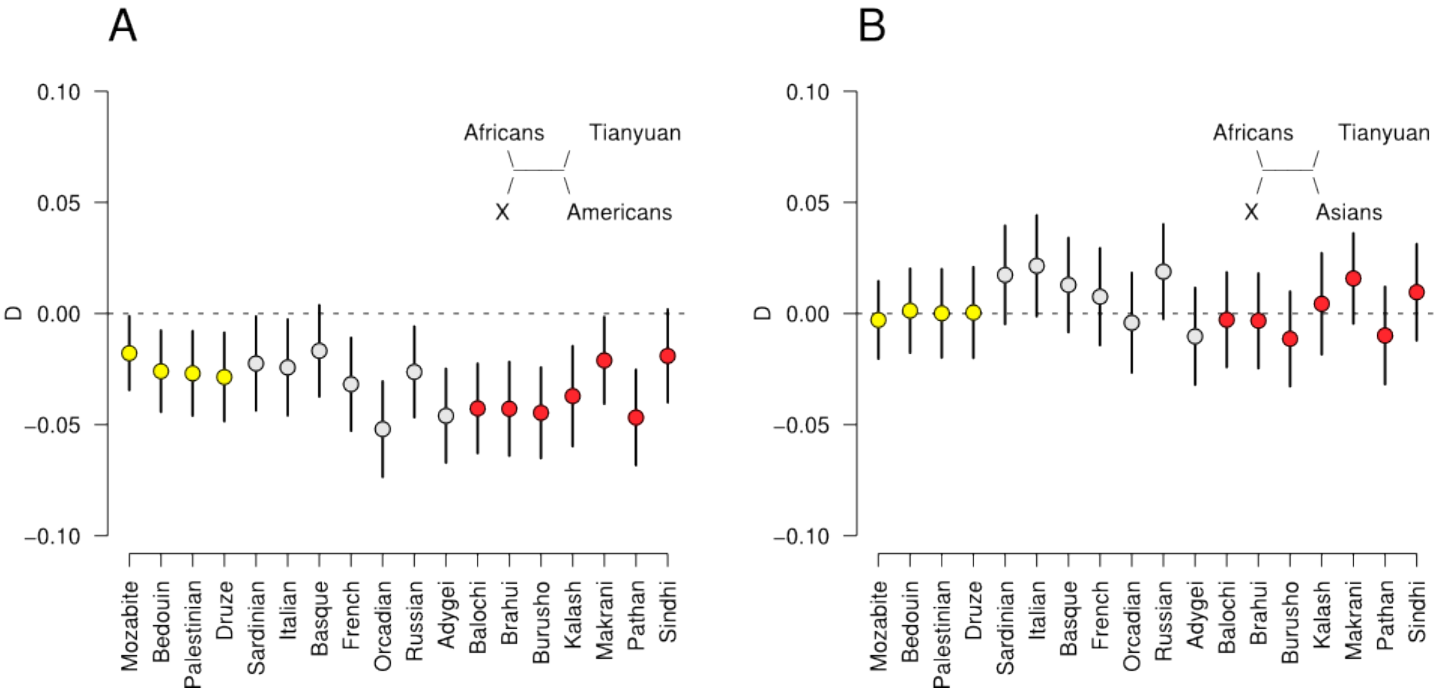


Figure SI 25 Tests for affinity between western Eurasians and Native Americans using the Tianyuan individual in place of East Asians. **A)** Tendency of Middle Eastern, European, and Central and South Asian populations to all be closer to Native Americans than to Tianyuan in the test D_{freq}(Africans, *X*; Tianyuan, Americans). "Americans" denotes pooled data from Maya1, Maya2, Karitiana, Surui, Aymara, Algonquin, Cree, and Ojibwa. **B)** No significant evidence of modern-day Asians being closer to Middle Eastern, European or Central and South Asian population than Tianyuan in the test D_{freq}(Africans, *X*; Tianyuan, Asians). "Asians" denotes pooled data from Han, Yizu, Dai, Lahu, and Miaozu. "Africans" denotes pooled data from San and Yoruba for both panels.

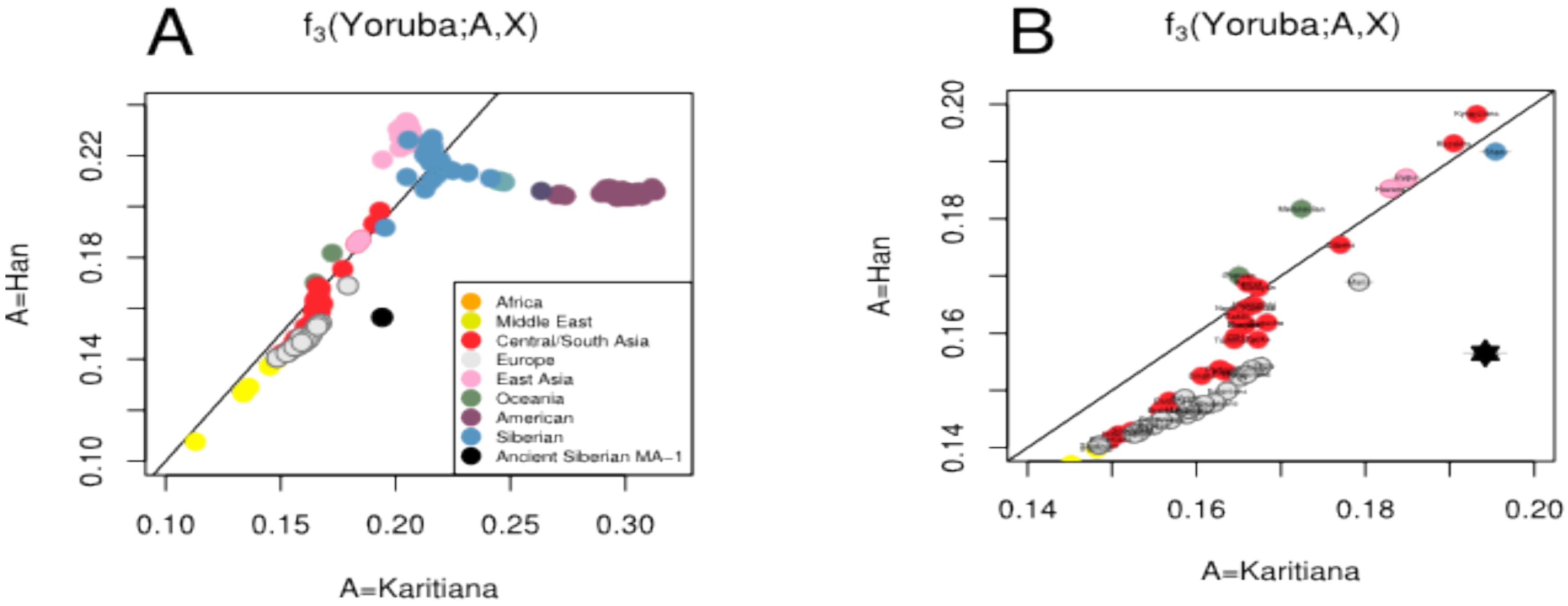


Figure SI 26 Contrasting shared drift between different modern populations and Han and Karitiana using an outgroup *f*$_3$-statistic. All western Eurasian populations are closer to Karitiana than to Han, but most variation between populations is seen regardless of whether Han or Karitiana are used. **A)** presents analysis of data from worldwide populations with unknown ascertainment. **B)** presents a zoom-in of the analysis in A) with population labels.

doi:10.1038/nature12736

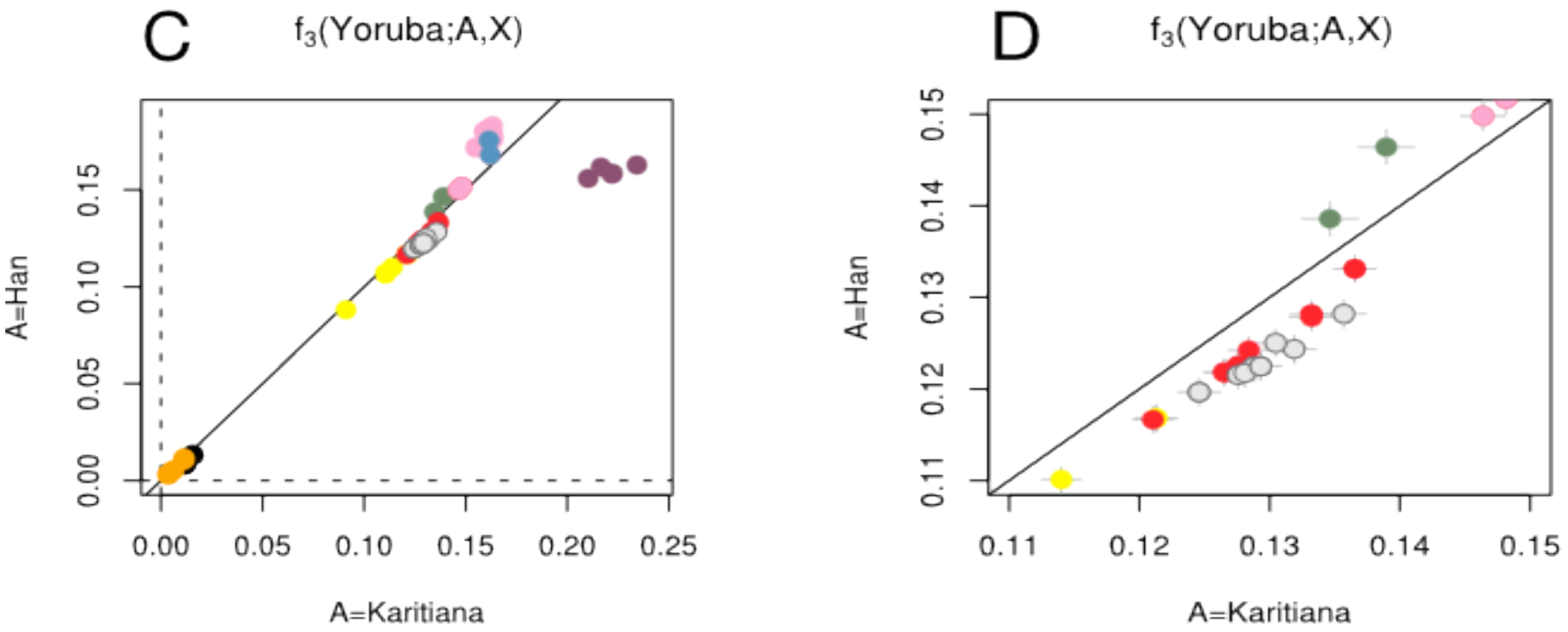


Figure SI 26 Contrasting shared drift between different modern populations and Han and Karitiana using an outgroup *f*$_3$-statistic. All western Eurasian populations are closer to Karitiana than to Han, but most variation between populations is seen regardless of whether Han or Karitiana are used. **C)** presents analysis of outgroup-ascertained (San) data and **D)** is a zoom-in of C).

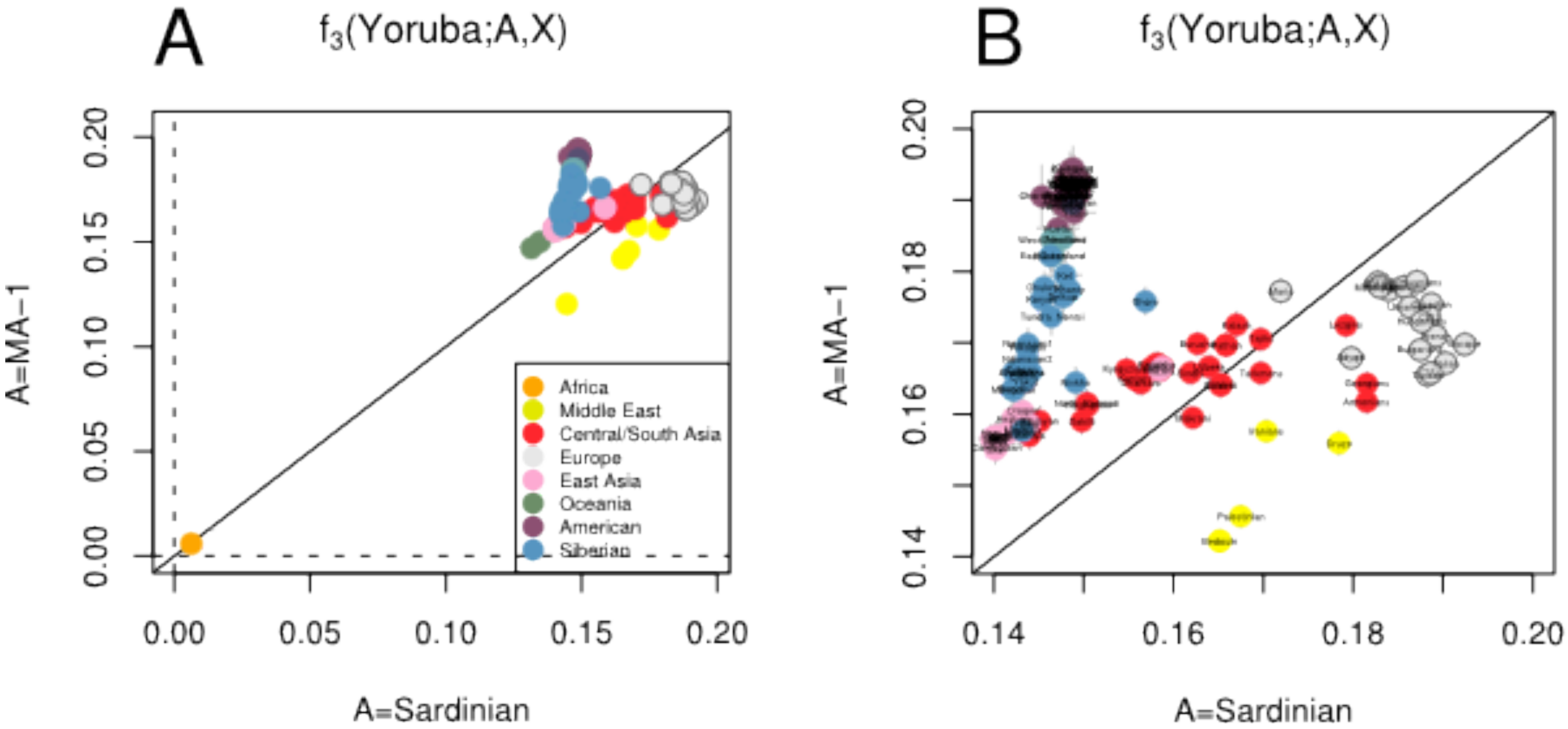


Figure SI 27 Contrasting shared drift between different modern populations and, Sardinian and MA-1 using an outgroup *f*$_3$-statistic. All East Asian populations (pink) are closer to MA-1 than to Sardinian, and no difference is seen with Oceanians (dark green). **A)** presents analysis of data from worldwide populations with unknown ascertainment. **B)** presents a zoom-in of the analysis in **A)** with population labels.

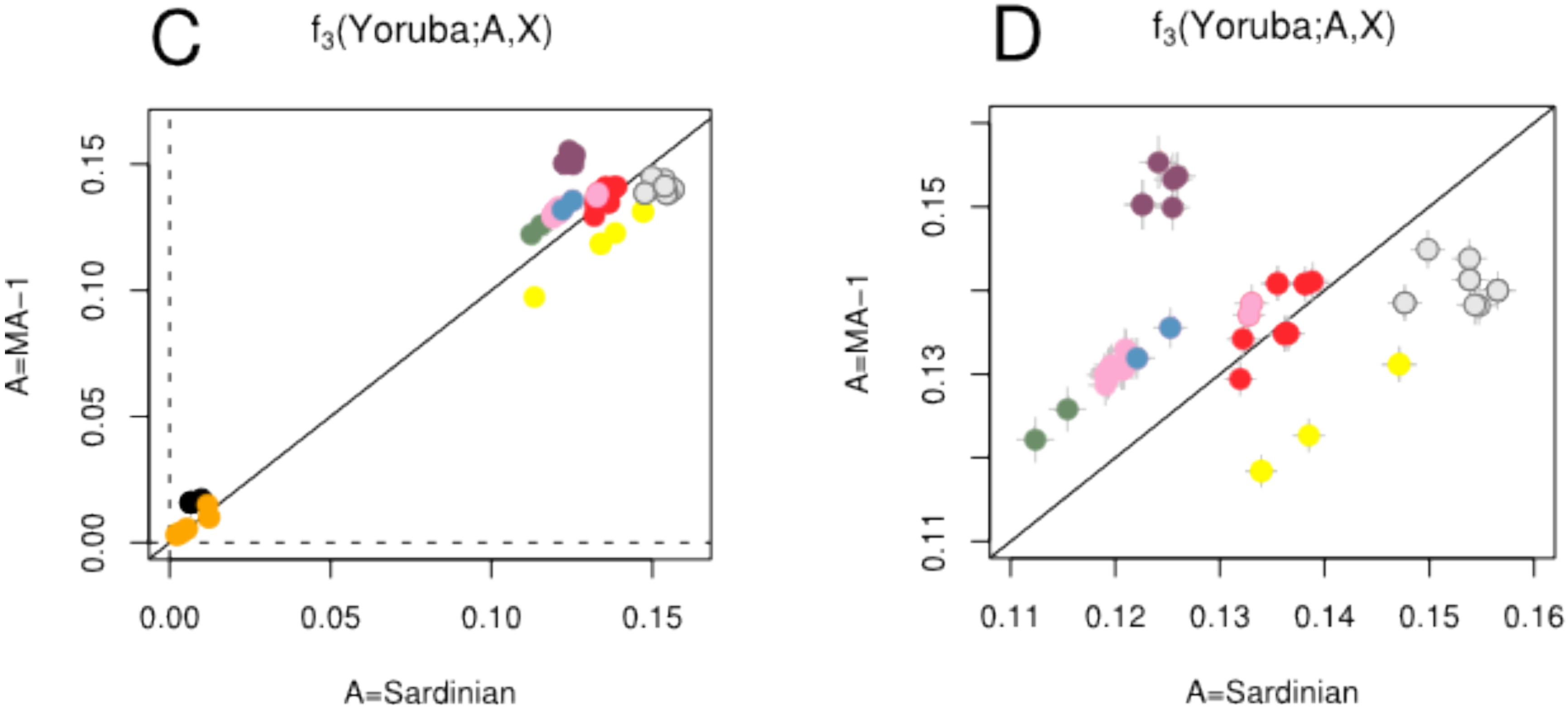


Figure SI 27 Contrasting shared drift between different modern populations and, Sardinian and MA-1 using an outgroup f_3-statistic. All East Asian populations (pink) are closer to MA-1 than to Sardinian, and no difference is seen with Oceanians (dark green). **C)** presents analysis of outgroup-ascertained (San) data and **D)** is a zoom-in of C).

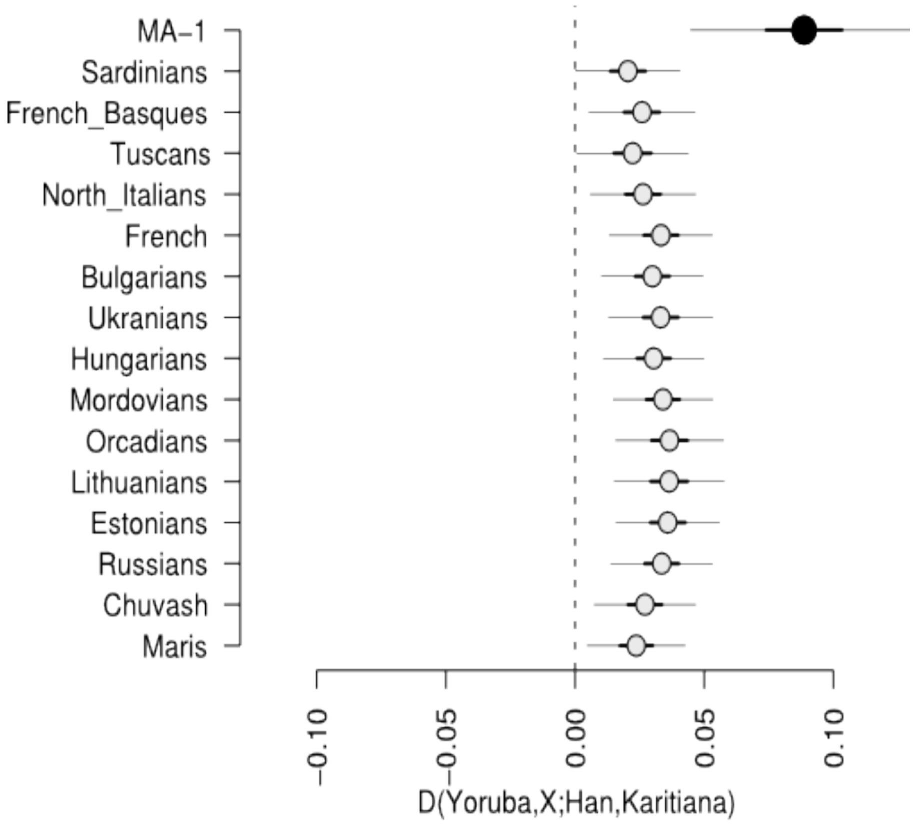


Figure SI 28 No modern-day European population is as close to Native Americans as is MA-1 in the test *D*(Yoruba, European; Han, Karitiana)

SI 15. Analysis of AG-2 using nucleotide misincorporation patterns

Both mitochondrial and X-chromosomal data suggest that the sequence data from AG-2 is affected by substantial present-day DNA contamination (SI 5). To look more closely at the affinities of the data from this individual, data from non-African populations[1,2,3] were used as a reference panel and PCA was performed as detailed in section SI 10. Data from AG-2 showed a similar genetic makeup as MA-1, being projected intermediate to the positions of Native Americans and Europeans (Figure SI 29). To compare the two PCA results more closely, Procrustes transformation was performed as in Skoglund *et al.* (2012)[4], rotating the PC1-PC2 configurations obtained for the two individuals to the configuration obtained using only the reference panel. In this analysis, AG-2 is slightly closer to Europeans than is MA-1, suggesting that the contamination detected is possibly from a European source.

The analysis was repeated using a previously described approach[4], where only those sequences which displayed a C → T mismatch consistent with *post-mortem* ancient DNA nucleotide misincorporations (PMD) in the first 5 bases of the sequence read (requiring a base quality of at least 30) were used. Nucleotide misincorporations are not extensively seen in modern-day DNA or contaminants, but are evident in both the AG-2 and MA-1 (SI 6.2) datasets and other ancient DNA samples studied so far[5,6,7], and have been used previously to test for contamination in low-coverage sequence data[4]. To test for potential reduction in contamination, we used data from a 100-year-old Australian Aborigine hair sample[8], which shows approximately the same level of *post-mortem* damage as AG-2 (~2-3%), and from a modern-day French individual. Out of 1 million sampled sequences from each individual, we found that only 2,805 (0.3%) sequence reads from the modern-day French data fulfilled the criteria of showing a C → T mismatch in the first 5 bases, compared to 40,675 (4%) of the Australian hair, a ~15 fold enrichment in the degraded DNA. Assuming that the damage in the Australian hair sample is similar to that of the endogenous DNA in AG-2 and that the damage in the French individual is the same as the contamination in AG-2, we would expect the restriction to sequences with PMD to result in a reduction of contamination by an order of magnitude, from ~30% to <5%.

While the resolution for AG-2 is poor on the PCA, the results are consistent with those from the full analysis (Figure SI 29). However, to assess this more closely, D_{freq} tests were performed on AG-2 and MA-1 using both the original data (D_{full}) and the data with evidence of post-mortem damage (D_{PMD}). We find:

D_{full}(Yoruba, AG-2; Han, Karitiana) = 0.078±0.0039 ($Z = 19.9$)

D_{PMD}(Yoruba, AG-2; Han, Karitiana) = 0.11±0.020 ($Z = 5.56$)

D_{full}(Yoruba, MA-1; Han, Karitiana) = 0.083±0.0038 ($Z = 21.96$)

D_{PMD}(Yoruba, MA-1; Han, Karitiana) = 0.086±0.0082 ($Z = 10.52$)

These results suggest that the signal of Native American affinity is present in both the full data and the PMD-filtered data for both the ancient individuals. AG-2 also showed a stronger similarity to Europeans for both data subsets:

D_{full}(Yoruba, AG-2; Han, French) = 0.054±0.0030 ($Z = 18.18$)

D_{PMD}(Yoruba, AG-2; Han, French) = 0.039±0.015 ($Z = 2.58$)

D_{full}(Yoruba, MA-1; Han, French) = 0.030±0.0031 ($Z = 9.69$)

D_{PMD}(Yoruba, MA-1; Han, French) = 0.017±0.0065 ($Z = 2.56$)

This would be consistent with contamination of European origin in a sample with largely similar ancestry as MA-1.

References for SI 15

1. Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100-1104 (2008)

2. Surakka, I. *et al*. Founder population-specific HapMap panel increases power in GWA studies through improved imputation accuracy and CNV tagging. *Genome. Res.* **20**, 1344-1351 (2010)

3. International HapMap3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-58 (2010)

4. Skoglund, P. *et al*. Origins and Genetic Legacy of Neolithic Farmers and Hunter-Gatherers in Europe. *Science* **336**, 466-469 (2012)

5. Briggs, A.W. *et al*. Targeted Retrieval and Analysis of Five Neandertal mtDNA Genomes. *Science* **325**, 318-321 (2009)

6. Krause, J. *et al*. The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature* **464**, 894-897 (2010)

7. Sawyer, S., Krause, J., Guschanski, K., Savolainen, V., Pääbo,S. Temporal Patterns of Nucleotide Misincorporations and DNA Fragmentation in Ancient DNA. *PLoS One* **7**, e34131 (2012)

8. Rasmussen, M. *et al.* An Aboriginal Australian Genome Reveals Separate Human Dispersals in Asia. *Science* **334**, 94-98 (2011)

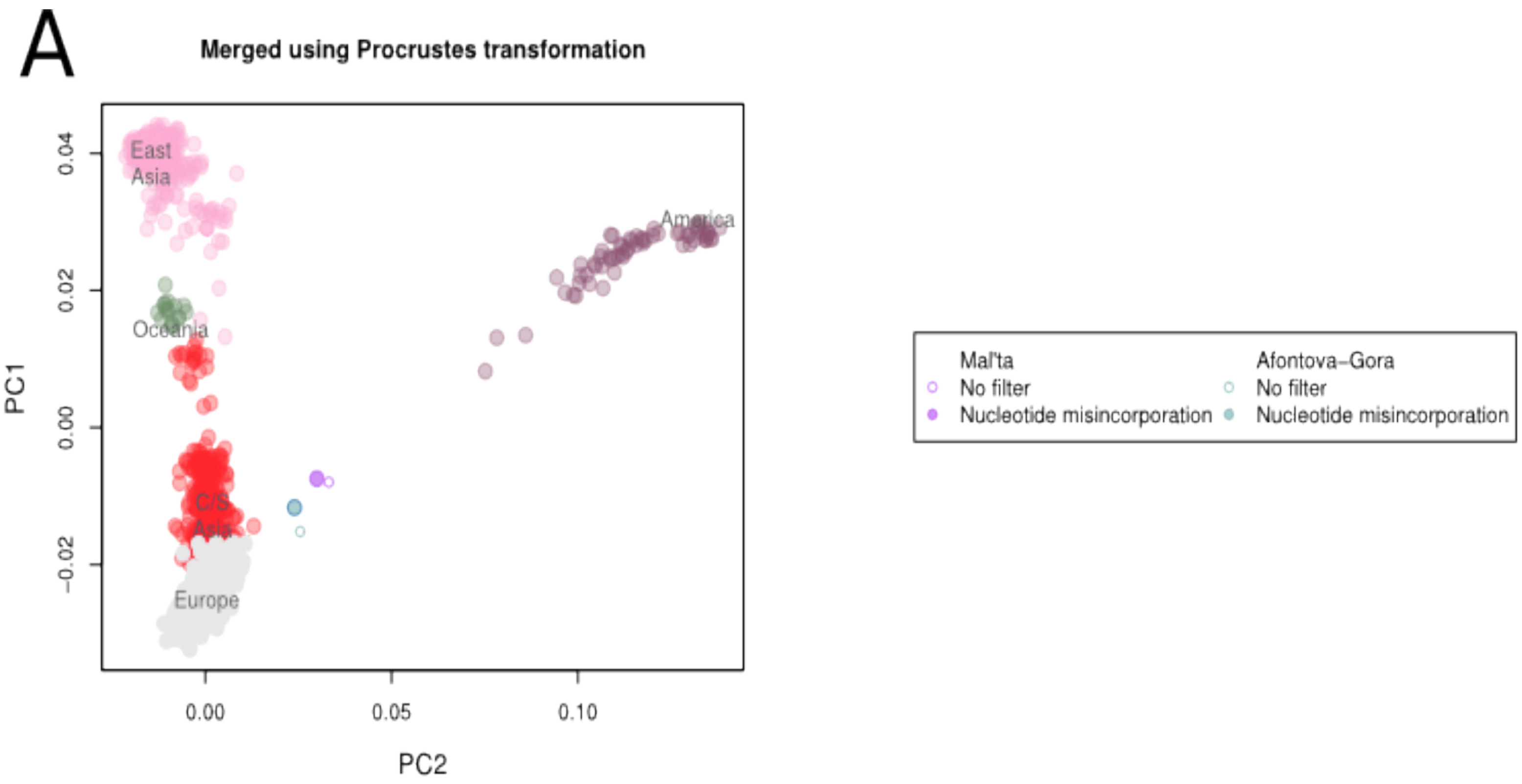


Figure SI 29 Principal component analysis of Afontova-Gora and MA-1 individuals restricted to sequences with nucleotide misincorporations. **A)** Merged PC1-PC2 configurations using Procrustes transformation.

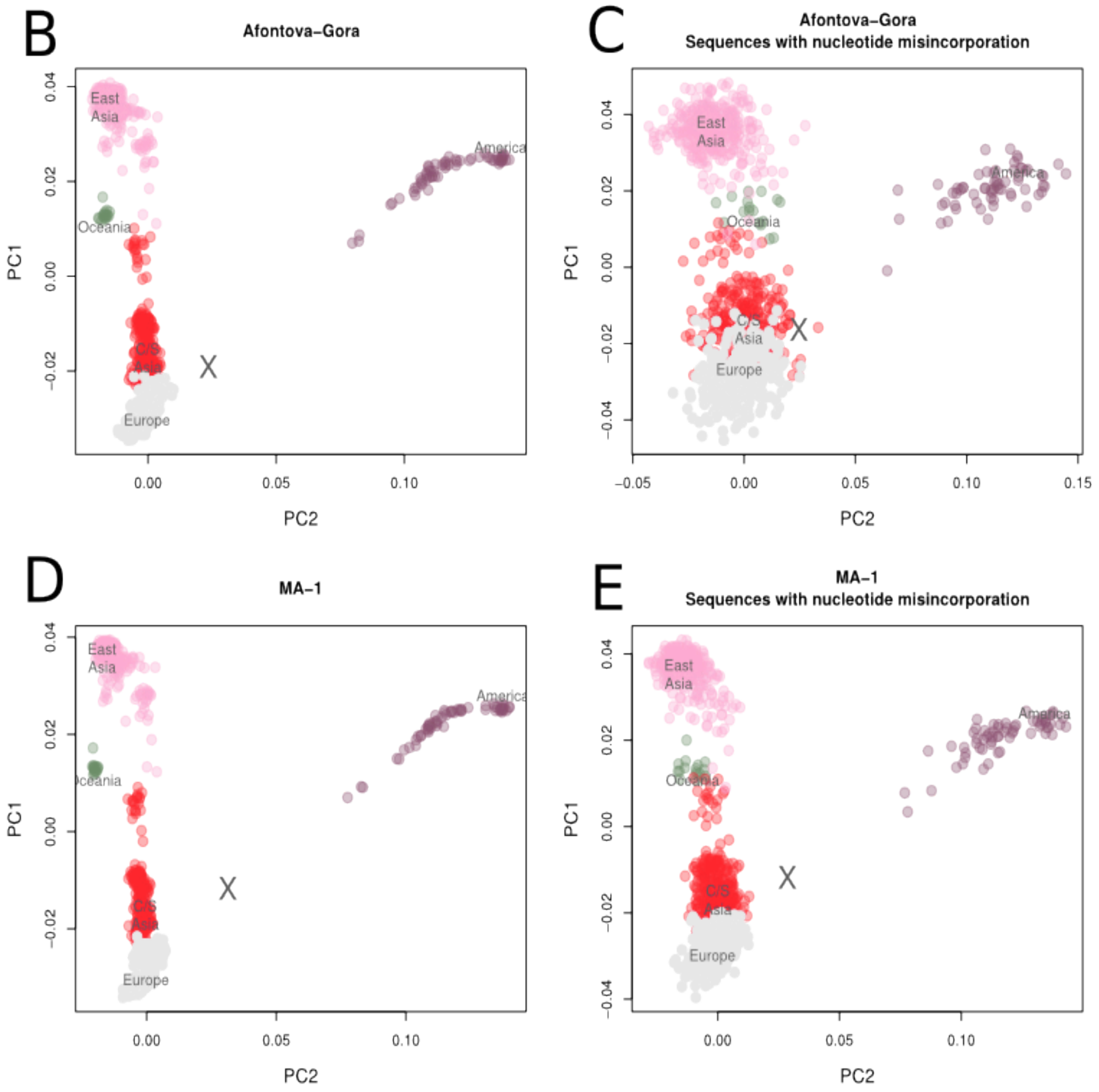


Figure SI 29 Principal component analysis of the Afontova-Gora and MA-1 individuals restricted to sequences with nucleotide misincorporations. **B)** PCA of the full Afontova Gora data. **C)** PCA of Afontova Gora data restricted to sequences with nucleotide misincorporations. **D)** PCA of the full MA-1 data. **E)** PCA of MA-1 data restricted to sequences with nucleotide misincorporations.

SI 16. Phenotype reconstruction

Ancient genomic sequencing can provide information about prehistoric individuals for traits that are not visible in fossil material. One suite of traits for which several genetic variants associated with different phenotypes have been identified is pigmentation. Light pigmentation supposedly arose after the appearance of anatomically modern humans outside of Africa, possibly as an adaptation to decreased UV radiation, but the exact timing of the appearance of these traits is unknown. Since the MA-1 individual represents the oldest anatomically modern human genome to date and shows a genetic affinity to northern Europeans, populations with the highest frequency of light pigmentation phenotypes, its pigmentation could be informative on the origin of this phenotype.

A set of 124 SNPs identified by Cerquira *et al*. (2011)[1] as informative on skin, hair and eye pigmentation was investigated. Following Cerquira *et al.*[1], competing hypotheses about these traits were tested using a majority rule estimate for the individual's phenotype. MA-1 individual was predicted as having relatively dark pigmentation for hair, skin and eyes. However, we note that this analysis is based on only 17 SNPs at low depth-of-coverage, and even with the full set of SNPs the method has limited prediction accuracy.

Among other frequently cited genetic variants that have been characterized in ancient genomes (e.g. Rasmussen *et al*. 2010[2]), MA-1 has the G allele at rs6152 that is associated with high risk for male pattern baldness. As expected, MA-1 also carries the ancestral C allele at rs4988235, and not the T allele that confers lactose persistence in western Eurasians. It does not have the allele associated with hair thickness and shovel-shaped incisors in Asians (rs3827760), while the position that is associated with dry earwax in East Asians (rs17822931) is not covered in the draft genome data. Results are presented in Table SI 14, along with pigmentation predictions for the Tyrolean Iceman[3] for comparison.

References for SI 16

1. Cerqueira, C.C.S. *et al*. Predicting *homo* pigmentation phenotype through genome data: From neanderthal to James Watson. *Am. J. Hum. Biol.* **24**, 705-709 (2012)

2. Rasmussen, M. *et al*. Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463**, 757-762 (2010)

3. Keller, A. *et al*. New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nat. Commun*. **3**, doi: 10.1038/ncomms1701 (2012)

Table SI 14 Pigmentation prediction in the MA-1 individual contrasted with the Tyrolean Iceman (Keller *et al.* 2012)[3]. The method is based on assessing a total of 124 SNPs associated with pigmentation traits (Cerquira *et al.* 2011)[1].

Phenotype	MA-1 (17 SNPs)	Iceman (24 SNPs)
darker skin vs fairer skin	darker skin (6/9)	fairer skin (8/15)
blue eyes vs non-blue eyes	Non-blue eyes (4/4)	blue eyes (7/9)
brown eyes vs non-brown eyes	brown eyes (2/3)	non-brown eyes (5/8)
green or blue eyes vs brown or black eyes	brown or black eyes (2/2)	brown or black eyes (3/3)
red hair vs non-red hair	-	red hair (2/3)
freckles vs non-freckles	freckles (1/1)	freckles (2/3)
blond hair vs non-blond hair	Non-blond hair (1/1)	blond hair (2/3)
brown hair vs non-brown hair	brown hair (5/7)	brown hair (8/11)
lighter brown or blond hair vs darker brown or black hair	darker brown or black hair (2/3)	lighter brown or blond hair (2/3)
blond or red hair vs brown hair	brown hair (7/7)	brown hair (11/13)
lighter brown hair vs darker brown hair	lighter brown hair (2/2)	lighter brown hair (2/3)
blond or red hair vs non-blond or non-red hair	-	blond or red hair (2/3)