Supplementary Information

Ancient human genomes suggest three ancestral populations for present-day Europeans

Supplementary Information 1

Sampling, Library Preparation and Sequencing

Alissa Mittnik*, Susanna Sawyer, Ruth Bollongino, Christos Economou, Dominique Delsate, Michael Francken, Joachim Wahl, Johannes Krause

* To whom correspondence should be addressed (amittnik@gmail.com)

Samples and extraction

Loschbour

The Mesolithic Loschbour sample stems from a male skeleton recovered from the Loschbour rock shelter in Heffingen, Luxembourg.

The skeleton we analyzed was excavated in 1935 by Nicolas Thill. The *in situ* find is not documented, but was described retrospectively by Heuertz (1950, 1969). According to his reports it seemed to be a primary burial, as the skeleton was lying on its back in a flexed position and with arms crossed over the chest. Based on morphological, radiological and histological data, the estimated age of death is 34 to 47 years (Delsate *et al.* 2009). Pathological finds are summarized in slight dorsal and lumbar vertebral osteoarthritic lesions, minimal unsystematized enthesopathies and an osteo-dental discharge fistula (Delsate *et al.* 2009). The skull was most likely decorated with ocher (Delsate *et al.* 2009). A second, cremated individual was discovered in a nearby pit (Toussaint *et al.* 2009). The archaeological layer in which the skeleton was found contained rich lithic assemblages, including microlithic artefacts (points with retouched and unretouched bases, points with bilateral retouch, an obliquely truncated point, a point with a slanted base and surface retouch, mistletoe points with surface retouch, a scalene triangle, narrow backed bladelets and a truncated bladelet with a narrow back), massive antler tools, perforated and burnt snails (*Bayania lactea*), and faunal remains from aurochs, red deer, wild boar, and roe deer. Additional excavations were carried out in 1981 and 2003 which revealed additional information on the stratigraphy (Gob 1982, Gob *et al.* 1984) and palaeoenvironment (Brou 2006).

The skeleton was AMS radiocarbon dated to 7,205 ± 50 years before present (BP) (OxA-7338; 6,220-5,990 cal BC; Toussaint *et al.* 2009).

The DNA extraction was performed on a molar (M12) sampled from the skull, pictured in Extended Data Fig. 1A, in as sterile as possible conditions in 2009. After sampling, the tooth was UV-irradiated, and the surface was removed and again irradiated with UV-light in the Palaeogenetics Laboratory in Mainz. Subsequently, the sample was pulverized in a mixer mill (Retsch).

The initial extraction was performed using 80 mg of tooth powder by a silica protocol after Rohland & Hofreiter (2007), resulting in 100 µl of extract (extract LB1, Table S1.1).

Two more extracts with a volume of 100 µl each were prepared from an additional 90 mg of tooth powder each following the protocol of Dabney *et al.* 2013 (extracts LB2 and LB3, Table S1.1).

Stuttgart

The Stuttgart sample stems from a female skeleton (LBK380, Extended Data Fig. 1B) that was excavated in 1982 at the site Viesenhäuser Hof, Stuttgart-Mühlhausen, Germany. The site reflects a long period of habitation starting from the earliest Neolithic to the Iron Age. The early Neolithic at this site is represented by a large number of well preserved burials belonging to the Linear Pottery

Culture (*Linearbandkeramik*, LBK), dated to 5,500-4,800 BC, as inferred from artifacts such as pottery associated with the graves of the female skeletons as well as surrounding graves (Kurz, in prep.). The Neolithic part of the graveyard separates into two large areas including burials from the early (area-2) and middle and late phases (area-1) of the LBK. The relative chronology of the burials from area-1 has been corroborated by calibrated radiocarbon dates of 5,100-4,800 BC (Stäuble 2005).

Based on morphology, Stuttgart (LBK380) is a female who died at an estimated age of 20 to 30 years. The skeleton derives from a grave (I-78, area-1) excavated among 83 others from area I of the cemetery and is well preserved but partially fragmented. The skeleton was buried in the characteristic way of the LBK, lying in a seated position on the right side. The burial was oriented from East-North-East to West-North-West with the skull facing north. Most of the body parts were represented (Burger-Heinrich, in prep.). Strontium isotope analysis suggests that the female was of local origin (Price *et al.* 2003). Noticeable pathological changes were present including multiple osseous lesions, compression fractures, and an angular kyphosis affecting several vertebrae. These may be due to a diagnosis of primary hyperparathyroidism in this individual (Zink *et al.* 2005).

For DNA analysis the M47 molar was removed. A total of 40 mg of tooth powder were taken from the inner part of the Stuttgart molar by a sterile dentistry drill in the clean room facilities of the University of Tübingen and extracted according to the protocol of Dabney *et al.* (2013) resulting in 100 µl of DNA extract (extract LBK1, Table S1.1).

Table S1.1: *Summary of extractions*

Extract	Individual	Tissue	Amount (mg)	Extraction protocol
LB1	Loschbour	Molar	80	Rohland & Hofreiter (2007)
LB2	Loschbour	Molar	90	Dabney *et al.* (2013)
LB3	Loschbour	Molar	90	Dabney *et al.* (2013)
LBK1	Stuttgart (LBK380)	Molar	40	Dabney *et al.* (2013)
MOT1	Motala 1	Molar	100	Yang *et al.* (1998)
MOT2	Motala 2	Molar	100	Yang *et al.* (1998)
MOT3	Motala 2	Skull	100	Yang *et al.* (1998)
MOT4	Motala 3	Molar	100	Yang *et al.* (1998)
MOT5	Motala 4	Molar	100	Yang *et al.* (1998)
MOT6	Motala 5	Molar	100	Yang *et al.* (1998)
MOT7	Motala 6	Molar	100	Yang *et al.* (1998)
MOT8	Motala 6	Skull	100	Yang *et al.* (1998)
MOT9	Motala 7	Skull	100	Yang *et al.* (1998)
MOT10	Motala 8	Molar	100	Yang *et al.* (1998)
MOT11	Motala 8	Skull	100	Yang *et al.* (1998)
MOT12	Motala 9	Molar	100	Yang *et al.* (1998)
MOT13	Motala 12	Molar	100	Yang *et al.* (1998)
MOT14	Motala 12	Maxilla	100	Yang *et al.* (1998)
MOT15	Motala 4170	Tibia	100	Yang *et al.* (1998)
MOT16	Motala MkA	Femur	100	Yang *et al.* (1998)

Motala

The Motala samples come from the site of Kanaljorden in the town of Motala, Östergötland, Sweden. The site was excavated between 2009 and 2013. The samples that we analyzed in the present study were retrieved in 2009 and 2010 (Extended Data Figs. 1C, 1D).

The human remains are part of ritual depositions that were made on a 14 × 14 meter stone-packing, constructed on the bottom of a small lake. The stone-packing was completely submerged and covered by at least 0.5m of water at the time of use. The ritual depositions include human bones: mostly skulls and fragments of skulls but also some stray bones from other parts of the body. The minimal number of individuals is inferred to be ten adults and one infant. The infant is the only individual that has bone representation from the entire body. Two of the skulls were mounted on wooden stakes still imbedded in the crania at the time of discovery. Apparently the skulls were put on display prior to the deposition in the lake. In addition to human bones, the ritual depositions also includes artifacts of antler, bone, wood and stone, animal carcasses/bones, as well as nuts, mushrooms and berries.

Direct dates on 11 human bones range between 7,013 ± 76 and 6,701 ± 64 BP (6,361-5,516 cal BC), with a twelfth outlier at 7,212 ± 109 BP. Dates on animal bones (N=11) and resin, bark and worked wood (N=6) range between 6,954 ± 50 and 6,634 ± 45 BP (5,898 - 5,531 cal BC). These dates correspond to a late phase of the Middle Mesolithic of Scandinavia.

Table S1.2: *Summary of libraries sequenced as part of this study*

Library	**From extract**	**Extract vol. in lib. (ul)**	**Library prep. Protocol**	**UDG treatment**	**Insert size fractionation**
ALB1	LB1	20	Meyer & Kircher (2010)	no	none
ALB2-10	LB1	28.5 (total)	Meyer *et al.* (2012)	yes	55-300bp
ALB11-12	LB2	25	Briggs *et al.* (2010)	yes	80-180bp
ALB13-14	LB3	25	Briggs *et al.* (2010)	yes	80-180bp
ALBK1	LBK1	5	Meyer & Kircher (2010)	no	none
ALBK2	LBK1	50	Briggs *et al.* (2010)	yes	70-180bp
AMOT1	MOT1	10	Meyer & Kircher (2010)	no	none
AMOT2	MOT2	10	Meyer & Kircher (2010)	no	none
AMOT3	MOT3	10	Meyer & Kircher (2010)	no	none
AMOT4	MOT4	10	Meyer & Kircher (2010)	no	none
AMOT5	MOT5	10	Meyer & Kircher (2010)	no	none
AMOT6	MOT6	10	Meyer & Kircher (2010)	no	none
AMOT7	MOT7	10	Meyer & Kircher (2010)	no	none
AMOT8	MOT8	10	Meyer & Kircher (2010)	no	none
AMOT9	MOT9	10	Meyer & Kircher (2010)	no	none
AMOT10	MOT10	10	Meyer & Kircher (2010)	no	none
AMOT11	MOT11	10	Meyer & Kircher (2010)	no	none
AMOT12	MOT12	10	Meyer & Kircher (2010)	no	none
AMOT13	MOT13	10	Meyer & Kircher (2010)	no	none
AMOT14	MOT14	10	Meyer & Kircher (2010)	no	none
AMOT15	MOT15	10	Meyer & Kircher (2010)	no	none
AMOT16	MOT16	10	Meyer & Kircher (2010)	no	none
AMOT17	MOT1	15	Briggs *et al.* (2010)	yes	none
AMOT18	MOT2	15	Briggs *et al.* (2010)	yes	none
AMOT19	MOT4	15	Briggs *et al.* (2010)	yes	none
AMOT20	MOT5	15	Briggs *et al.* (2010)	yes	none
AMOT21	MOT7	15	Briggs *et al.* (2010)	yes	none
AMOT22	MOT12	15	Briggs *et al.* (2010)	yes	none
AMOT23	MOT13	15	Briggs *et al.* (2010)	yes	none

Teeth from nine of the better-preserved skulls were selected for DNA analysis, as well as a femur and a tibia (Motala MkA and 4170, from the first two human bones found in 2009). Extraction of the

samples from Motala took place in the clean-room facilities of the Ancient DNA laboratory at the Archaeological Research Laboratory, Stockholm University. Bone powder was removed from the inner parts of the bones or teeth with a sterile dentistry drill and extracted according to a protocol by Yang *et al.* (1998) resulting in 16 extracts (MOT 1 to 16, Table S1.1).

Library Preparation

For screening and mtDNA capture, libraries for all samples were prepared using either double- or single-stranded library preparation protocols (Table S1.2) (Meyer & Kircher 2010; Meyer *et al.* 2012). For large scale shotgun sequencing, additional libraries were produced including a DNA repair step with Uracil-DNA-glycosylase (UDG) and endonuclease VIII (endo VIII) treatment (Briggs *et al.* 2010). Libraries ALB 2-14 and ALBK1 were furthermore size fractionated on a PAGE gel according to Meyer *et al.* 2012 (Table S1.2).

Shotgun Sequencing

All non-UDG-treated libraries were random shotgun sequenced. For libraries ALB1 and AMOT 1, 2, 3, 4, 6, 9, and 12, sequencing was carried out on an Illumina Genome Analyzer IIx with 2×76 + 7 cycles. For library ALBK1 we carried out the sequencing on an Illumina MiSeq with 2×150 + 8 + 8 cycles. In all cases, we followed the manufacturer's protocol for multiplex sequencing.

Raw reads were analyzed as described in Kircher (2012) and were mapped with BWA 0.6.1 (Li & Durbin 2009) to the human reference genome (hg19/GRCh37/1000Genomes) in order to calculate the fraction of endogenous human DNA. After duplicate removal, 0.82% to 66.4% of reads were estimated to map to the human reference genome with a mapping quality of at least > 30 (Table S1.3).

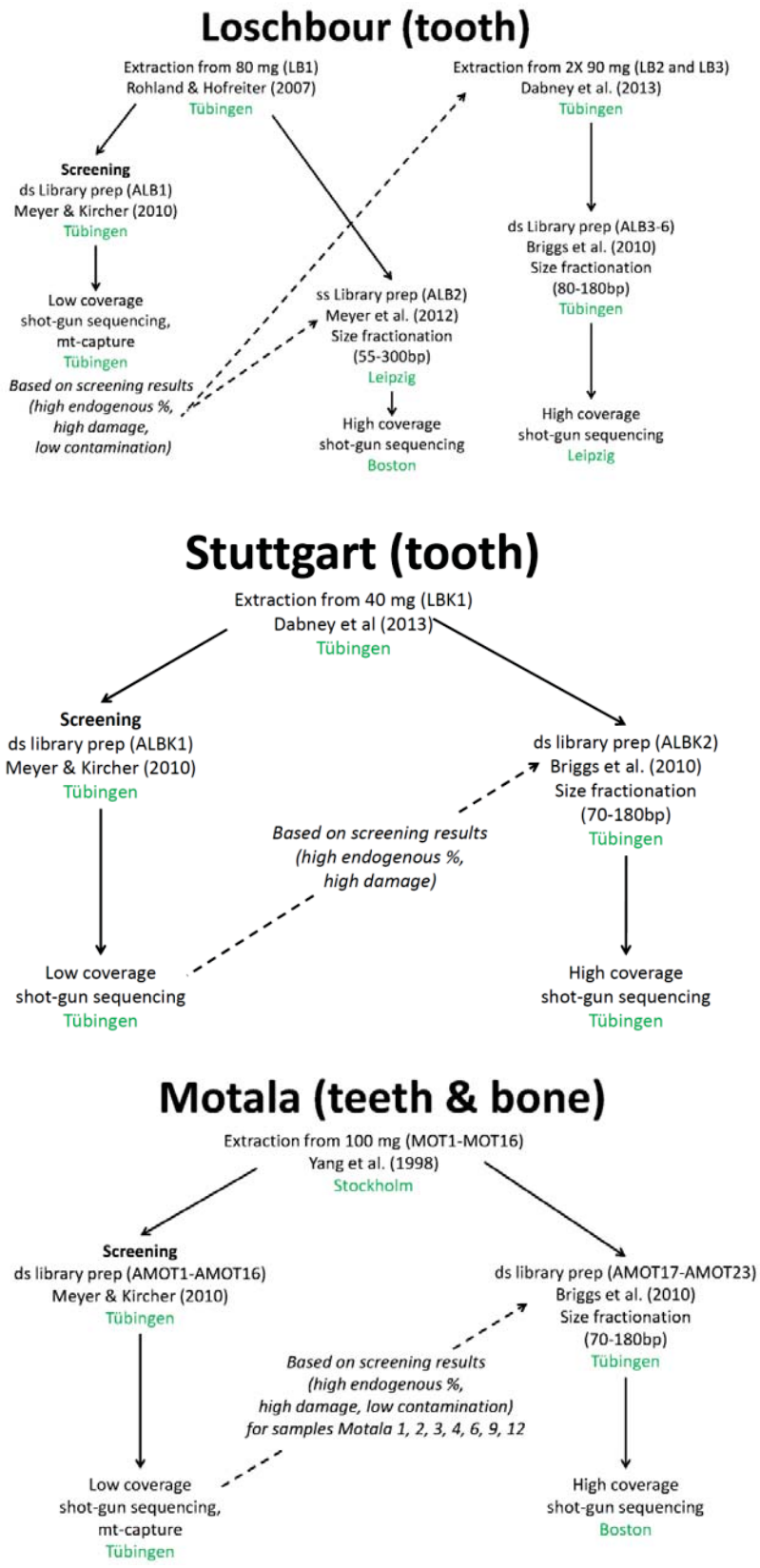
Based on the results, the extracts LB 1-3, LBK1, MOT 1, 2, 4, 5, 7, 12 and 13—representing individuals Loschbour, Stuttgart, and Motala 1, 2, 3, 4, 6, 9, and 12, respectively—were chosen for UDG-treatment and possible further deep sequencing.

Table S1.3: *Summary of whole-genome deep sequencing runs*

Library	Pooled	No. lanes	Read length	Facility
ALB2	no	3	50 bp	HMS, Boston
ALB2	no	3	100 bp	Illumina, San Diego
ALB3-10	LB Pool 1	5	100 bp	Illumina, San Diego
ALB11-14	LB Pool 2	8	101 bp	MPI, Leipzig
ALBK2	no	8	101 bp	MPI, Tübingen
AMOT17, 18, 23	Motala Pool 1	4	100 bp	Illumina, San Diego
AMOT19-22	Motala Pool 2	4	100 bp	Illumina, San Diego
AMOT23	no	8	100 bp	Illumina, San Diego

The UDG treated library ALB2 was sequenced on 3 Illumina HiSeq 2000 lanes with 50-bp single-end reads in the Harvard Medical School Biopolymers Facility, followed by 3 Illumina HiSeq 2000 lanes of 100-bp paired-end reads at Illumina, San Diego. We also carried out sequencing at Illumina, San Diego of 5 HiSeq 2000 lanes of 100-bp paired-end reads of pooled libraries ALB3-10.

Figure S1.1: Visualization of sample preparation process. (Top) *Loschbour,* ***(Middle)*** *Stuttgart and* ***(Bottom)*** *Motala. The responsible research group for each step is marked in green.*



We prepared UDG and endo VIII damage repaired libraries (USER-enzyme; Briggs *et al.* 2010) for ALB 3-6. We then size fractionated these libraries to an insert size of about 80-180bp. We performed 8 lanes of shotgun sequencing of these libraries on an Illumina HiSeq 2000 at the Max Planck Institute for Evolutionary Anthropology in Leipzig. Sequences were produced using 101-bp paired-end reads using CR2 forward (5' – TCTTTCCCTACACGACGCTCTTCCGATCTGTCT) and CR2 reverse (5' – GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTCT) custom primers. In addition, seven cycles were sequenced for a P7 index using the P7 Illumina Mulitplex primer. The P5 index was not sequenced. The instructions from the manufacturers were followed for multiplex sequencing on the HiSeq platform with a TruSeq PE Cluster Kit v3 - cBot – HS cluster generation kit and a TruSeq SBS Kit v3 sequencing chemistry. An indexed control library of ϕX 174 was spiked into each library prior to sequencing, contributing to 0.5% of the sequences from each lane.

We also prepared UDG-treated libraries for the Stuttgart sample, and size fractionated them to an insert size of about 70-180bp. ALBK2 was sequenced on 8 HiSeq 2000 lanes and 101-bp paired-end reads plus seven cycles for a P7 index using the P7 Illumina Mulitplex sequencing primer at the Max Planck Institute for Developmental Biology in Tübingen. Instructions from the manufacturers were followed using a TruSeq PE Cluster Kit v3 - cBot – HS cluster generation kit and a TruSeq SBS Kit v3 sequencing chemistry.

The UDG-treated libraries for Motala (AMOT17-23) were sequenced on 8 HiSeq 2000 lanes of 100-bp paired-end reads, with 4 lanes each for two pools (one of 3 individuals and one of 4 individuals), through contract sequencing at Illumina, San Diego of 100-bp paired-end reads. We also sequenced an additional 8 HiSeq 2000 lanes of AMOT23 at Illumina, San Diego through contract sequencing. This was the library with the highest percentage of endogenous human DNA (from Motala12).

A visual overview of sample processing, including library preparation, capture methods and sequencing results is shown in Figure S1.1.

References

Briggs, A.W., Stenzel, U., Meyer, M., Krause, J., Kircher, M., and Pääbo, S.(2010). Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. Nucleic Acids Res. 38, e87.

Brou, L. 2006, Abri d'Heffingen-Loschbour (G.-D. de Luxembourg), sondages programmés. Rapport d'archéologie programmée n° 9. Service d'archéologie préhistorique, Rapport interne MNHA. 14 p. 42 fig. 2 annexes. MNHA 2006.

Dabney, J., Knapp, M., Glocke, I., Gansauge, M.T., Weihmann, A., Nickel, B., Valdiosera, C., García, N., Pääbo, S., Arsuaga, J.L., Meyer, M. (2013). Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proc Natl Acad Sci USA. 2013 Sep 24;110(39):15758-63. doi: 10.1073/pnas.1314445110.

Delsate, D., Guinet J.-M. & Saverwyns, S. 2009, De l'ocre sur le crâne mésolithique (haplogroupe U5a) de Reuland-Loschbour (Grand-Duché de Luxembourg) ? Bull. Soc. Préhist. Luxembourgeoise 31, 2009, 7-30.

Fu, Q., Mittnik, A., Johnson, P. L., Bos, K., Lari, M., Bollongino, R. *et al.* (2013). A Revised Timescale for Human Evolution Based on Ancient Mitochondrial Genomes. Current Biology, 23(7), 553-559.

Gob, A. 1982, L'occupation mésolithique de l'abri du Loschbour près de Reuland (G. D. de Luxembourg). *In :* Le Mésolithique entre Rhin et Meuse. Actes du Colloque sur le Paléolithique

supérieur final et le Mésolithique dans le Grand-Duché de Luxembourg et dans les régions voisines (Ardenne, Eifel, Lorraine) tenu à Luxembourg le 18 et 19 mai 1981. Edité par A. GOB et F. SPIER. Publication de la Société Préhistorique Luxembourgeoise 1982, 91-118.

Gob, A., Heim, J., Spier, F & Ziesaire, P. 1984, Nouvelles recherches à l'abri du Loschbour près Reuland (G.-D. de Luxembourg). Bull. Soc. Préhist. Luxembouregoise 6 (1984), 87-99.
Heuertz, M. 1950, Le gisement préhistorique n° 1 (Loschbour) de la vallée de l'Ernz-Noire (G.-D. de Luxembourg). Ed. Musée d'Histoire Naturelle, Luxembourg 1950. Extrait des « Archives » Tome 19, Nouvelle série (Année du Centenaire 1950) de l'Institut Grand-ducal de Luxembourg, Section des Sciences naturelles, physiques et mathématiques, N.S., 19, 1950, 409-441.

Heuertz, M. 1969, Documents préhistoriques du territoire luxembourgeois. Le milieu naturel. L'homme et son œuvre. Fascicule 1. Publications du Musée d'Histoire naturelle et de la Société des Naturalistes luxembourgeois. Fasc. 1, Luxembourg 1969, 295 p., 190 fig.

Kircher M. (2012) Analysis of high-throughput ancient DNA sequencing data. Methods Mol Biol. 840: 197-228. doi: 10.1007/978-1-61779-516-9_23.

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

Meyer, M. and Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harbor Protoc., 10.1101/pdb.prot5448.

Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S. *et al.* (2012). A High-Coverage Genome Sequence from an Archaic Denisovan Individual. Science, 338(6104), 222-226.

Rohland, N. and Hofreiter, M. (2007): Ancient DNA extraction from bones and teeth. Nat Protoc 2(7):1756–1762.

Stäuble, H. (2005). Häuser und absolute Datierung der ältesten Bandkeramik. (Universitätsforsch. Prähist. Arch. 117). Bonn: Habelt.

Toussaint, M., Brou, L., Spier, F. & Le Brun-Ricaliens, F. 2009, Le gisement mésolithique d'Heffingen-Loschbour (Grand-Duché de Luxembourg). Une sépulture à incinération inédite de la culture RMS : implications anthropologiques, radiométriques et archéologiques. *In :* Ph. Crombé (dir.), Chronology and Evolution in the Mesolithic of North-West Europe. Congrès international de Bruxelles, 30-31 mai et 1er juin 2007. Cambridge Scholars Publishing.

Yang D.Y., Eng B., Waye J.S., Dudar J.C., and Saunders S.R. (1998). Improved DNA extraction from ancient bones using silica- based spin columns. American Journal of Physical Anthropology 105:539-543.

Zink, A., Panzer, S., Fesq-Martin, M., Burger-Heinrich, E., Wahl, J., Nerlich, A.G. (2005). Evidence for a 7000-year-old case of primary hyperparathyroidism. Journal of the American Medical Association 293: 40–42.

Supplementary Information 2

Processing of sequencing data and estimation of heterozygosity

Gabriel Renaud*, Cesare de Filippo, Swapan Mallick, Janet Kelso and Kay Pruefer

* To whom correspondence should be addressed (gabriel_renaud@eva.mpg.de)

Overview

This note describes the processing of the sequence data for the Loschbour, Stuttgart and Motala samples. It also describes the estimation of heterozygosity for the high coverage Stuttgart and Loschbour individuals. For Stuttgart, heterozygosity was estimated to be higher than in any of present-day 15 non-African and lower than in 10 present-day Africans. For Loschbour, heterozygosity was estimated to be lower than in any of 25 present-day humans.

Sequencing data

All ancient DNA (aDNA) libraries were sequenced on the Illumina HiSeq platform. Base-calling was carried out using the default Illumina basecaller, Bustard, except where noted. The following data were generated (summarized in Table S2.1):

Loschbour:

1. Four double-stranded libraries (ALB11-14) were sequenced for 101-cycles, paired-end, on a HiSeq 2500 platform (8 lanes). Base-calling was performed using freeIbis[1].
2. Nine single-stranded libraries (ALB2-10) were sequenced for 100-cycles, paired-end, on a HiSeq 2000 platform. This consisted of 3 lanes for ALB2 and 5 lanes for a pool of ALB3-10.

Stuttgart:

A double-stranded library (ALBK2) was sequenced for 101-cycles, paired-end, on a HiSeq 2000 platform (8 lanes).

Motala:

Seven double-stranded libraries (AMOT17-23) were sequenced for 100-cycles, paired-end, on a HiSeq 2000 platform (8 lanes). Motala12 (AMOT23), the sample with the highest percentage of endogenous DNA, was then sequenced on a further 8 lanes.

Processing of sequencing data prior to genotyping

Ancient DNA molecules are often short enough that the paired-end reads carry the flanking sequencing adaptors at the ends. The reads were therefore pre-processed to trim adaptors and to merge overlapping paired-end reads using the merger program in aLib[2] (*-mergeoverlap* option). The merged sequences and unmerged read pairs were then mapped. Sequences with more than five bases with quality less than 10 were flagged as "QC failed" and were removed.

The sequences from Loschbour and Stuttgart were mapped to the *hg19* genome assembly (1000 Genomes version) using BWA[3] version 0.5.10, parameters "-n 0.01 -o 2", with the seed disabled. Sequences that were merged and pairs that were flagged as properly paired were retained for analysis. The mappings were sorted and duplicates were removed using bam-rmdup2 version 0.4.9. Indel realignment was performed using GATK[4] version 1.3-25. To restore the MD field in the BAM files, "samtools fillmd" was used (samtools[5] 0.1.18). Sequences produced from libraries prepared using the single-stranded protocol still carry uracils at the first or last two bases of the molecules. These are read as thymines during sequencing, and cannot be identified or corrected using metrics such as base quality. Since they can influence variant calling we reduced to a PHRED score of 2 the base quality of

any 'T' in the first base or last two bases of sequence reads from all single-stranded libraries. Similarly, sequences produced from libraries prepared using the double-stranded protocol may carry uracils at the first base causing C-to-T changes and G-to-A changes on the last base. Qualities of thymines in the first and adenines in the last base were reduced to a PHRED score of 2[6].

The seven Motala samples had to be treated slightly differently. Initial light shotgun sequencing of seven Motala libraries was performed to determine candidate libraries for deeper sequencing. The samples were sequences as a pool, so we de-multiplexed the data by searching among the sequences for ones that had no more than one mismatch compared with each of the expected P7 and P5 indices for the seven samples. Reads were stripped of adapters, merged using SeqPrep[7], and aligned with BWA[3] version 0.5.10, with parameters "-n 0.01 –o 2" (seed disabled). Duplicates were removed using samtools[5] 0.1.18. PCA indicated that the Motala samples were relatively homogenous in ancestry and we therefore merged the data for all of the samples except for Motala3 and Motala12 (using samtools *merge*5) to increase coverage for population genetic analysis (labeled 'Motala_merge' in Fig. 1B).

Comparisons of the endogenous rates for all Motala samples indicated that the library from Motala12 had the highest percentage of endogenous DNA, and thus a further eight lanes of sequencing were generated for this individual.

Table S2.1 reports summary statistics for all the libraries we sequenced. Figure S2.1 reports base-specific substitution patterns per library.

Table S2.1. Sequencing results by library for Loschbour, Stuttgart and Motala

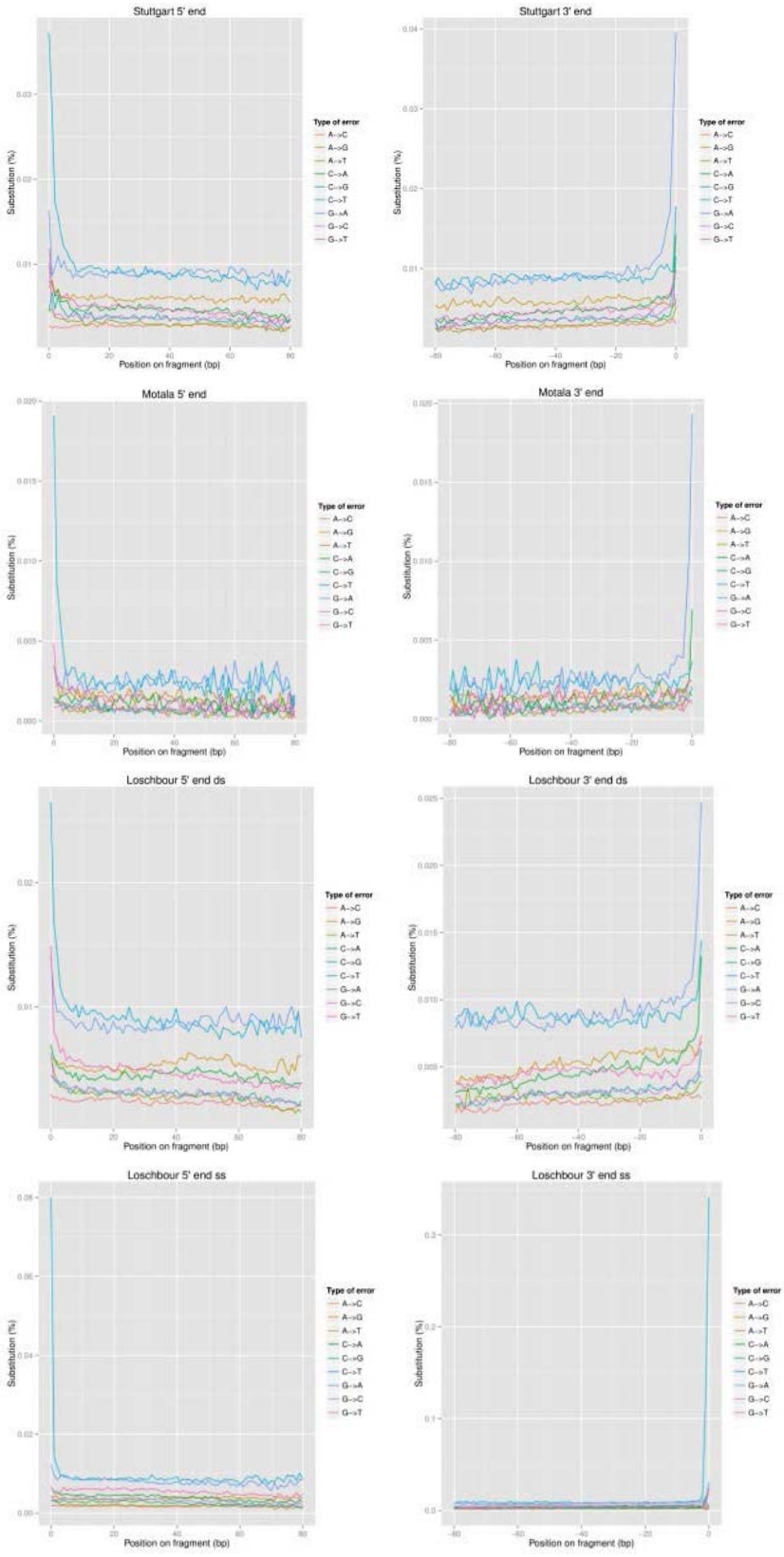
Sample	Library ID	Library type	Mapped sequences	Mean insert size (bp)	Std. Dev. in insert size (bp)	Genome coverage
Loschbour1	ALB11	Double strand + UDG	93,342,792	87	23	2.8
Loschbour2	ALB12	Double strand + UDG	111,474,060	80	22	3.1
Loschbour3	ALB13	Double strand + UDG	146,593,852	78	23	4.0
Loschbour4	ALB14	Double strand + UDG	161,736,672	80	23	4.5
Loschbour	ALB2-10	Single strand + UDG	345,350,969	61	20	7.2
Stuttgart	ALBK2	Double strand + UDG	788,244,122	70	17	19.1
Motala1	AMOT17	Double strand + UDG	8,050,873	68	29	0.18
Motala2	AMOT18	Double strand + UDG	6,670,241	70	31	0.15
Motala3	AMOT19	Double strand + UDG	23,622,338	73	32	0.55
Motala4	AMOT20	Double strand + UDG	3,369,460	64	29	0.070
Motala6	AMOT21	Double strand + UDG	1,032,460	71	31	0.024
Motala9	AMOT22	Double strand + UDG	484,149	64	27	0.010
Motala12	AMOT23	Double strand + UDG	94,818,771	73	32	2.4

Note: "Mapped" refers to the number of merged and properly paired sequences after duplicate removal.

Diploid genotyping

For the Loschbour and Stuttgart high coverage individuals, diploid genotype calls were obtained using the Genome Analysis Toolkit (GATK)[8] version 1.3-25, using the parameters: "--output_mode EMIT_ALL_SITES --genotype_likelihoods_model BOTH --baq OFF". Because GATK does not call heterozygous sites in cases in which neither allele matches the reference, a second round of genotyping was carried out, providing as input a modified reference sequence that carried the bases called in the first round of genotyping. The genotype calls from both rounds were then combined to obtain a final variant call format (VCF) file.

Figure S2.1: Substitution patterns for Loschbour, Stuttgart and Motala 12 (measured on chromosome 21). *Single- and double-stranded Loschbour libraries are reported separately.*



Estimation of heterozygosity

Two approaches were used to estimate heterozygosity:

1. *mlRho*[9], which estimates heterozygosity as the maximum likelihood of the population mutation parameter (θ) from high-coverage data of one individual, assuming an infinite sites model of mutation. The program also estimates the sequencing error rate per site (ε).
2. *GATK*. The GATK genotype calls are viewed as correct, and the number of called heterozygous sites divided by the total number of screened nucleotides is interpreted as the heterozygosity.

Heterozygosity was estimated in the high-coverage genome sequences from 29 individuals: 25 diverse present-day humans, Altai Neandertal, Denisova, Stuttgart and Loschbour. This is the same dataset described in the paper on the high coverage Neandertal genome, here supplemented by Stuttgart and Loschbour[10]. Analysis was restricted to ~629 million sites on the autosomes that passed the following filters in all 29 individuals (the filters are described in more detail in ref. 10):

1. Fall in the most stringent mappability track (Map35_100%): positions where all overlapping 35mers align only to one location in the genome allowing for up to one mismatch.
2. A mapping quality (MQ) of 30.
3. In the 2.5% - 97.5% interval of the coverage distribution specific to each sample. For the ancient samples, coverage is computed by binning sites according to their local GC content (i.e. the number of GC bases in a 51 bp window centered at the site).
4. Do not overlap insertion / deletion polymorphisms (indels).
5. Not a simple repeat as specified by the UCSC Tandem Repeat Finder track[11] for *hg19*.

Table S2.2: Heterozygosity and error estimates per 10,000 screened sites

Sample*	***mlRho***	***mlRho* error (ε)**	**GATK**	***mlRho*/GATK ratio**
Altai Neandertal	1.68	11.20	1.75	0.96
Denisova	1.82	7.44	2.14	0.85
Loschbour	4.75	9.43	6.62	0.72
Karitiana_B	4.99	0.99	5.52	0.90
Papuan_B	5.02	0.95	5.98	0.84
Mixe_B	5.85	1.60	6.12	0.96
Karitiana_A	5.87	2.66	5.76	1.02
Australian1_B	6.03	0.98	6.59	0.91
Papuan_A	6.03	2.47	6.38	0.94
Australian2_B	6.42	0.97	6.66	0.96
Dai_A	6.46	2.53	7.44	0.87
Han_B	6.62	0.96	7.23	0.92
Dai_B	6.67	1.22	7.19	0.93
Sardinian_B	6.69	1.07	7.34	0.91
French_B	6.92	1.18	7.58	0.91
Han_A	7.04	2.48	7.45	0.95
Sardinian_A	7.07	2.47	7.79	0.91
French_A	7.38	2.82	7.81	0.94
Stuttgart	7.42	10.20	10.59	0.70
Dinka_B	8.26	2.50	9.68	0.85
Mandenka_B	9.14	1.19	10.01	0.91
Dinka_A	9.23	2.78	9.99	0.92
Mbuti_B	9.35	1.03	10.09	0.93
Mbuti_A	9.38	2.35	10.23	0.92
San_B	9.44	1.14	10.21	0.92
Mandenka_A	9.50	2.66	10.31	0.92
Yoruba_B	9.50	1.03	10.06	0.94
San_A	9.64	2.90	10.69	0.90
Yoruba_A	9.78	2.70	10.18	0.96

* The suffix for the 25 present-day samples indicates whether the individual is from the A or B panel.

GATK calls were extracted from the VCF files[8]. GATK heterozygosity was defined as the number of heterozygous sites divided by the number of bases screened.

mlRho was run directly on the BWA alignments, restricting to sites that passed the filters above and additionally restricting to DNA sequencing data with a minimum base quality of 30.

Inspection of Table S2.2 indicates that the *mlRho* estimates are smaller than the GATK estimates for nearly all samples. However, the reduction below one is most marked for Stuttgart and Loschbour:

0.92	Mean of 25 present-day humans
0.96	Altai Neandertal
0.84	Denisova
0.72	Loschbour
0.70	Stuttgart

The discrepancy between GATK and *mlRho* estimates is plausibly due to a higher error rate in the Stuttgart and Loschbour diploid genotype calls due to these two genomes' lower sequencing coverage. Specifically, the Stuttgart and Loschbour sequencing coverage is ~ ~20× compared with >30× for most other samples. The GATK estimates do not correct for the genotyping error that occurs in the context of low coverage, and hence are likely to produce artifactual overestimates of heterozygosity.

It is important to note that although the diploid genotype calls for both Stuttgart and Loschbour have a higher error rate than for the other genomes, these error rates are not likely to be sufficient to bias the analyses of population history reported in this study. The reason for this is that the Loschbour and Stuttgart diploid genotypes are used in this study largely for the purpose of determining allelic state at sites that are already known to be polymorphic in present-day humans: SNPs that are part of the Affymetrix Human Origins array (SI 4). At these sites, the probability of polymorphism is much higher than the likely error rate of 1/1000 to 1/10000 in the Stuttgart and Loschbour data, and thus error does not contribute much to the observed variability of the inferred allelic state at these sites.

Using the *mlRho* estimates of heterozygosity which are likely to be more accurate than those from GATK because *mlRho* co-estimates and correct for error, Loschbour is inferred to have an average of 4.75 heterozygous sites per 10,000 base pairs. This is lower than in any of 25 diverse present-day human samples to which Loschbour was compared although it is still about three times higher than the heterozygosity reported for the Denisovan and Altai Neandertal (Table S2.2). In contrast, Stuttgart has 7.42 heterozygous sites per 10,000 base pairs. This is higher than the heterozygosity measured any of 15 diverse non-Africans, although only slightly higher than the most diverse present-day non-African in the panel (French_A at 7.38 heterozygous sites per 10,000 base pairs) (Table S2.2).

Inferring allelic state for the low coverage ancient genomes

Many of the analyses reported in this study include ancient samples that had too-low sequencing coverage to permit confident diploid genotype calls. To analyze these samples in conjunction with genotyping data from the Affymetrix Human Origins array (SI 4), a single allele was picked at random for each individual from each site in the genome for which there was a high quality sequence. That allele was then used to represent that individual at that nucleotide position (the individual was treated as homozygous there). This procedure has the effect of (artifactually) inferred a high level of genetic drift on the lineage specific to the individual. However, it is not expected to induce correlations in drift with other samples, and thus it is not expected to bias inferences about population relationships.

The ancient DNA sequences to whom this procedure was applied are listed in Table S2.3, and discussed in more detail below:

(1) *Motala:* The number of Human Origins array SNPs for which there was sequencing coverage after this procedure was 352,966 for Motala_merge and 411,453 for Motala12.

(2) *Swedish farmers and hunter-gathers.* BAM files mapped to *hg19* were downloaded from ref. 12 for one Swedish Neolithic farmer (Gök4 in ref. 9; Skoglund_farmer in this paper), and three

Swedish Neolithic hunter-gatherers (Ajv52, Ajv70 and Ire8; combined as Skoglund_mergeHG in this paper). The number of SNPs with coverage after this procedure was 4,548 for Skoglund_farmer and 18,261 for Skoglund_mergeHG.

(3) *Iceman:* The *hg18*-mapped genotype calls for this individual were downloaded from the VCF file reported by ref. 13. liftOver[11] was used to convert the coordinates to *hg19*. There was coverage on 518,229 Human Origins array SNPs after this procedure.

(4) *Iberian Mesolithic hunter-gatherers:* BAM files mapped to *hg18* were downloaded from ref. 14 for two Iberian Mesolithic hunter-gathers from the La Braña site. liftOver[15] was used to convert the coordinates to *hg19*. Because of the extremely low coverage of the data, no additional filtering was applied with the goal of maximizing the number of retained SNPs. The number of SNPs for which there was sequencing coverage after applying this procedure was 9,868 for La Braña 1, and 4,525 for La Braña 2.

(5) *Upper Paleolithic Siberians*: BAM files mapped to *hg19* were downloaded from ref. 16. The number of SNPs for which there was sequencing coverage after this procedure was 427,211 for MA1 and 92,486 for AG2.

Table S2.3: Number of autosomal SNPs with an allele call

Sample	**SNPs†**
Motala12	411,453
Motala_merge*	352,966
Skoglund_farmer	4,548
Skoglund_mergeHG	18,261
Tyrolean Iceman	518,229
La Braña 1	9,868
La Braña 2	4,525

* Motala_merge includes reads from five Motala samples (all except Motala3 and Motala12).
† The maximum number of SNPs was the 594,924 autosomal sites in the Human Origins genotyping data (SI 4).

References

[1] Renaud, G., Kircher, M., Stenzel, U. & Kelso, J. freeIbis: an efficient basecaller with calibrated quality scores for Illumina sequencers. *Bioinformatics* **29**, 1208-1209, doi:10.1093/bioinformatics/btt117 (2013).

[2] https://github.com/grenaud/aLib master branch with revision # : 3c552c66da8f049324485122fefaf2297a072930

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

[4] McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

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

[6] decrQualDeaminated and decrQualDeaminatedDoubleStranded from https://github.com/grenaud/libbam

[7] https://github.com/jstjohn/SeqPrep

[8] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297-303.

[9] Haubold, B., Pfaffelhuber, P. & Lynch, M. mlRho - a program for estimating the population mutation and recombination rates from shotgun-sequenced diploid genomes. *Molecular ecology* 19 Suppl 1, 277-284, doi:10.1111/j.1365-294X.2009.04482.x (2010).

[10] Prüfer, K. et al. (2013) The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* Advance online publication December 18 2013.

[11] http://genome.ucsc.edu

[12] Skoglund P, Malmström H, Raghavan M, Storå J, Hall P, Willerslev E, Gilbert MT, Götherström A, Jakobsson M (2012) Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. Science 336, 466-9.

[13] Keller A, Graefen A, Ball M, Matzas M, Boisguerin V, Maixner F, Leidinger P, Backes C, Khairat R, Forster M, Stade B, Franke A, Mayer J, Spangler J, McLaughlin S, Shah M, Lee C, Harkins TT, Sartori A, Moreno-Estrada A, Henn B, Sikora M, Semino O, Chiaroni J, Rootsi S, Myres NM, Cabrera VM, Underhill PA, Bustamante CD, Vigl EE, Samadelli M, Cipollini G, Haas J, Katus H, O'Connor BD, Carlson MR, Meder B, Blin N, Meese E, Pusch CM, Zink A (2012) New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. Nat Commun. 3, 698

[14] Sánchez-Quinto F, Schroeder H, Ramirez O, Avila-Arcos MC, Pybus M, Olalde I, Velazquez AM, Marcos ME, Encinas JM, Bertranpetit J, Orlando L, Gilbert MT, Lalueza-Fox C (2012) Genomic affinities of two 7,000-year-old Iberian hunter-gatherers. Curr Biol. 22, 1494-9.

[15] http://genome.ucsc.edu/util.html

[16] Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, Rasmussen S, Stafford Jr TW, Orlando L, Metspalu E, Karmin M, Tambets K, Rootsi S, Mägi R, Campos PF, Balanovska E, Balanovsky O, Khusnutdinova E, Litvinov S, Osipova LP, Fedorova SA, Voevoda MI, Degiorgio M, Sicheritz-Ponten T, Brunak S, Demeshchenko S, Kivisild T, Villems R, Nielsen R, Jakobsson M, Willerslev E (2013) Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature Advance Online Publication November 20.

Supplementary Information 3
Ancient DNA authenticity

Alissa Mittnik*, Gabriel Renaud, Qiaomei Fu, Janet Kelso and Johannes Krause

* To whom correspondence should be addressed (amittnik@gmail.com)

Overview
This note describes the analyses that were performed to test the authenticity of the ancient DNA obtained from each of the ancient human samples. Contamination estimates were carried out for the mitochondrial DNA as well as for nuclear DNA sequences.

To identify suitable ancient human samples for deep sequencing, libraries for targeted mtDNA capture from Loschbour and all Motala individuals were prepared without the use of Uracil DNA glycosylase (UDG) in order to preserve DNA damage patterns that are an indication of authentic ancient DNA (Krause *et al.* 2010).

MtDNA capture, sequencing and processing was performed as described in SI 4. DNA extracts that showed high proportions of apparently authentic mtDNA were used for preparation of UDG-treated libraries for deep sequencing as described in SI 1. The mtDNA contamination rate for the deeply sequenced shotgun data from UDG-treated libraries was estimated by direct comparison to the mtDNA consensus from the targeted mtDNA enrichment.

No mtDNA capture was performed for Stuttgart. For this sample, deep sequencing data was used to analyze DNA damage patterns from a non-UDG-treated library (ALBK1). The mtDNA contamination estimate was obtained from high coverage shotgun data from a UDG-treated library (ALBK2).

Assessment of ancient DNA authenticity
Authenticity of aDNA from ancient human DNA extracts was assessed as part of the screening procedure described in SI 1. To assess authenticity the following criteria were applied.

1. Consistency of reads mapping to the mitochondrial genome consensus sequence (summarized in Stoneking & Krause 2011), showing that the majority of reads (>95%) derive from a single biological source.
2. Presence of aDNA-typical C-to-T damage patterns at the 5'-ends of DNA fragments, caused by post-mortem miscoding lesions (summarized in Stoneking & Krause 2011).
3. Plausibility of mitochondrial sequences in the broader context of the human mitochondrial phylogeny and contemporary population diversity, e.g. branch shortening, due to missing substitutions in ancient mtDNA (Fu *et al.* 2013; see also SI 4.)
4. In the case of the male sample, Loschbour, an absence of polymorphic sites on chromosome X (Rasmussen *et al.* 2011).
5. A maximum-likelihood-based estimate of autosomal contamination for Loschbour and Stuttgart that uses variation at sites that are fixed in the 1000 genomes humans to estimate error, heterozygosity and contamination (Fu *et al.*, in preparation).

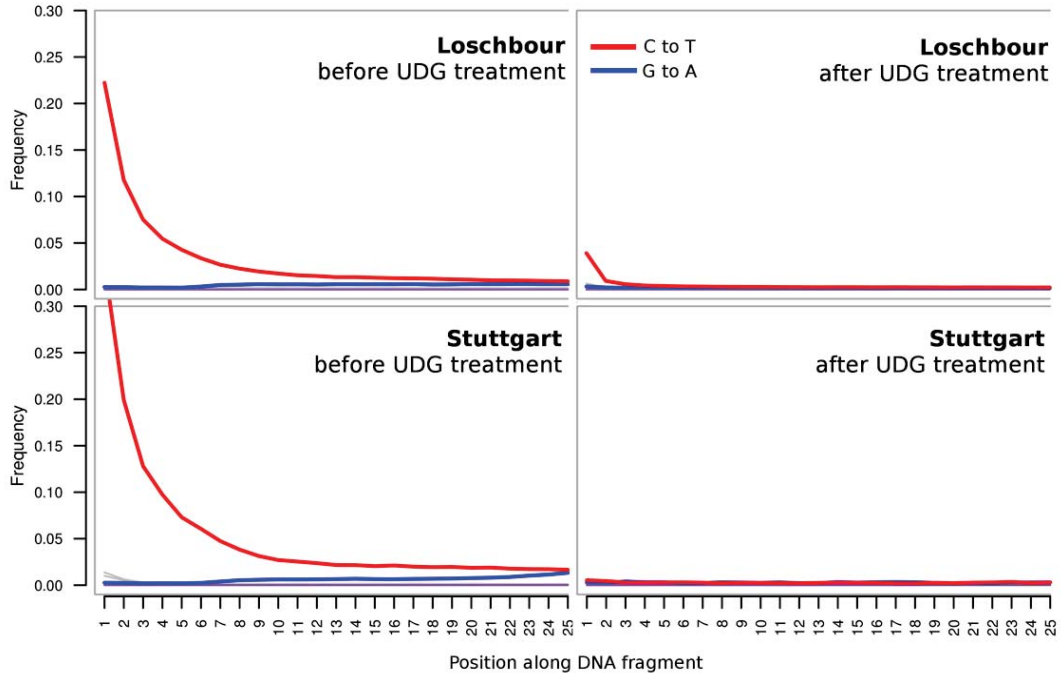
MtDNA contamination estimate and damage patterns for non-UDG-treated libraries
For sample screening, non-UDG-treated libraries from 17 ancient humans were used for estimating mtDNA contamination levels as described in Green *et al.* (2008). In total 8 of 17 samples show 5% or less inconsistent fragments (Table S3.1), suggesting that the DNA largely originated from a single biological source. Using a recently published Bayesian approach that compares the read sequences to

a set of 311 modern human mtDNAs and checks for consistency among the reads (Fu *et al.* 2013), similar results are obtained for those 8 samples (Table S3.1).

The percentage of C-to-T changes at the 5'-ends of endogenous mtDNA fragments from libraries that were non-UDG-treated was estimated using mapDamage2.0 (Jónsson *et al.* 2013) (the exception is ALBK1, where nuclear DNA from shotgun data was used to estimate damage patterns). Figure S3.1 shows the difference in damage patterns for libraries from Loschbour and Stuttgart with and without UDG-treatment. Based on previous evidence that samples older than 100 years typically have at least 20% deamination at the 5' ends (Sawyer *et al.* 2012), only samples that show more than 20% damage were considered as good candidates for harboring authentic ancient DNA (Table S3.1). All 8 samples that show internally consistent mtDNA show more than 20% damage at the 5'-ends and therefore meet the criteria of aDNA authenticity for further processing. Stuttgart was considered authentic due to its high level of DNA damage of more than 35% (Fig. S3.1).

In total 9 samples were shotgun sequenced: Loschbour, Stuttgart, and Motala 1, 2, 3, 4, 6, 9 and 12.

Figure S3.1: Frequency of nucleotide misincorporations due to deamination of ancient DNA. *UDG-treated libraries* ***(right)*** *show low frequencies of C-to-T changes because UDG removes uracils.*



mtDNA contamination estimate for UDG-treated libraries

The mtDNA consensus from the targeted mtDNA enrichment for Loschbour, Motala 1, 2, 3, 4, 6, 9 and 12 was used to estimate mtDNA contamination levels from the deep sequencing of UDG-treated libraries. Reads that mapped to the human mtDNA genome with a mapping quality of at least 30 were extracted from the deep-sequencing data for all above mentioned samples. Based on the rate at which reads mismatched the consensus, we estimated contamination rates of 0.3% for Loschbour (0.24% - 0.39%, 95% HPD) and 0.02%-3.35% for the Motala individuals (Table S3.2). The contamination for Motala 3 and 9 could not be accurately estimated due to low mtDNA coverage. For Stuttgart, the mtDNA consensus sequence was directly built from high coverage shotgun data and used to estimate the number of reads that mismatch the consensus. The contamination estimate for Stuttgart for the deep-sequencing data was found to be 0.43% (0.29% - 0.62%, 95% HPD) showing that less than 1% of the human mitochondrial DNA sequences for Loschbour, Stuttgart and Motala 1, 2, 4, 6 and 12 are likely to come from a contaminating source with a different mitochondrial DNA.

Table S3.1. Summary of screening results from non-UDG-treated libraries*. Samples with high levels of authentic DNA that were chosen for deep sequencing are marked in grey. Samples with relatively high amounts of endogenous DNA are marked in bold*

		Capture for human mtDNA							Shotgun screening data	
library	**sample**	**Total reads**	**Unique mapping reads**	**Average coverage of mtDNA**	**nt covered at 5· fold coverage (% of mtDNA)**	**Damage at 5' (%)**	**Green et al. 2008 Contamination estimate (%)**	**Fu et al. 2013 Contamination estimate (%)**	**Total reads**	**endogenous DNA (%)**
ALB1	**Loschbour**	3916672	74435	320.8	16,569 (100%)	27.6	0 - 0.5	1.3 - 1.9	5901087	**22.5**
ALBK1	**Stuttgart**	n/a	n/a	n/a	n/a	n/a	n/a	n/a	824725	**66.4**
AMOT1	Motala 1	3557100	84947	321.1	16568 (100%)	31.5	0 - 0.6	0.7 - 2.2	531393	3.25
AMOT2	**Motala 2**	1590655	42962	172	16572 (100%)	27.5	0 - 1.4	1.7 - 4.0	1502464	**9.62**
AMOT3	Motala 2	3114034	774	2.98	3610 (21.8%)	18.1	30 - 90.3	5.5 - 64.5	-	-
AMOT4	Motala 3	5268145	107797	412.8	16567 (100%)	29.9	0 - 0.3	5.0 - 7.8	1576248	1.07
AMOT5	Motala 4	4692878	72187	248.5	16569 (100%)	34.7	0 - 0.6	0 - 1.5	860818	2.3
AMOT6	Motala 5	3628834	1592	6.1	10608 (64.3%)	17.3	0 - 27.8	1.8 - 19.1	-	-
AMOT7	Motala 6	2253825	57614	252.7	16570 (100%)	28.5	0 - 1.1	0 - 0.8	799225	0.82
AMOT8	Motala 6	1837405	23306	116.7	16570 (100%)	14.1	1 - 4.7	0.6 - 2.7	-	-
AMOT9	Motala 7	4265963	364	1.3	231 (1.4%)	12.9	4.5 - 32.1	0.8 - 70.9	-	-
AMOT10	Motala 8	948122	206	0.8	203 (1.23%)	7.5	4.4 - 16.1	0.4 - 41.4	-	-
AMOT11	Motala 8	1265744	1403	6.4	129 (0.78%)	n/a	n/a	0.4 - 45.8	-	-
AMOT12	Motala 9	1754892	30147	115.9	16569 (100%)	35	0.6 - 2.9	1.2 - 4.3	555139	1.92
AMOT13	**Motala 12**	1622517	99154	355.5	16570 (100%)	20.4	0.0 - 1.0	0.8 - 2.3	416757	**9.3**
AMOT14	Motala 12	1207779	2552	11.6	16088 (97.1%)	27.7	0.7 - 20.2	0.4 - 6.4	-	-
AMOT15	Motala 4170	2323981	142	0.5	117 (0.72%)	0	1.4 - 16.5	0.8 - 70.6	-	-
AMOT16	Motala MKA	3905092	717	2.3	1984 (12.1%)	8.8	2.4 - 7.2	0.3 - 30.9	-	-

Table S3.2. Summary of contamination estimates for UDG-treated libraries.

library	sample	mtDNA contamination estimate	average mtDNA coverage	ratio mtDNA/ nuclear DNA	autosomal estimates	X Chr estimates
ALB2-14	Loschbour	0.3% (0.24-0.39, 95% HPD)	1519.6	76.9	0.44% (CI: 0.35-0.53%)	0.45%
ALBK2	Stuttgart	0.43% (0.29-0.62, 95% HPD)	371.5	20.9	0.30% (CI: 0.22-0.39%)	-
AMOT17	Motala1	0.6% (0.16-1.65, 95% HPD)	32.8	241.2	-	-
AMOT18	Motala2	0.02% (0.00-0.29, 95% HPD)	85	525.2	-	-
AMOT19	Motala3	-	<1	78.9	-	-
AMOT20	Motala4	0.91% (0.22-4.03, 95% HPD)	9.12	111.9	-	-
AMOT21	Motala6	0.18% (0.03-3.81, 95% HPD)	4.79	171	-	-
AMOT22	Motala9	-	2.35	212.4	-	-
AMOT23	Motala12	0.34% (0.18-0.71 95% HPD)	144.7	64.4	-	-

mtDNA to nuclear DNA ratio

The ratio of mtDNA to nuclear DNA was calculated by dividing the average coverage of the mtDNA by the average coverage across all autosomes, effectively giving the number of copies of the mitochondrial genome per cell. The copy number ranges from 21 to 525 between samples, and is substantially lower than previous aDNA studies on bone (Green *et al.* 2008). This could be due to differential mitochondrial density in different tissues (Veltri *et al.* 1990). As all samples were taken from molars, this suggests that dental tissue may have a comparatively low mitochondrial copy number.

NuclearDNA contamination estimates

We used two approaches to estimate the proportion of nuclear contamination in Loschbour and Stuttgart

1. In the case of the male sample, Loschbour, the presence of polymorphic sites on the X chromosome was used to estimate contamination (similar to the approach taken in Rasmussen *et al.* 2011).
2. For Loschbour and Stuttgart, we used a maximum-likelihood-based estimate of autosomal contamination that uses variation at sites fixed in the 1000 Genomes project humans to co-estimate error, heterozygosity and contamination (Fu *et al.*, in preparation).

As Loschbour is very likely a male (SI5), heterozygous sites along the X chromosome are not expected. Sites where a second allele is observed are then due to:

1. Contamination
2. Sequencing errors
3. Mismapping

In an approach similar to that used for the Australian Aboriginal Genome[1], we computed the frequency of each base at positions that are polymorphic on chromosome X in the 1000 Genomes[2] dataset.

To reduce the effect of mismapping, only genomic regions with high mappability (SI2) were analyzed. Reads were required to have a mapping quality of at least 30, and only bases with a quality of at least 30 were considered for this analysis. Sites were required to fall within the 95th percentile of the coverage distribution for chromosome X, resulting in a minimum coverage of 4× and a maximum of 21×.

Assuming that contamination and error are both low, the true Loschbour allele will be the majority allele at each site. The observation of minor alleles on chromosome X may arise from either contamination or error. To determine the contamination we recorded the allele frequencies at each site for the Eurasian 1000 genomes populations: British (GBR), Tuscan (TSI), Chinese (CHB, CHS), Japanese (JPT), Iberian (IBS), Finnish (FIN), and Central European (CEU). To determine the sequencing error rate, the nucleotides adjacent to each tested site were considered likely to be monomorphic. The observation of multiple alleles at these sites was assumed to approximate the background sequencing error rate.

For each site that is polymorphic among the 1000 genomes individuals the numbers of major and minor alleles were computed. Triallelic or tetraallelic sites were discarded. The number of major and minor alleles was computed for adjacent sites. The tally of minor and major alleles is presented in Table S3.3. The background probability of error is determined by the base quality cutoff.

Table S3.3. Divergence at assumed polymorphic and monomorphic sites for Loschbour

Type	**Sample**	**Computed value**	**Observed probability of error**
Polymorphic	h_i	7,138,322	0.002891
	e_i	20,697	
Adjacent	h_i'	7,123,114	0.001393
	e_i'	9,934	

For any polymorphic sites we use the observed probability of error (ε) at adjacent sites (ε = 0.001393) as the background error rate. For a given contamination rate, the probability of an allele occurring at frequency f_i at position i is given by cf_i.Therefore, the probability of observing one minor allele is given by:

$$cf_i + (1 - c)\varepsilon$$

The total probability of seeing the aforementioned read distribution at position i where h_i is the major count and minor allele count is given by:

$$[cf_i + (1 - c)\varepsilon]^{e_i}[1 - (cf_i + (1 - c)\varepsilon)]^{h_i}$$

We compute the likelihood of the data given the parameters. The total likelihood for all sites is then given by:

$$p(data|c) = \prod_i [cf_i + (1 - c)\varepsilon]^{e_i}[1 - (cf_i + (1 - c)\varepsilon)]^{h_i}$$

By analyzing the logarithm of the likelihood surface, we infer a maximum of 0.45% contamination in Loschbour.

Autosomal contamination estimate

For all samples, contamination rates on the autosomes were estimating using a method based on that of Meyer *et al.* 2012 that is based on the observation that some sites are more susceptible to error than others. The method is a maximum likelihood-based co-estimation of sequence error, contamination and two population parameters, and assumes that present-day human contaminants will contribute derived alleles to the archaic human sequences. The analysis conditions on sites where the derived allele is fixed in the 1000 Genomes individuals as compared to great ape outgroups. Low frequency allele counts at these homozygous positions are used to infer contamination and sequence error.

Reads were required to have a minimal length of 35 and a mapping quality of at least 30. We condition on sites where the derived allele is fixed in the 1000 Genomes individuals as compared to great ape outgroups. Low frequency allele counts at these homozygous positions are used to infer contamination and sequence error.

Reads were required to have a minimal length of 35 and a mapping quality of at least 30. The method estimates low contamination in both samples; the estimated contamination for Loschbour and Stuttgart are 0.44% (CI: 0.35-0.53%), 0.30% (CI: 0.22-0.39%), respectively (Table S3.2).

References

Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., *et al.*(2012). An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56-65, doi:10.1038/nature11632.

Briggs, W. et al. (2010).Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. Nucleic Acids Res. 38, e87.

Fu, Q., Mittnik, A., Johnson, P. L., Bos, K., Lari, M., Bollongino, R., et al. (2013). A Revised Timescale for Human Evolution Based on Ancient Mitochondrial Genomes. Current Biology, 23(7), 553-559.

Green, R.E., Malaspinas, A.S., Krause, J., Briggs, A.W., Johnson, P.L., Uhler, C., Meyer, M.,Good, J.M., Maricic, T., Stenzel, U., et al. (2008). A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing.Cell 134, 416-426.

Krause, J., Fu, Q., J.M. Good, B. Viola, M.V. Shunkov, A.P. Derevianko, S. Pääbo (2010).The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. Nature, 464 , pp. 894–897

Jónsson H, Ginolhac A, Schubert M, Johnson P, Orlando L. (2013). mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* 2013.

Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., et al. (2012). A High-Coverage Genome Sequence from an Archaic Denisovan Individual.Science, 338(6104), 222-226.

Rasmussen, M., Guo, X., Wang, Y., Lohmueller, K.E., Rasmussen, S., Albrechtsen, A., *et al.* (2011) An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia. Science 334, 94-98, doi:DOI 10.1126/science.1211177.

Veltri, K.L., Espiritu, M., Singh, G. (1990). Distinct genomic copy number in mitochondria of different mammalian organs. J Cell Physiol 143: 160–164. doi: 10.1002/jcp.1041430122

Supplementary Information 4

Mitochondrial genome analysis

Alissa Mittnik* and Johannes Krause

* To whom correspondence should be addressed (amittnik@gmail.com)

This note describes the enrichment and phylogenetic analysis of mtDNA from 17 ancient human libraries derived from the Loschbour, Stuttgart and Motala individuals.

Enrichment of complete mtDNAs and high throughput sequencing

To test for DNA preservation and contemporary modern human contamination, mitochondrial DNA from 17 ancient human samples was analyzed using a long-range PCR-product based hybridization capture protocol (Maricic *et al.* 2010). Libraries (see also SI1, SI3.1) for targeted DNA capture were not treated using the UDG protocol in order to observe DNA damage patterns as additional indication for authenticity (Krause et al. 2010). The mtDNA capture was carried out as described previously in Fu *et al.* (2013a). The resulting captured mtDNA libraries were pooled and sequenced on the Illumina Genome Analyzer IIx platform with 2 × 76 + 7 cycles. Sequencing data was treated following Kircher 2012. In short; raw reads were filtered according to the individual indices, adapter and index sequences were removed, and paired-end reads overlapping by at least 11 nucleotides were collapsed to one fragment where the base with the higher quality score was called in the overlapping sequence. The sequences enriched for human mtDNA were mapped to the Reconstructed Sapiens Reference Sequence (RSRS, Behar *et al.* 2012) using a custom iterative mapping assembler (Green *et al.* 2008). Between 142 and 107,797 mtDNA fragments were found to map to the reference genome, resulting in average mtDNA coverage of 0.5 to 421 fold (Table S3.1).

Phylogenetic analysis of the mitochondrial genomes

The consensus sequences of all samples that fulfilled the authenticity criteria were assigned to haplogroups using HaploFind (Vianello *et al*. 2013, Table S4.1). All Mesolithic genomes belong to haplogroups U2 or U5, which are common among pre-Neolithic Europeans as has been shown earlier (Caramelli *et al.* 2003, Caramelli *et al.* 2008, Bramanti *et al.* 2009, Krause *et al.* 2010, Hervella *et al.* 2012, Sánchez-Quinto *et al.* 2012, Fu *et al.* 2013a, Der Sarkissian *et al.* 2013, Bollongino *et al.* 2013, Figure S4.1). Motala 2 and 12 share the same haplotype, suggesting a close relationship through the maternal lineage. The Neolithic sample Stuttgart belongs to haplogroup T2, which is common among early European farmers (Haak *et al.* 2005, Haak *et al.* 2008, Bramanti *et al.* 2009, Malmström *et al.* 2009, Haak *et al.* 2010, Lacan 2011, Hervella *et al.* 2012, Gamba *et al.* 2012, Brandt *et al.* 2013, Bollongino *et al.* 2013) as well as present-day Europeans (Fu *et al.* 2012).

Table S4.1. Haplogroup assignments.

Sample	haplogroup	Additional substitutions
Loschbour	U5b1a	T16189C!, A6701G
Motala 1	U5a1	G5460A
Motala 2	U2e1	C16527T
Motala 3	U5a1	G5460A, A9389G
Motala 4	U5a2d	A13158G
Motala 6	U5a2d	C152T!, G6480A
Motala 9	U5a2	G228A, G1888A, A2246G, C3756T, G6917A, A9531G
Motala 12	U2e1	C16527T
Stuttgart	T2c1d1	T152C!, C6340T, T16296C!

Figure S4.1: Haplogroup frequencies of ancient and modern Europeans. (Top left) *shows haplogroup frequencies in Europe before the onset of the Neolithic,* ***(top right)*** *during the Neolithic and* ***(bottom left)*** *today.*

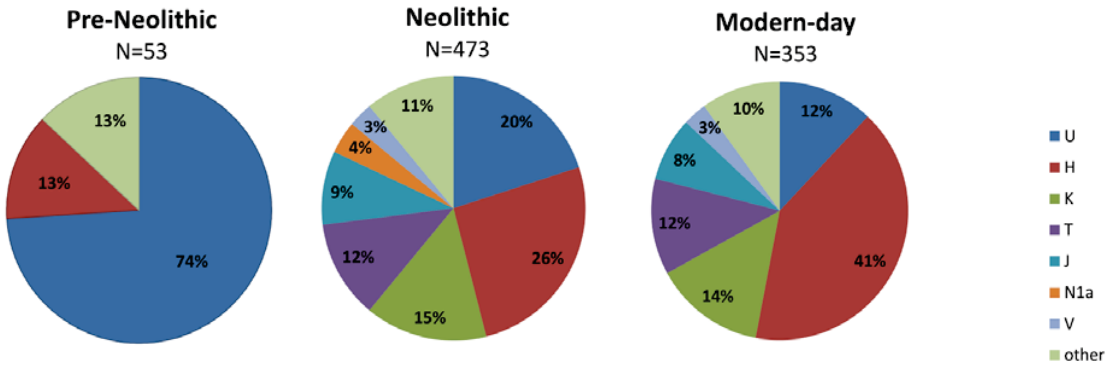
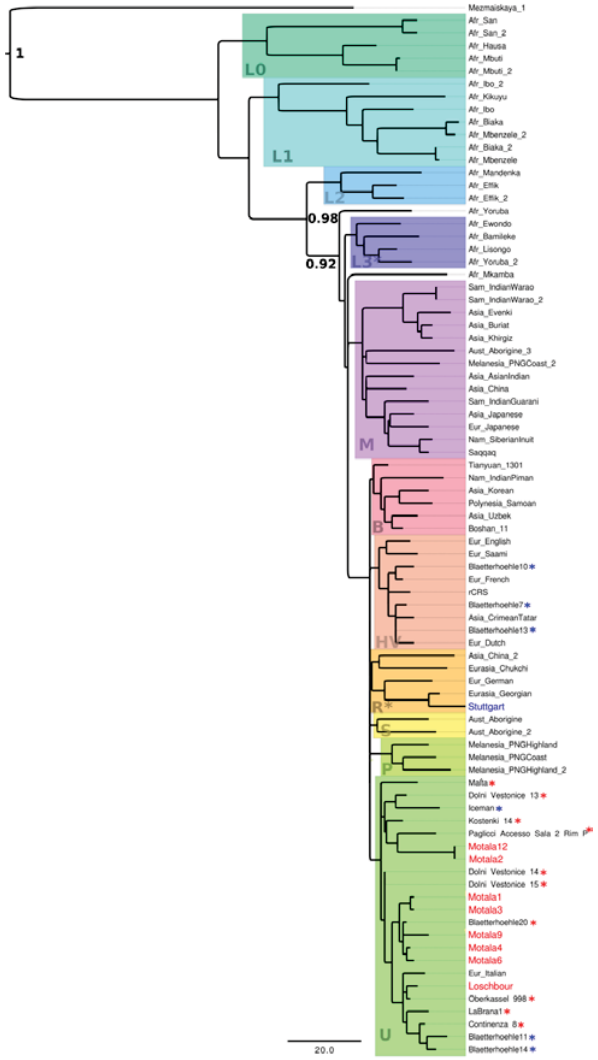


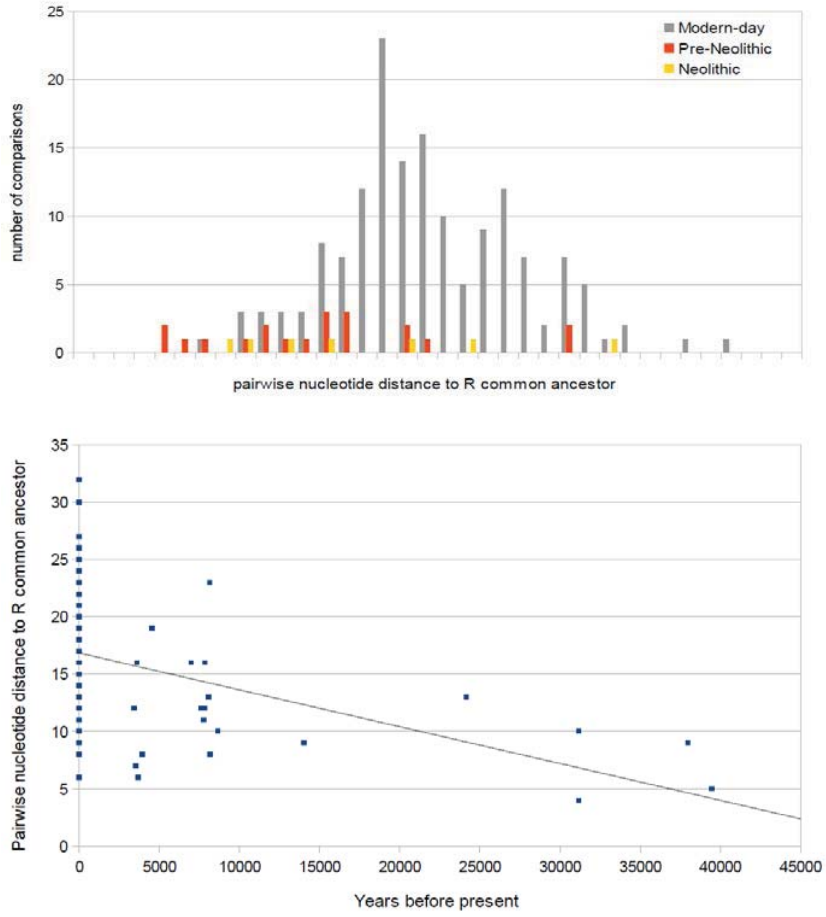
Figure S4.2: Maximum Parsimony tree of 54 modern and 27 ancient mtDNA genomes. *The Mesolithic genomes studied here in red, the Stuttgart sample is in blue, and previously published European pre-Neolithic and Neolithic genomes are marked with red and blue asterisks, respectively. Bootstrap values above 0.9 are given at major nodes.*



The mtDNA consensus sequences were aligned using the software MUSCLE (Edgar 2004). MEGA 5.2 (Tamura et al. 2011) was used to generate a Maximum Parsimony tree, which included the mtDNA sequences obtained here along with previously published complete early modern human mtDNAs (Ermini *et al.* 2008, Gilbert *et al.* 2008, Krause *et al.* 2010, Sánchez-Quinto *et al.* 2012, Fu *et al.* 2013a, Fu *et al.* 2013b, Bollongino *et al.* 2013, Raghavan *et al.* 2013) and 54 present-day human mtDNAs from a worldwide dataset (Ingman *et al.* 2000). Figure S4.2 shows that the Mesolithic genomes studied here cluster together with previously published pre-Neolithic genomes.

MEGA 5.2 was also used to calculate nucleotide distances to the root of haplogroup R for the ancient sequences belonging to this clade as well as to 154 modern-day sequences falling into haplogroup R (Green *et al.* 2008). The mean nucleotide distance of the prehistoric samples to the most recent common ancestor of haplogroup R is significantly shorter than that of all modern mtDNAs (Student's t-test, $p = 0.02$, Figure S4.3), demonstrating the effect of branch shortening (ancient mtDNA has accumulated fewer substitutions over time; Fu *et al.* 2013a). The early Neolithic individual Stuttgart falls at the upper end of the prehistoric distribution. Plotting the age of the samples against the pairwise nucleotide distance and calculating the slope of the regression (Figure S4.3) gives an estimate of the mitochondrial substitution rate of $1.94 \pm 0.36 \times 10^{-8}$ substitutions per bp per year for the mtDNA genome, comparable to previous estimates (Fu *et al.* 2013a, Brotherton *et al*. 2013).

Figure S4.3: Pairwise distance comparisons to the root of haplogroup R. (Top) *Pairwise nucleotide distance to the root of hg R for the complete mtDNA of 154 present-day and 20 prehistoric humans that fall inside the R clade.* ***(Bottom)*** *Plot of nucleotide distance against age of the sequence, slope of the linear regression gives substitution rate of the whole mitochondrial genome ($1.94 \pm 0.36 \times 10^{-8}$ substitutions per bp per year).*



References

Behar D.M., van Oven M., Rosset S., Metspalu M., Loogvali E.L. et al. (2012) A "Copernican" reassessment of the human mitochondrial DNA tree from its root. Am J Hum Genet 90: 675-684. doi:10.1016/j.ajhg.2012.03.002.

Bollongino R, Nehlich O, Richards P, Orschiedt J, Thomas MG, Sell C, Fajkošová Z, Powell A, Burger J (2013) 2000 Years of Parallel Societies in Stone Age Central Europe. Science 342: 479-481.

Bramanti, B., Thomas, M. G., Haak, W., Unterlaender, M., Jores, P., Tambets, K., Antanaitis-Jacobs, I., Haidle, M. N., Jankauskas, R., Kind, C.-J., Lueth, F., Terberger, T., Hiller, J., Matsumura, S., Forster, P. and Burgerg, J. (2009): Genetic discontinuity between local hunter-gatherers and central Europe's first farmers. Science 326(5949):137–140.

Brandt, G., Haak, W., Adler, C., Roth, C., Szécsényi-Nagy, A., Karimnia, S., Meller, H., Ganslmeier, R., Friederich, S., Dresely, V., *et al* (2013). Ancient DNA reveals key stages in the formation of central European mitochondrial genetic diversity. *Science*, 342(6155), 257-261.

Briggs, W. et al. (2010).Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA.Nucleic Acids Res. 38, e87.

Brotherton, P., Haak, W., Templeton, J., Brandt, G., Soubrier, J., Jane Adler, C. et al. (2013), Neolithic mitochondrial haplogroup H genomes and the genetic origins of Europeans. *Nature Communications.* 4, 1764.

Der Sarkissian C, Balanovsky O, Brandt G, Khartanovich V, Buzhilova A, Koshel S, V Zaporozhchenko, Gronenborn D, Moiseyev V, Kolpakov E, Shumkin V, Alt KW, E Balanovska E, Cooper A, Haak W, The Genographic Consortium (2013). Ancient DNA Reveals Prehistoric Gene-Flow From Siberia in the Complex Human Population History of North East Europe. PLoS Genet. 2013; 9(2):e1003296s.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res., 32 , pp. 1792–1797

Ermini, L. C. Olivieri, E. Rizzi, G. Corti, R. Bonnal, P. Soares, S. Luciani, I. Marota, G. De Bellis, M.B. Richards, F. Rollo (2008). Complete mitochondrial genome sequence of the Tyrolean Iceman Curr. Biol., 18, pp. 1687–1693

Fu, Q., Mittnik, A., Johnson, P. L., Bos, K., Lari, M., Bollongino, R., et al. (2013). A Revised Timescale for Human Evolution Based on Ancient Mitochondrial Genomes. Current Biology, 23(7), 553-559.

Fu, Q., M. Meyer, X. Gao, U. Stenzel, H.A. Burbano, J. Kelso, S. Pääbo (2013). DNA analysis of an early modern human from Tianyuan Cave, China ProcNatlAcadSci U S A., 110, pp. 2223–2337

Gamba C, Fernández E, Tirado M, Deguilloux MF, Pemonge MH, Utrilla P, Edo M, Molist M, Rasteiro R, Chikhi L, Arroyo-Pardo E. (2012). Ancient DNA from an Early Neolithic Iberian population supports a pioneer colonization by first farmers. Mol Ecol. 2012;21(1):45-56.

Gilbert, M.T., T. Kivisild, B. Grønnow, P.K. Andersen, E. Metspalu, M. Reidla, E. Tamm, E. Axelsson, A. Götherström, P.F. Campos *et al. (2008)*. Paleo-Eskimo mtDNAgenome reveals matrilineal discontinuity in Greenland Science, 320, pp. 1787–1789

Green, R.E., Malaspinas, A.S., Krause, J., Briggs, A.W., Johnson, P.L., Uhler, C., Meyer, M.,Good, J.M., Maricic, T., Stenzel, U., et al. (2008).A complete Neandertalmitochondrialgenome sequence determined by high-throughput sequencing.Cell 134, 416-426.

Haak, W., Balanovsky, O., Sanchez, J. J., Koshel, S., Zaporozhchenko, V., Adler, C. J., Sarkissian, C. S. I. D., Brandt, G., Schwarz, C., Nicklisch, N., Dresely, V., Fritsch, B., Balanovska, E., Villems, R., Meller, H., Alt, K. W., Cooper, A. and of the Genographic Consortium, M. (2010): Ancient DNA from European early neolithic farmers reveals their near eastern affinities. PLoSBiol 8(11):e1000536.

Haak, W., Forster, P., Bramanti, B., Matsumura, S., Brandt, G., Tänzer, M., Villems, R., Renfrew, C., Gronenborn, D., Alt, K. W. and Burger, J. (2005): Ancient DNA from the first European farmers in 7500-year-old Neolithic sites. Science 310(5750):1016–1018.

Haak W, Brandt G, de Jong HN, Meyer C, Ganslmeier R, Heyd V, Hawkesworth C, Pike AW, Meller H, Alt KW. (2008). Ancient DNA, Strontium isotopes, and osteological analyses shed light on social and kinship organization of the Later Stone Age. ProcNatlAcadSci U S A. 2008; 105(47):18226-18231.

Hervella M, Izagirre N, Alonso S, Fregel R, Alonso A, Cabrera VM, de la Rúa C. (2012). Ancient DNA from Hunter-Gatherer and Farmer Groups from Northern Spain Supports a Random Dispersion Model for the Neolithic Expansion into Europe. PLoS one. 2012; 7(4):e34417.

Kircher, M. (2012) Analysis of high-throughput ancient DNA sequencing data. Methods Mol Biol. 840: 197-228.doi: 10.1007/978-1-61779-516-9_23.

Krause, J., Briggs, A. W., Kircher, M., Maricic, T., Zwyns, N., Derevianko, A. and Pääbo, S. (2010a): A complete mtDNA genome of an early modern human from Kostenki, Russia. CurrBiol 20(3):231–236.

Lacan M. (2011). La Néolithisation du bassinméditerranéen: Apports de l'ADNancien (University of Toulouse, 2011)

Malmström H, Gilbert MT, Thomas MG, Brandström M, Storå J, Molnar P, Andersen PK, Bendixen C, Holmlund G, Götherström A, Willerslev E. (2009). Ancient DNA reveals lack of continuity between neolithic hunter-gatherers and contemporary Scandinavians. Curr Biol. 2009; 19(20):1758-1762.

Maricic, T., Whitten, M. and Pääbo, S. (2010): Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. PLoS One 5(11):e14004.

Raghavan, M., Skoglund, P., Graf, K.E., Metspalu, M., Albrechtsen, A., Moltke, I., Rasmussen, S., Stafford, Jr. T.W., Orlando, L., Metspalu, E., Karmin, M., Tambets, K., Rootsi, S., Magi, R., Campos, P.F., Balanovska, E., Balanovsky, O., Khusnutdinova, E., Litvinov, S., Osipova, L.P., Fedorova, S.A., Voevoda, M.I.,

Sánchez-Quinto, F., Schroeder, H., Ramirez, O., Avila Arcos, M.D.C., Pybus, M., Olalde, I., Velazquez, A.M.V., Marcos, M.E.P., Encinas, J.M.V., Bertranpetit, J. , Orlando, L.A.A. , Gilbert, M.T.P., Lalueza-Fox, C.(2012). Genomic Affinities of Two 7,000-Year-Old *Iberian* Hunter-Gatherers.Current Biology, vol 22, no. 16, pp. 1494-1499.

Tamura, K., Peterson D., Peterson, N., Stecher, G., Nei, M., Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol., 28 , pp. 2731–2739

Vianello D., Sevini F., Castellani G., Lomartire L., Capri M. and Franceschi C. (2013) HAPLOFIND: ANew Method for High-Throughput mtDNAHaplogroup Assignment. Hum. Mutat. doi:10.1002/humu.22356

Supplementary Information 5

Sex determination and Y chromosome analysis

Iosif Lazaridis† and Gabriel Renaud†*

* To whom correspondence should be addressed (gabriel_renaud@eva.mpg.de)
† Contributed equally to this section

We infer which of the archaic human individuals in this study are likely to be male, and determine their Y chromosome haplogroups using publicly available Y chromosome SNPs. For Loschbour, which has the highest coverage of all samples, we also determine its phylogenetic placement in a larger Y-SNP dataset of modern humans.

Sex determination

Based on morphological elements of the skeleton such as the pelvis and the skull, the sex of an individual can be inferred with high accuracy; however ancient skeletons are often fragmentary and morphologically altered. It is therefore of interest to be able to use genetic information to determine the sex of an individual. As males have a single X and Y chromosome, the coverage of X chromosome nucleotides is expected to be about half of the autosomal coverage, and a significant number of reads are expected to map to the Y chromosome (Skoglund et al. 2013). Conversely, females will have X chromosome coverage comparable to the autosomal coverage, and few reads mapping to the Y chromosome (largely due to regions of similarity between the Y and the X chromosomes). The ratio of reads mapping to the X and Y can thus be used to infer sex.

We extracted reads of high map quality (MAPQ ≥ 30) using *samtools* 0.1.18, and identified Loschbour and five of the Motala individuals (#2, 3, 6, 9, 12) as males using the ratio of chrY to (chrX+chrY) reads, using a tool recently developed for this purpose (Table S5.1).[1]

Table S5.1: Sex determination using the number of reads (N) aligning to the X- and Y-chromosomes after MAPQ filtering. *The ratio (R_y) of NchrY/(NchrY+NchrX) with its standard error (SE) and 95% CI is presented.*

Sample	NchrY+NchrX	NchrY	R_y	SE	95% CI	Assignment
Loschbour	22,068,747	1,873,062	0.0849	0.0001	0.0848−0.085	XY
Stuttgart	34,997,784	87,882	0.0025	0.00001	0.0025−0.0025	XX
Motala1	349,043	1,095	0.0031	0.0001	0.003-0.0033	XX
Motala2	162,747	13,560	0.0833	0.0007	0.082−0.0847	XY
Motala3	588,788	49,687	0.0844	0.0004	0.0837−0.0851	XY
Motala4	143,005	483	0.0034	0.0002	0.0031-0.0037	XX
Motala6	25,549	2,176	0.0852	0.0017	0.0817−0.0886	XY
Motala9	11,571	932	0.0805	0.0025	0.0756−0.0855	XY
Motala12	2,384,534	200,346	0.084	0.0002	0.0837−0.0844	XY

Y chromosome haplogroup determination

We used Y-chromosome SNPs included in the Y chromosome phylogeny of the International Society of Genetic Genealogy (ISOGG) to determine the haplogroups of the ancient samples. We removed SNPs with incomplete information (e.g., lacking physical position) and SNPs marked by ISOGG as "Investigation". For each SNP we examined the reference allele in the same physical position (*hg19/GRCh37*) to correct strand assignment errors. Since C/G and A/T SNPs cannot be fixed in this

manner we did not use these for further analysis. We also excluded apparently heterozygous sites since these are not expected on chromosome Y and might reflect contamination, mapping error, or deamination. We intersected the set of called sites for each individual with the physical positions of ISOGG SNPs, and used this to determine the Y-chromosome haplogroup for each individual.

Loschbour belonged to Y chromosome haplogroup I2a1b, defined by two mutations M423 and L178. Here we list a number of upstream mutations that securely place this individual within the I branch of the phylogeny, as well as a number of downstream sites for which it is ancestral (Table S5.2).

Table S5.2: Diagnostic Loschbour alleles place it in haplogroup I2a1b*(xI2a1b1, I2a1b2, I2a1b3).

Haplogroup	SNP	Ancestral	Derived	GRCh37	Read Depth	State
I2a1b	L178	G	A	15574052	12	+
I2a1b	M423	G	A	19096091	13	+
I2a1	P37.2	T	C	14491684	7	+
I2a	L460	A	C	7879415	7	+
I2	M438	A	G	16638804	14	+
I2	L68	C	T	18700150	12	+
I	P38	A	C	14484379	2	+
I	M170	A	C	14847792	14	+
I	M258	T	C	15023364	5	+
I	U179	G	A	16354708	9	+
I	L41	G	A	19048602	3	+
I2a1b1	M359.2	T	C	14491671	9	−
I2a1b2	L161.1	C	T	22513718	7	−
I2a1b3	L621	G	A	18760081	15	−
I2a1b3a	L147.2	T	C	6753258	5	−

Note: State (+) here and in following tables indicates presence of the derived allele; state (−)the ancestral allele.

Table S5.3: Diagnostic Motala2 alleles place it in haplogroup I*(xI1, I2a2,I2a1b3)

Haplogroup	SNP	Ancestral	Derived	GRCh37	Read Depth	State
I	P38	A	C	14484379	1	+
I	U179	G	A	16354708	1	+
I	L41	G	A	19048602	1	+
I1	M253	C	T	15022707	1	−
I1a2a1a	Z140	G	A	17863355	1	−
I2a1b3	L621	G	A	18760081	1	−
I2a2	L37	T	C	17516123	1	−
I2a2a1b2	L703	G	A	14288983	1	−
I2a2a1c1b1a1a	S434	G	A	17147721	1	−

Motala2 (Table S5.3) belongs to Y-haplogroup I on the basis of three mutations.

Motala3 (Table S5.4) belongs to Y-haplogroup I2 on the basis of L68+, with three additional mutations placing it in Y-haplogroup I.

Motala6 was L55+ (19413335 G>A), placing it in Y-haplogroup Q1a2a, but L232− (17516095 G>A), which contradicts the hypothesis that it belongs to haplogroup Q1. These two observations are phylogenetically inconsistent, and we are unable to assign a haplogroup to this individual.

Motala9 (Table S5.5) belongs to Y-haplogroup I on the basis of P38+ but not on the I1 branch on the basis of P40−. P40 is a C-to-T mutation and might reflect ancient DNA damage. I1 occurs at high frequencies in present-day Swedes[2], but has not been detected in prehistoric Europe including here.

Table S5.4: Diagnostic Motala3 alleles place it in haplogroup I2*(xI2a1a, I2a2, I2b).

Haplogroup	SNP	Ancestral	Derived	GRCh37	Read Depth	State
I2	L68	C	T	18700150	1	+
I	M258	T	C	15023364	2	+
I	U179	G	A	16354708	1	+
I2a1a	M26	G	A	21865821	1	−
I2a1b1	M359.2	T	C	14491671	1	−
I2a1b3	L621	G	A	18760081	1	−
I2a2	L181	G	T	19077754	1	−
I2a2a	L59	C	T	7113556	1	−
I2a2a	L622	C	A	13718315	1	−
I2a2a1a1a	L137	G	A	9791250	1	−
I2a2a1a1a	L369	T	C	14850314	1	−
I2a2a1a1a	L126	C	T	14901633	1	−
I2a2a1b1	P78	G	A	6740387	1	−
I2a2a1b2	L699	A	G	2663920	1	−
I2a2a1c1a1	P95	G	T	14869706	1	−
I2a2a1c1b1a1	Z190	G	T	17473966	1	−
I2a2a1c1b1a1a	S434	G	A	17147721	1	−
I2a2a1d1a	L812	G	A	14850035	1	−
I2a2b	L38	A	G	15668070	1	−
I2a2b	L40	T	C	16202267	1	−
I2a2b	L65.1	A	G	16626617	2	−
I2b	L417	T	C	8426321	1	−

Table S5.5: Diagnostic Motala9 alleles place it in haplogroup I*(xI1).

Haplogroup	SNP	Ancestral	Derived	GRCh37	Read Depth	State
I	P38	A	C	14484379	1	+
I1	P40	C	T	14484394	1	−

Motala12 (Table S5.6) belonged to Y-haplogroup I2a1b on the basis of L178+ (15574052 G>A) and was L621− and M359.2− and thus assigned to I2a1b*(xI2a1b1, I2a1b3). A number of upstream mutations securely place it in haplogroup I. It appears that the L178 clade was present in at least two locations of pre-Neolithic Europe, as both Motala12 and Loschbour belonged to it.

Table S5.6: Diagnostic Motala12 alleles place it in haplogroup I2a1b*(xI2a1b1, I2a1b3).

Haplogroup	SNP	Ancestral	Derived	GRCh37	Read Depth	State
I2a1b	L178	G	A	15574052	2	+
I2a1	P37.2	T	C	14491684	1	+
I2a	L460	A	C	7879415	2	+
I2	L68	C	T	18700150	1	+
I	M170	A	C	14847792	1	+
I	M258	T	C	15023364	2	+
I	U179	G	A	16354708	1	+
I2a1b1	M359.2	T	C	14491671	1	−
I2a1b3	L621	G	A	18760081	2	−

Phylogenetic analysis of the Loschbour Y chromosome

Given that Loschbour carried a Y chromosome belonging to haplogroup I, we sought to investigate how this individual's Y-chromosome compares to the diversity of present-day humans. We used a dataset from Lippold *et al.* (submitted) which contains the genotype at 2,799 positions for a worldwide panel of 623 Y chromosomes. Using BEAST v1.7.51[3] with a coalescence prior of 60,000 years for all non-Africans and a tip age for Loschbour of 8,000 years, we reconstructed a Bayesian inference tree. Compared to the haplotype assignment described above, the correctness of a published phylogeny is not assumed; the phylogeny is instead reconstructed based on Y chromosome polymorphism data.

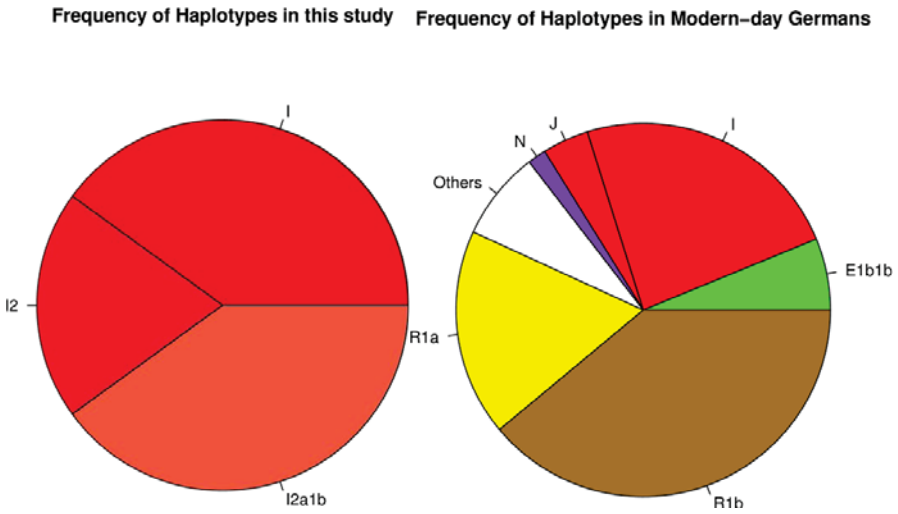
The Y chromosome of Loschbour clusters with present-day haplogroup I individuals (Figure S5.1), confirming the placement based on diagnostic alleles for this haplogroup (Table S5.2). The coalescence of the I and J2 haplogroups is infered to have occurred 31 kya whereas the coalescence of this group to the R haplogroup is estimated at 40 kya. Such numbers are somewhat consistent with the current models of population expansion in Europe[4]. On a finer scale, a modern Russian individual (HGDP00887) was found to share high similarity with the Loschbour individual. Out of 2,790 informative positions in both individuals (9 were not covered by reads in Loschbour), only 5 sites were different, including 3 transversions and 2 transitions. We used another Russian individual (HGDP00894) to show that all five of these five mutations fell on the HGDP00887 branch rather than on the Loschbour branch. These derived mutations may either have occurred on the HGDP00887 branch after divergence from Loschbour, or they might represent errors in HGDP00887.

***Figure S5.1**: **Phylogenetic position of Loschbour Y chromosome within present-day haplogroup I.** The highlighted branch (yellow) displays the Loschbour individual and its closest relative for the Y chromosome in the dataset, a present-day Russian. The inferred coalescence of the sub-tree here with with the R haplogroup (not shown) is 40,661 years, consistent with some previous estimates[5].*

Frequency of Y chromosome haplotypes

All 5 of the male individuals in this study belonged to the I haplogroup. Among present-day Germans, this is found at a much lower frequency of ~24% (Figure S**5**.2). At present, the limited number of ancient samples for which Y chromosome data is currently available makes it difficult to assess how statistically surprising it is that the Y haplogroup group occurs in all five of the ancient Mesolithic males but in only a quarter of present-day German males.

Figure S5.2: Pie chart of Y chromosome haplogroups of the individuals in this study and present-day Germans. *For the ancient individuals, only haplogroup I was found. However, in present-day Europeans from Germany, I is now a minority haplogroup.*



Discussion

Our finding that Loschbour and all four Motala males whose haplogroups we could determine belong to Y-haplogroup I is not entirely unexpected, as this clade of the human Y-chromosome phylogeny is found almost exclusively in Europe[6], with much rarer occurrences elsewhere. Its sister clade (haplogroup J) is thought to have a Near Eastern origin[7]. It has been hypothesized that I was common in pre-agricultural Europeans[8], and our study confirms this directly as it documents its presence in two European hunter-gatherer groups from the period immediately antedating the Neolithic transition.

We cannot, at present, determine when Y chromosome haplogroup I entered Europe, although its occurrence in two Mesolithic European hunter-gatherer populations (Loschbour and Motala) suggest an old origin, potentially entering Europe during the Upper Paleolithic around 40,000 years ago.

It is tempting to speculate that haplogroup I might be the dominant European Y chromosome haplogroup in Palaeolithic Europe, as the male counterpart of maternally inherited mitochondrial haplogroup U (SI4). Y chromsome haplogroup I[9] as well as mitochondrial haplogroup U were also identified in Neolithic Europeans, and are found throughout Europe in present-day populations. Thus, both maternally- and paternally-inherited genetic components of present-day Europeans may reflect a history of major admixture: genetic contribution from both the hunter-gatherers and early farmers of Europe. We further note that Y chromosome haplogroup I is scarce in the Near East today, with only sporadic occurrences of this haplogroup in the North Caucasus (~3% in frequency)[10], consistent with very limited gene flow from Europe into this area.

The present-day frequency of haplogroup I in Europe is variable, with local maxima in Scandinavia[2] and the western Balkans which might reflect more recent expansions. Our finding that Loschbour, a

Mesolithic west European, was M423+ contrasts with a previous suggestion[11] that this lineage diffused during the Neolithic from south-eastern Europe.

The absence of Y-haplogroup R1b in our two sample locations is striking given that it is, at present, the major west European lineage. Importantly, however, it has not yet been found in ancient European contexts prior to a Bell Beaker burial from Germany (2,800-2,000BC)[12], while the related R1a lineage has a first known occurrence in a Corded Ware burial also from Germany (2,600BC)[13]. This casts doubt on early suggestions associating these haplogroups with Paleolithic Europeans[14], and is more consistent with their Neolithic entry into Europe at least in the case of R1b[15, 16]. More research is needed to document the time and place of their earliest occurrence in Europe. Interestingly, the Mal'ta boy belonged[17] to haplogroup R* and we tentatively suggest that some haplogroup R bearers may be responsible for the wider dissemination of Ancient North Eurasian ancestry into Europe, as their haplogroup Q relatives may have plausibly done into the Americas[17].

This work provides a first glimpse into the the pre-Neolithic Y chromosomes of Europe. Despite the fact that our sample is limited to two locations and five male individuals, the results in this section are consistent with haplogroup I representing a major pre-Neolithic European clade, and hint at subsequent events during and after the Neolithic transition as important contributors to the Y chromosomal variation of living Europeans.

References

1 Pontus Skoglund, Jan Storå, Anders Götherström, and Mattias Jakobsson, 'Accurate Sex Identification of Ancient Human Remains Using DNA Shotgun Sequencing', *Journal of Archaeological Science,* 40 (2013), 4477-82.

2 Andreas O. Karlsson, Thomas Wallerstrom, Anders Gotherstrom, and Gunilla Holmlund, 'Y-Chromosome Diversity in Sweden - a Long-Time Perspective', *Eur J Hum Genet,* 14 (2006), 963-70.

3 Alexei Drummond, and Andrew Rambaut, 'Beast: Bayesian Evolutionary Analysis by Sampling Trees', *BMC Evolutionary Biology,* 7 (2007), 214.

4 S. Benazzi, K. Douka, C. Fornai, C. C. Bauer, O. Kullmer, J. Svoboda, I. Pap, F. Mallegni, P. Bayle, M. Coquerelle, S. Condemi, A. Ronchitelli, K. Harvati, and G. W. Weber, 'Early Dispersal of Modern Humans in Europe and Implications for Neanderthal Behaviour', *Nature,* 479 (2011), 525-U249.

5 T. M. Karafet, F. L. Mendez, M. B. Meilerman, P. A. Underhill, S. L. Zegura, and M. F. Hammer, 'New Binary Polymorphisms Reshape and Increase Resolution of the Human Y Chromosomal Haplogroup Tree', *Genome Res,* 18 (2008), 830-8.

6 Siiri Rootsi, Toomas Kivisild, Giorgia Benuzzi, Hela Help, Marina Bermisheva, Ildus Kutuev, Lovorka Barać, Marijana Peričić, Oleg Balanovsky, Andrey Pshenichnov, Daniel Dion, Monica Grobei, Lev A. Zhivotovsky, Vincenza Battaglia, Alessandro Achilli, Nadia Al-Zahery, Jüri Parik, Roy King, Cengiz Cinnioğlu, Elsa Khusnutdinova, Pavao Rudan, Elena Balanovska, Wolfgang Scheffrahn, Maya Simonescu, Antonio Brehm, Rita Goncalves, Alexandra Rosa, Jean-Paul Moisan, Andre Chaventre, Vladimir Ferak, Sandor Füredi, Peter J. Oefner, Peidong Shen, Lars Beckman, Ilia Mikerezi, Rifet Terzić, Dragan Primorac, Anne Cambon-Thomsen, Astrida Krumina, Antonio Torroni, Peter A. Underhill, A. Silvana Santachiara-Benerecetti, Richard Villems, Chiara Magri, and Ornella Semino, 'Phylogeography of Y-Chromosome Haplogroup I Reveals Distinct Domains of Prehistoric Gene Flow in Europe', *The American Journal of Human Genetics,* 75 (2004), 128-37.

7 Ornella Semino, Chiara Magri, Giorgia Benuzzi, Alice A. Lin, Nadia Al-Zahery, Vincenza Battaglia, Liliana Maccioni, Costas Triantaphyllidis, Peidong Shen, Peter J. Oefner, Lev A. Zhivotovsky, Roy King, Antonio Torroni, L. Luca Cavalli-Sforza, Peter A. Underhill, and A. Silvana Santachiara-Benerecetti, 'Origin, Diffusion, and Differentiation of Y-Chromosome Haplogroups E and J: Inferences on the Neolithization of Europe and Later Migratory Events in the Mediterranean Area', *American journal of human genetics,* 74 (2004), 1023-34.

8 Pedro Soares, Alessandro Achilli, Ornella Semino, William Davies, Vincent Macaulay, Hans-Jürgen Bandelt, Antonio Torroni, and Martin B. Richards, 'The Archaeogenetics of Europe', *Current Biology,* 20 (2010), R174-R83.

9 Marie Lacan, Christine Keyser, François-Xavier Ricaut, Nicolas Brucato, Francis Duranthon, Jean Guilaine, Eric Crubézy, and Bertrand Ludes, 'Ancient DNA Reveals Male Diffusion through the Neolithic Mediterranean Route', *Proceedings of the National Academy of Sciences* (2011).

10 Bayazit Yunusbayev, Mait Metspalu, Mari Järve, Ildus Kutuev, Siiri Rootsi, Ene Metspalu, Doron M. Behar, Kärt Varendi, Hovhannes Sahakyan, Rita Khusainova, Levon Yepiskoposyan, Elza K. Khusnutdinova, Peter A. Underhill, Toomas Kivisild, and Richard Villems, 'The Caucasus as an Asymmetric Semipermeable Barrier to Ancient Human Migrations', *Molecular Biology and Evolution* (2011).

11 Vincenza Battaglia, Simona Fornarino, Nadia Al-Zahery, Anna Olivieri, Maria Pala, Natalie M. Myres, Roy J. King, Siiri Rootsi, Damir Marjanovic, Dragan Primorac, Rifat Hadziselimovic, Stojko Vidovic, Katia Drobnic, Naser Durmishi, Antonio Torroni, A. Silvana Santachiara-Benerecetti, Peter A. Underhill, and Ornella Semino, 'Y-Chromosomal Evidence of the Cultural Diffusion of Agriculture in Southeast Europe', *Eur J Hum Genet,* 17 (2008), 820-30.

12 Esther J. Lee, Cheryl Makarewicz, Rebecca Renneberg, Melanie Harder, Ben Krause-Kyora, Stephanie Müller, Sven Ostritz, Lars Fehren-Schmitz, Stefan Schreiber, Johannes Müller, Nicole von Wurmb-Schwark, and Almut Nebel, 'Emerging Genetic Patterns of the European Neolithic: Perspectives from a Late Neolithic Bell Beaker Burial Site in Germany', *American Journal of Physical Anthropology,* 148 (2012), 571-79.

13 Wolfgang Haak, Guido Brandt, Hylke N. de Jong, Christian Meyer, Robert Ganslmeier, Volker Heyd, Chris Hawkesworth, Alistair W. G. Pike, Harald Meller, and Kurt W. Alt, 'Ancient DNA, Strontium Isotopes, and Osteological Analyses Shed Light on Social and Kinship Organization of the Later Stone Age', *Proceedings of the National Academy of Sciences* (2008).

14 Ornella Semino, Giuseppe Passarino, † Peter J. Oefner, Alice A. Lin, Svetlana Arbuzova, Lars E. Beckman, Giovanna De Benedictis, Paolo Francalacci, Anastasia Kouvatsi, Svetlana Limborska, Mladen Marcikiæ, Anna Mika, Barbara Mika, Dragan Primorac, A. Silvana Santachiara-Benerecetti, L. Luca Cavalli-Sforza, and Peter A. Underhill, 'The Genetic Legacy of Paleolithic Homo Sapiens Sapiens in Extant Europeans: A Y Chromosome Perspective', *Science,* 290 (2000), 1155-59.

15 Patricia Balaresque, Georgina R. Bowden, Susan M. Adams, Ho-Yee Leung, Turi E. King, Zoë H. Rosser, Jane Goodwin, Jean-Paul Moisan, Christelle Richard, Ann Millward, Andrew G. Demaine, Guido Barbujani, Carlo Previderè, Ian J. Wilson, Chris Tyler-Smith, and Mark A. Jobling, 'A Predominantly Neolithic Origin for European Paternal Lineages', *PLoS Biol,* 8 (2010), e1000285.

16 Per Sjödin, and Olivier François, 'Wave-of-Advance Models of the Diffusion of the Y Chromosome Haplogroup R1b1b2 in Europe', *PLoS ONE,* 6 (2011), e21592.

17 Maanasa Raghavan, Pontus Skoglund, Kelly E. Graf, Mait Metspalu, Anders Albrechtsen, Ida Moltke, Simon Rasmussen, Thomas W. Stafford Jr, Ludovic Orlando, Ene Metspalu, Monika Karmin, Kristiina Tambets, Siiri Rootsi, Reedik Magi, Paula F. Campos, Elena Balanovska, Oleg Balanovsky, Elza Khusnutdinova, Sergey Litvinov, Ludmila P. Osipova, Sardana A. Fedorova, Mikhail I. Voevoda, Michael DeGiorgio, Thomas Sicheritz-Ponten, Soren Brunak, Svetlana Demeshchenko, Toomas Kivisild, Richard Villems, Rasmus Nielsen, Mattias Jakobsson, and Eske Willerslev, 'Upper Palaeolithic Siberian Genome Reveals Dual Ancestry of Native Americans', *Nature,* advance online publication (2013).

Supplementary Information 6

Exome deleterious mutation loads

Sergi Castellano*

* To whom correspondence should be addressed (sergi.castellano@eva.mpg.de)

To quantify departures from neutrality in the coding regions of the genomes of Loschbour, Stuttgart and three present-day humans (Yoruba, HGDP00927; French, HGDP00521; and Han, HGDP00778), we classified coding derived alleles by functional class in heterozygous (Table S6.1) and homozygous derived (Table S6.2) positions in each individual.

Table S6.1. *Distribution of derived alleles by functional class in heterozygous positions*

		Ancient modern humans		Present-day humans		
Category		**Loschbour (Europe)**	**Stuttgart (Europe)**	**Yoruba (Africa)**	**French (Europe)**	**Han (Asia)**
Synonymous		**1,784 (54.1%)**	**2,790 (45.7%)**	2,906 (55.4%)	2,200 (54.2%)	2,112 (55.6%)
Non-synonymous		**1,512 (45.9%)**	**3,320 (54.3%)**	2,340 (44.6%)	1,859 (45.8%)	1,687 (44.4%)
Phast Cons	**Benign**	**885 (58.7%)**	**1,674 (50.7%)**	1,614 (65.0%)	1,186 (64.1%)	1,070 (63.6%)
	Deleterious	**623 (41.3%)**	**1,631 (49.3%)**	817 (35.0%)	665 (35.9%)	612 (36.4%)

Table S6.2. *Distribution of derived alleles by functional class in homozygous derived positions*

		Early modern humans		Present-day humans		
Category		**Loschbour (Europe)**	**Stuttgart (Europe)**	**Yoruba (Africa)**	**French (Europe)**	**Han (Asia)**
Synonymous		**21,149 (60.2%)**	**20,940 (60.3%)**	20,714 (60.5%)	20,936 (60.3%)	20,984 (60.3%)
Non-synonymous		**13,967 (39.8%)**	**13,765 (39.7%)**	13,515 (39.5%)	13,760 (39.7%)	13,804 (39.7%)
Phast Cons	**Benign**	**10,402 (74.8%)**	**10,281 (75.0%)**	10,889 (75.0%)	10,272 (75.0%)	10,306 (75.0%)
	Deleterious	**3,502 (25.2%)**	**3,424 (25.0%)**	3,365 (25.0%)	3,426 (25.0%)	3,436 (25.0%)

Counts in these tables were obtained from genotype calls for each individual in a combined VCF file. A coding site in the longest transcript of 17,367 genes from the CDDS[2], RefSeq[3] and GENCODE[4] annotations was considered for analysis when the following set of filters is met: (1) a GATK call was made; (2) the genotype quality (GQ) is at least 20; (3) there is a mapability score of 1 in the Duke 20mer uniqueness score (Map20); 4) the fraction of mapped reads with Mapping Quality (MQ) of zero is less than 10%; (5) coverage is within the 95% of the exome coverage; (6) the site is not flagged as a systematic error; (7) the site is not flagged as LowQuality (SNP quality, QUAL, is at least 30); (8) the site is derived according to human-chimpanzee ancestry information from the EPO 6 primate genome alignments[5,6]; (9) the human-chimpanzee ancestral allele matches one of the two alleles at heterozygous sites; (10) human and chimpanzee appear no more than once in the EPO alignment block.

We used only those coding sites that passed all quality filters in all the individuals compared. In this way, the absolute number of derived sites can be compared between individuals with different sequence

coverage in Tables S6.1 and S6.2. The derived allele in these sites was classified as synonymous and non-synonymous based on the gene annotations described above. We assumed that non-synonymous derived alleles in a position with a PhastCons[1] posterior probability larger than 0.9 are likely to alter protein structure or function. These alleles are likely to be slightly deleterious as they segregate in highly conserved positions from mammalian alignments that exclude the human reference sequence.

The fraction of non-synonymous derived alleles in both heterozygous and homozygous positions is equal or larger in the protein-coding regions of the Loschbour and Stuttgart genomes than in present-day humans (Tables S6.1 and S6.2). Using the same data, we also computed the Neutrality Index (NI)[7] for each individual analyzed (Table S6.3). The neutrality index is defined as the ratio of the number of non-synonymous and synonymous polymorphism (Pn/Ps) and substitution (Dn/Ds) ratios:

$$NI = (Pn/Ps)/(Dn/Ds).$$

Under strict neutrality, the two ratios Pn/Ps and Dn/Ds are expected to be the same. Thus, NI > 1 indicates an excess of amino acid polymorphism and NI < 1 an excess of amino acid substitutions. All individuals show an excess of amino acid polymorphism (Table S6.3), but the excess in Stuttgart is larger (NI = 1.80). The excess of amino acid polymorphism in Loschbour is in the upper range of present-day humans.

Table S6.3. *Neutrality Index*

	Loschbour (Europe)	Stuttgart (Europe)	Yoruba (Africa)	French (Europe)	Han (Asia)
Pn/Ps	**0.85**	**1.19**	0.81	0.85	0.80
Dn/Ds	**0.66**	**0.66**	0.65	0.66	0.66
NI	**1.29**	**1.80**	1.25	1.29	1.21

We tested whether Loschbour and Stuttgart have a larger fraction of derived alleles inferred to be deleterious in heterozygous positions than the present-humans analyzed. The fraction of non-synonymous derived alleles in heterozygous positions inferred to be deleterious in Loschbour (41.3%) is significantly larger than in Yoruba (35.0%; G-test; $P = 1.13 \times 10^{-6}$), French (35.9%; $P = 1.4 \times 10^{-3}$) and Han (36.4%; $P = 4.3 \times 10^{-3}$). Similarly, the fraction of non-synonymous derived alleles in heterozygous positions inferred to be deleterious in Stuttgart (49.3%) is significantly larger than in Yoruba (35.0%; G-test; $P < 1 \times 10^{-10}$), French (35.9%; $P < 1 \times 10^{-10}$) and Han (36.4%; $P < 1 \times 10^{-10}$).

Because C to T and G to A deaminations have highly elevated error rates in ancient DNA, and can still affect data even after UDG-treatment, we tested whether the higher proportion of deleterious alleles in early modern humans holds for substitutions other than deaminations. The fraction of non-synonymous derived alleles in heterozygous positions inferred to be deleterious in the Loschbour (41.3%), but not Stuttgart (37.2%), remains larger than in the Yoruba (39.8%), French (39.9%) and Han (37.6%) individuals. The fact that there is a large change in this quantity for Stuttgart but not for Loschbour suggests that DNA damage may be causing more false-polymorphisms in Stuttgart.

In conclusion, these results point to either a less effective removal of slightly deleterious mutations in Loschbour or a population bottleneck in Loschbour history, which would increase the relative fraction of deleterious mutations in protein-coding regions[8]. In either case, the observation points to a history of smaller population size in Loschbour than in Stuttgart and present-day humans since their separation.

References

1. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-50 (2005).
2. Pruitt, K.D. *et al.* The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res* **19**, 1316-23 (2009).
3. Pruitt, K.D., Tatusova, T., Klimke, W. & Maglott, D.R. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res* **37**, D32-6 (2009).
4. Coffey, A.J. *et al.* The GENCODE exome: sequencing the complete human exome. *Eur J Hum Genet* **19**, 827-31 (2011).
5. Paten, B., Herrero, J., Beal, K., Fitzgerald, S. & Birney, E. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res* **18**, 1814-28 (2008).
6. Paten, B. *et al.* Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res* **18**, 1829-43 (2008).
7. Rand, D.M. & Kann, L.M. Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from Drosophila, mice, and humans. *Mol Biol Evol* **13**, 735-48 (1996).
8. Lohmueller, K.E. *et al.* Proportionally more deleterious genetic variation in European than in African populations. *Nature* **451**, 994-7 (2008).

Supplementary Information 7
Phenotypic inference

Karola Kirsanow*

* To whom correspondence should be addressed (kirsanow@uni-mainz.de)

Introduction
Walsh et al. (*1-4*), among others (*5-7*), demonstrated that it is possible to predict human eye hair, and skin color phenotypes with accuracy using a small number of DNA variants.

We made predictions of pigmentation using two models (*2, 3, 6, 7*) that have been validated in present-day populations (*1, 4, 7, 8*) as well as on skeletal remains (*9*). We used these models to infer the most likely iris, hair and skin pigmentation for the Loschbour and Stuttgart individuals.

Table S7.1. Genotypes for SNPs associated with pigmentation in Loschbour and Stuttgart

8plex SNPs			
SNP	***Gene***	**Loschbour**	**Stuttgart**
rs1291382	*HERC2*	G/G	A/A
rs1545397	*OCA2*	A/A	A/A
rs16891982	*SLC45A2*	C/C	C/C
rs885479	*MC1R*	G/G	G/G
rs1426654	*SLC24A5*	G/G	A/A
rs12896399	*SLC24A4*	G/G°	T/T
rs6119471	*ASIP*	C/C	C/C
rs12203592	*IRF4*	T/T	C/C
Hiriplex SNPs			
SNP	**Gene**	**Loschbour**	**Stuttgart**
n29insa	*MC1R*	C/C	C/C
rs11547464	*MC1R*	G/G	G/G
rs885479	*MC1R*	G/G	G/G
rs18050008	*MC1R*	C/C	C/C
rs18050005	*MC1R*	G/G	G/G
rs18050006	*MC1R*	C/C	C/C
rs18050007	*MC1R*	C/C	C/C
rs1805009	*MCIR*	G/G	G/G
y152och	*MC1R*	C/C	C/C
rs2228479	*MC1R*	G/G°	G/G
rs1110400	*MC1R*	T/T	T/T
rs28777	*SLC45A2*	C/A	C/C
rs16891982	*SLC45A2*	C/C	C/C
rs12821256	*KITLG*	T/T	T/T
rs4959270	*EXOC2*	A/A	C/C
rs12203592	*IRF4*	T/T	C/C
rs1042602	*TYR*	C/C	C/A
rs1800407	*OCA2*	C/C	C/C°
rs2402130	*SLC24A4*	G/A	A/A
rs12913832	*OCA2/HERC2*	G/G	A/A
rs2378249	*PIGU/ASIP*	A/A	A/A
rs12896399	*SLC24A4*	G/G°	T/T
rs1393350	*TYR*	G/G	G/G
rs683	*TYRP1*	A/A	A/A

**These SNPs had genotype quality between 20 and 30, but passed other quality filters.*

We also analyzed the data for the two ancient humans at 35 single nucleotide polymorphisms (SNPs) known from genome-wide association studies (GWAS) to be reproducibly associated with susceptibility to the Metabolic Syndrome (MetS) and compared the results of two different diabetes-risk score models incorporating 24 of these SNPs (*10, 11*). MetS-related SNPs have evidence of being

under recent selection (*12, 13*), possibly because of pressures related to changes in diet and climate associated with human migration and the adoption of agriculture.

We finally analyzed the genotypes of Loschbour and Stuttgart at sites that are known to affect human phenotypes and have been identified as affected by selection in recent human history (*14-18*).

Methods

We analyzed DNA polymorphism data stored in the VCF format (*19*) using the VCFtools software package (htt://vcftoools.sourceforge.net/). We included data from sites not flagged as LowQuality, with genotype quality (GQ) of ≥30, and SNP quality (QUAL) of ≥50.

We carried out three sets of phenotypic analyses:

(1) We assessed the Loschbour forager and Stuttgart farmer individuals for their genotype at pigmentation SNPs included in the 8-plex system and the Hirisplex system (Table S7.1). We assigned probabilities to phenotypes using the Hirisplex Microsoft Excel macro (*2*).

(2) We assessed the Loschbour and Stuttgart samples for the genotypes at a panel of SNPs with evidence for recent natural selection, including several known to show high allele frequency differentiation between European and East Asian populations (Table S7.2).

(3) We assessed the Loschbour and Stuttgart samples for their genotypes at a panel of SNPs associated with risk for Metabolic Syndrome (Table S4.3) and that form the basis for two type 2 diabetes (T2D) risk score models (*10, 11*). We computed weighted genotype risk scores using the methods described in Meigs (2008) and Cornelis (2009).

We caution that the pigmentation phenotype models and metabolic syndrome risk scoring models are not independent. In particular, seven core pigmentation SNPs are shared with the pigmentation prediction models, and four metabolic syndrome-associated SNPs are shared between the diabetes risk score models.

Table S7.2. Genotypes for SNPs known to be under selection in Loschbour and Stuttgart

SNP	Gene	Loschbour	Stuttgart
rs182549	*LCTb*	C/C	C/C
rs4988235	*MCM6/LCTa*	G/G	G/G
rs699	*AGT*	A/G	A/G
rs4590952	*KITLG*	A/G	G/G
rs2740574	*CYP3A4*	T/T	T/T
rs776746	*CYP3A5*	C/C	C/C
rs3827760	*EDAR*	A/A	A/A
rs671	*ALDH2*	G/G	G/G

** The Loschbour forager could not be genotyped at the ADH1Bb locus.*

Results

Pigmentation

For hair color, the integrated results of the genotype-based pigmentation models indicate that there is at least a 99% probability that both the Stuttgart and Loschbour individuals had dark (brown or black) hair. The Hirisplex model assigns the highest probability to black hair color for both individuals (Table S7.4).

The results of the 8-plex skin pigmentation model were inconclusive for both the Loschbour and Stuttgart individuals. However, the Loschbour and Stuttgart genotypes at rs1426654 in *SLC24A5* indicate that the Stuttgart individual may have had lighter skin than the Loschbour hunter and gatherer.

The Loschbour individual is homozygous for the rs1426654 ancestral allele, while Stuttgart is homozygous for the derived skin-lightening allele (*22, 23*).

Table S7.3. Metabolic syndrome SNPs assessed in Loschbour and Stuttgart, by risk score model.

Core metabolic syndrome associated SNPs			
SNP	**Gene**	**Loschbour**	**Stuttgart**
rs7923837	*HHEX*	G/G	A/A
rs5015480	*HHEX/IDE*	C/C	T/T
rs3802678	*GBF1*	A/A	A/T
rs6235	*PCSK1*	C/C	G/G
rs7756992	*CDKAL1*	A/G	A/G
rs6446482	*WFS1*	C/G	C/G
rs11037909	*EXT2*	T/C	T/C
rs6698181	*PKN2*	T/T	C/T
rs17044137	*FLJ39370*	T/A	T/A
rs12255372	*TCF7L2*	G/G	G/G
rs7480010	*LOC387761*	A/A	A/A
rs11634397	*ZFAND6*	A/G	G/G
rs10946398	*CDKAL1*	A/C	C/C
rs8050136	*FTO*	A/A	C/A
Meigs 2008			
SNP	**Gene**	**Loschbour**	**Stuttgart**
rs7903146	*TCF7L2*	C/C	C/C
rs1470579	*IGF2BP2*	A/C	A/A
rs10811661	*CDKN2A/B*	T/C	T/T
rs864745	*JAZF1*	T/C	C/C
rs5219	*KCNJ11*	*	T/C
rs5215*	*KCNJ11*	C/T	C/T
rs12779790	*CDC123/CAMK1D*	A/G	A/A
rs7578597	*THADA*	T/T	T/T
rs7754840	*CDKAL1*	G/C	C/C
rs7961581	*TSPAN8/LGR5*	T/T	C/T
rs4607103	*ADAMTS9*	C/C	C/C
rs1111875	*HHEX*	C/C	T/T
rs10923931	*NOTCH2*	G/T	G/T
rs13266634	*SLC30A8*	C/C	C/C
rs1153188	*DCD*	T/T	T/A
rs1801282	*PPARG*	C/C	C/C
rs9472138	*VEGFA*	C/C	C/C
rs10490072	*BCL11A*	T/C	T/T
rs689	*INS*	A/T	A/T
Weighted genotype risk score		**118.0**	**101.6**
Metabolic-syndrome associated SNPs			
SNP	**Gene**	**Loschbour**	**Stuttgart**
rs564398	*CDKN2A/B*	C/T	T/T
rs1001031	*WFS1*	A/G	A/G
rs7754840	*CDKAL1*	G/C	C/C
rs4402960	*IGF2BP2*	G/T	G/G
rs1801282	*PPARG*	C/C	C/C
rs5219	*KCNJ11*	*	T/C
rs5215*	*KCNJ11*	C/T	C/T
rs1111875	*HHEX*	C/C	T/T
rs13266634	*SLC30A8*	C/C	C/C
rs10811661	*CDKN2A/B*	T/C	T/T
rs7901695	*TCF7L2*	T/T	T/T°
Rs7903146°	*TCF7L2*	C/C	C/C
Weighted genotype risk score		**10.6**	**10.7**

**For the purpose of computing the Weighted Genotype Risk Score, we use rs5215 as a proxy for rs5219, which failed to pass the quality filter for the Loschbour sample. These two SNPs are in strong LD (r^2=0.90)(20).*

°rs7903146 was used as a proxy for rs901695, which for the Stuttgart individual failed to pass the quality filter. The two SNPs are in strong LD (r^2=0.98) (21).

For eye color, the single most significant determinant is the rs12913832 SNP in the *HERC2* gene. The genotype at this site excludes the possibility that the Stuttgart farmer had blue eyes. Positive iris color determinations are less secure. The Loschbour forager is homozygous for the derived allele at rs12913832, indicating that this individual is likely to have had blue (52% probability) or intermediate iris color (27% probability). It has been suggested that this mutation arose within the last 6,000 to 10,000 years, and thus the Loschbour individual would have been a relatively early carrier (*24*).

Table S7.4. Hirisplex model probability scores for pigmentation.	**Loschbour**	**Stuttgart**
HAIR	**Probability**	**Probability**
Brown	0.256	0.079
Red	0	0
Black	0.734	0.917
Blond	0.01	0.004
HAIR SHADE	**Probability**	**Probability**
Light	0.025	0.002
Dark	0.975	0.998
EYE	**Probability**	**Probability**
Blue	0.524	0
Intermediate	0.268	0.006
Brown	0.207	0.994

Metabolic Syndrome Risk Score

Complex human disease phenotypes are less amenable to genotype-based prediction than externally visible characteristics such as pigmentation. The diabetes risk scoring systems developed to date thus do not have strong predictive power at the population level (*10*). Nevertheless, we used these scoring systems to begin to characterize the metabolic genotypes of the two ancient modern humans in comparison with the average modern non-diabetic genotype.

We find that the two ancient modern humans display metabolic syndrome-associated allele spectra comparable to those observed in present-day Europeans.

The Meigs 2008 model indicates a higher T2D risk for the Loschbour individual relative to Stuttgart. The weighted genotype risk scores for both Loschbour and Stuttgart fall within the overlapping one standard deviation ranges of the present-day diabetic and non-diabetic ranges as predicted by this model.

The Cornelis 2009 model predicts a roughly equal risk for both individuals (Table S7.3). According to this model, the weighted genotype risk scores of both the Loschbour and Stuttgart individuals (10.7 and 10.6, respectively) are within the 95% CI of that of the median present-day non-diabetic individual (10.4).

Overall, the risk allele is the ancestral allele at 19 out of the 35 MetS-associated SNPs whose genotypes we evaluate. The Loschbour and Stuttgart individuals carried similar numbers of ancestral MetS-associated risk alleles (21 for Loschbour and 19 for Stuttgart), and derived MetS risk alleles (14 for Loschbour and 15 for Stuttgart). Moreover, the MetS risk scores of the ancient forager and farmer do not indicate any significant departures from the MetS risk score averages in present-day Europeans.

Other phenotypic characteristics

We also assessed the Loschbour and Stuttgart individuals for their genotype at nine SNPs with well-validated phenotypic associations and evidence for recent positive selection.

Both ancient modern humans are homozygous for the ancestral alleles at the LCTa and LCTb polymorphisms and as a result are predicted to have been unable to digest lactose as adults. The LCTa mutation has been estimated to have first experienced positive selection between 6,256 and 8,683 years ago in central Europe (*25*). Thus, although the allele is associated with the spread of the LBK culture, it is likely to have been uncommon in early LBK populations, consistent with our results.

The heterozygous state of both the Stuttgart and Loschbour individuals at a SNP in the *AGT* gene suggests that they may have had a slightly increased risk of hypertension. The risk allele in the AGT gene is an ancestral allele. The derived protective allele is estimated to have arisen 22,500-44,500 years ago (*15*).

Both ancient modern humans were homozygous for a derived allele at *CYP3A4*, which is thought to confer protection from certain forms of cancer and is also possibly associated with protection from rickets (*26*). Both samples are also homozygous for the derived allele at *CYP3A5*, which is estimated to have arisen ~75,000 years ago (*27*), and which affects drug metabolism.

Finally, we evaluated the Loschbour and Stuttgart individuals for their genotypes at SNPs in *EDAR*, *ADH1B*, *ABCC1*, and *ALDH2* that are known to have high allele frequency differentiation between present-day European and East Asian populations. Both Stuttgart and Loschbour are homozygous for alleles associated with wet earwax (*ABCC1*) and non-shoveled incisors (*EDAR*), which are phenotypes known to occur at higher frequency in Europeans (*28-31*). Neither of the ancient modern humans carried the derived alleles at three loci associated with alcohol metabolism (*ALDH2*, *ADH1Ba* and *ADH1Bb*), which are known to have been under recent positive selection in East Asian populations (*17, 18, 32*) .

References

1. S. Walsh, A. Wollstein, F. Liu, U. Chakravarthy, M. Rahu, J. H. Seland, G. Soubrane, L. Tomazzoli, F. Topouzis, J. R. Vingerling, J. Vioque, A. E. Fletcher, K. N. Ballantyne, M. Kayser, DNA-based eye colour prediction across Europe with the IrisPlex system. *Forensic Science International: Genetics* **6**, 330-340 (2012)http://dx.doi.org/10.1016/j.fsigen.2011.07.009).
2. S. Walsh, F. Liu, A. Wollstein, L. Kovatsi, A. Ralf, A. Kosiniak-Kamysz, W. Branicki, M. Kayser, The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. *Forensic science international. Genetics* **7**, 98-115 (2013); published online EpubJan (10.1016/j.fsigen.2012.07.005).
3. S. Walsh, F. Liu, K. N. Ballantyne, M. van Oven, O. Lao, M. Kayser, IrisPlex: A sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. *Forensic Science International: Genetics* **5**, 170-180 (2011)http://dx.doi.org/10.1016/j.fsigen.2010.02.004).
4. S. Walsh, A. Lindenbergh, S. B. Zuniga, T. Sijen, P. de Knijff, M. Kayser, K. N. Ballantyne, Developmental validation of the IrisPlex system: Determination of blue and brown iris colour for forensic intelligence. *Forensic Science International: Genetics* **5**, 464-471 (2011)http://dx.doi.org/10.1016/j.fsigen.2010.09.008).
5. W. Branicki, F. Liu, K. van Duijn, J. Draus-Barini, E. Pośpiech, S. Walsh, T. Kupiec, A. Wojas-Pelc, M. Kayser, Model-based prediction of human hair color using DNA variants. *Hum Genet* **129**, 443-454 (2011)10.1007/s00439-010-0939-8).
6. K. L. Hart, S. L. Kimura, V. Mushailov, Z. M. Budimlija, M. Prinz, E. Wurmbach, Improved eye- and skin-color prediction based on 8 SNPs. *Croatian Medical Journal* **54**, 248-256 (2013)10.3325/cmj.2013.54.248).

7. O. Spichenok, Z. M. Budimlija, A. A. Mitchell, A. Jenny, L. Kovacevic, D. Marjanovic, T. Caragine, M. Prinz, E. Wurmbach, Prediction of eye and skin color in diverse populations using seven SNPs. *Forensic science international. Genetics* **5**, 472-478 (2011); published online EpubNov (10.1016/j.fsigen.2010.10.005).
8. A. Pneuman, Z. M. Budimlija, T. Caragine, M. Prinz, E. Wurmbach, Verification of eye and skin color predictors in various populations. *Leg Med (Tokyo)* **14**, 78-83 (2012); published online EpubMar (10.1016/j.legalmed.2011.12.005).
9. J. Draus-Barini, S. Walsh, E. Pośpiech, T. Kupiec, H. Głąb, W. Branicki, M. Kayser, Bona fide colour: DNA prediction of human eye and hair colour from ancient and contemporary skeletal remains. *Investigative genetics* **4**, 3 (2013).
10. M. Cornelis, L. Qi, C. Zhang, P. Kraft, J. Manson, T. Cai, D. J. Hunter, F. B. Hu, Joint effects of common genetic variants on the risk for type 2 diabetes in U.S. men and women of European ancestry. *Ann Intern Med* **150**, 541-550 (2009).
11. J. B. Meigs, P. Shrader, L. M. Sullivan, J. B. McAteer, C. S. Fox, J. Dupuis, A. K. Manning, J. C. Florez, P. W. Wilson, R. B. D'Agostino, Sr., L. A. Cupples, Genotype score in addition to common risk factors for prediction of type 2 diabetes. *N Engl J Med* **359**, 2208-2219 (2008); published online EpubNov 20 (10.1056/NEJMoa0804742).
12. L. B. Barreiro, G. Laval, H. Quach, E. Patin, L. Quintana-Murci, Natural selection has driven population differentiation in modern humans. *Nat Genet* **40**, 340-345 (2008); published online EpubMar (Doi 10.1038/Ng.78).
13. E. Corona, J. T. Dudley, A. J. Butte, J. Hawks, Extreme Evolutionary Disparities Seen in Positive Selection across Seven Complex Diseases. *PLoS ONE* **5**, e12236 (2010).
14. J. Zeron-Medina, X. Wang, E. Repapi, Michelle R. Campbell, D. Su, F. Castro-Giner, B. Davies, Elisabeth F. P. Peterse, N. Sacilotto, Graeme J. Walker, T. Terzian, Ian P. Tomlinson, Neil F. Box, N. Meinshausen, S. De Val, Douglas A. Bell, Gareth L. Bond, A Polymorphic p53 Response Element in KIT Ligand Influences Cancer Risk and Has Undergone Natural Selection. *Cell* **155**, 410-422 (2013)10.1016/j.cell.2013.09.017).
15. T. Nakajima, S. Wooding, T. Sakagami, M. Emi, K. Tokunaga, G. Tamiya, T. Ishigami, S. Umemura, B. Munkhbat, F. Jin, J. Guan-jun, I. Hayasaka, T. Ishida, N. Saitou, K. Pavelka, J.-M. Lalouel, L. B. Jorde, I. Inoue, Natural Selection and Population History in the Human Angiotensinogen Gene (AGT): 736 Complete AGT Sequences in Chromosomes from Around the World. *The American Journal of Human Genetics* **74**, 898-916 (2004)10.1086/420793).
16. E. E. Thompson, H. Kuttab-Boulos, D. Witonsky, L. Yang, B. A. Roe, A. Di Rienzo, CYP3A variation and the evolution of salt-sensitivity variants. *American Journal of Human Genetics* **75**, 1059-1069 (2004); published online EpubDec (10.1086/426406).
17. H. Li, S. Gu, Y. Han, Z. Xu, A. J. Pakstis, L. Jin, J. R. Kidd, K. K. Kidd, Diversification of the ADH1B gene during expansion of modern humans. *Ann Hum Genet* **75**, 497-507 (2011); published online EpubJul (10.1111/j.1469-1809.2011.00651.x).
18. H. Oota, A. J. Pakstis, B. Bonne-Tamir, D. Goldman, E. Grigorenko, S. L. B. Kajuna, N. J. Karoma, S. Kungulilo, R.-B. Lu, K. Odunsi, F. Okonofua, O. V. Zhukova, J. R. Kidd, K. K. Kidd, The evolution and population genetics of the ALDH2 locus: random genetic drift, selection, and low levels of recombination. *Annals of Human Genetics* **68**, 93-109 (2004)10.1046/j.1529-8817.2003.00060.x).
19. P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, 1000 Genomes Project Analysis Group, The variant call format and VCFtools. (2011); published online Epub2011-08-01 (10.1093/bioinformatics/btr330).
20. C. Wellcome Trust Case Control, Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-678 (2007); published online EpubJun 7 (http://www.nature.com/nature/journal/v447/n7145/suppinfo/nature05911_S1.html).
21. J. Vcelak, D. Vejrazkova, M. Vankova, P. Lukasova, O. Bradnova, T. Halkova, J. Bestak, K. Andelova, H. Kvasnickova, P. Hoskovcova, K. Vondra, J. Vrbikova, B. Bendlova, T2D risk haplotypes of the TCF7L2 gene in the Czech population sample: the association with free fatty acids composition. *Physiol Res* **61**, 229-240 (2012); published online EpubJul 20 (

22. R. L. Lamason, M.-A. P. K. Mohideen, J. R. Mest, A. C. Wong, H. L. Norton, M. C. Aros, M. J. Jurynec, X. Mao, V. R. Humphreville, J. E. Humbert, S. Sinha, J. L. Moore, P. Jagadeeswaran, W. Zhao, G. Ning, I. Makalowska, P. M. McKeigue, D. O'Donnell, R. Kittles, E. J. Parra, N. J. Mangini, D. J. Grunwald, M. D. Shriver, V. A. Canfield, K. C. Cheng, SLC24A5, a Putative Cation Exchanger, Affects Pigmentation in Zebrafish and Humans. *Science* **310**, 1782-1786 (2005); published online EpubDecember 16, 2005 (10.1126/science.1116238).
23. S. Beleza, N. A. Johnson, S. I. Candille, D. M. Absher, M. A. Coram, J. Lopes, J. Campos, Araujo, II, T. M. Anderson, B. J. Vilhjalmsson, M. Nordborg, E. S. A. Correia, M. D. Shriver, J. Rocha, G. S. Barsh, H. Tang, Genetic architecture of skin and eye color in an African-European admixed population. *PLoS Genet* **9**, e1003372 (2013); published online EpubMar (10.1371/journal.pgen.1003372).
24. H. Eiberg, J. Troelsen, M. Nielsen, A. Mikkelsen, J. Mengel-From, K. W. Kjaer, L. Hansen, Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression. *Hum Genet* **123**, 177-187 (2008).
25. Y. Itan, A. Powell, M. A. Beaumont, J. Burger, M. G. Thomas, The origins of lactase persistence in Europe. *PLoS Computational Biology* **5**, e1000491 (2009); published online EpubAug (10.1371/journal.pcbi.1000491).
26. M. Schirmer, M. R. Toliat, M. Haberl, A. Suk, L. K. Kamdem, K. Klein, J. Brockmöller, P. Nürnberg, U. M. Zanger, L. Wojnowski, Genetic signature consistent with selection against the CYP3A4*1B allele in non-African populations. *Pharmacogenetics and genomics* **16**, 59-71 (2006).
27. R. K. Bains, M. Kovacevic, C. A. Plaster, A. Tarekegn, E. Bekele, N. N. Bradman, M. G. Thomas, Molecular diversity and population structure at the Cytochrome P450 3A5 gene in Africa. (2013).
28. R. Kimura, T. Yamaguchi, M. Takeda, O. Kondo, T. Toma, K. Haneji, T. Hanihara, H. Matsukusa, S. Kawamura, K. Maki, A Common Variation in< i> EDAR</i> Is a Genetic Determinant of Shovel-Shaped Incisors. *The American Journal of Human Genetics* **85**, 528-535 (2009).
29. P. C. Sabeti, P. Varilly, B. Fry, J. Lohmueller, E. Hostetter, C. Cotsapas, X. Xie, E. H. Byrne, S. A. McCarroll, R. Gaudet, S. F. Schaffner, E. S. Lander, K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, S. Pasternak, D. A. Wheeler, T. D. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu, H. Pan, X. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. Zhang, H. Zhao, J. Zhou, S. B. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta, J. Moore, H. Nguyen, R. C. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, D. Altshuler, Y. Shen, Z. Yao, W. Huang, X. Chu, Y. He, L. Jin, Y. Liu, W. Sun, H. Wang, Y. Wang, X. Xiong, L. Xu, M. M. Waye, S. K. Tsui, H. Xue, J. T. Wong, L. M. Galver, J. B. Fan, K. Gunderson, S. S. Murray, A. R. Oliphant, M. S. Chee, A. Montpetit, F. Chagnon, V. Ferretti, M. Leboeuf, J. F. Olivier, M. S. Phillips, S. Roumy, C. Sallee, A. Verner, T. J. Hudson, P. Y. Kwok, D. Cai, D. C. Koboldt, R. D. Miller, L. Pawlikowska, P. Taillon-Miller, M. Xiao, L. C. Tsui, W. Mak, Y. Q. Song, P. K. Tam, Y. Nakamura, T. Kawaguchi, T. Kitamoto, T. Morizono, A. Nagashima, Y. Ohnishi, A. Sekine, T. Tanaka, T. Tsunoda, P. Deloukas, C. P. Bird, M. Delgado, E. T. Dermitzakis, R. Gwilliam, S. Hunt, J. Morrison, D. Powell, B. E. Stranger, P. Whittaker, D. R. Bentley, M. J. Daly, P. I. de Bakker, J. Barrett, Y. R. Chretien, J. Maller, S. McCarroll, N. Patterson, I. Pe'er, A. Price, S. Purcell, D. J. Richter, P. Sabeti, R. Saxena, P. C. Sham, L. D. Stein, L. Krishnan, A. V. Smith, M. K. Tello-Ruiz, G. A. Thorisson, A. Chakravarti, P. E. Chen, D. J. Cutler, C. S. Kashuk, S. Lin, G. R. Abecasis, W. Guan, Y. Li, H. M. Munro, Z. S. Qin, D. J. Thomas, G. McVean, A. Auton, L. Bottolo, N. Cardin, S. Eyheramendy, C. Freeman, J. Marchini, S. Myers, C. Spencer, M. Stephens, P. Donnelly, L. R. Cardon, G. Clarke, D. M. Evans, A. P. Morris, B. S. Weir, T. A. Johnson, J. C. Mullikin, S. T. Sherry, M. Feolo, A. Skol, H. Zhang, I. Matsuda, Y. Fukushima, D. R. Macer, E. Suda, C. N. Rotimi, C. A. Adebamowo, I. Ajayi, T. Aniagwu, P. A. Marshall, C. Nkwodimmah, C. D. Royal, M. F. Leppert, M. Dixon, A. Peiffer, R. Qiu, A. Kent, K. Kato, N. Niikawa, I. F. Adewole, B. M. Knoppers, M. W. Foster,

E. W. Clayton, J. Watkin, D. Muzny, L. Nazareth, E. Sodergren, G. M. Weinstock, I. Yakub, B. W. Birren, R. K. Wilson, L. L. Fulton, J. Rogers, J. Burton, N. P. Carter, C. M. Clee, M. Griffiths, M. C. Jones, K. McLay, R. W. Plumb, M. T. Ross, S. K. Sims, D. L. Willey, Z. Chen, H. Han, L. Kang, M. Godbout, J. C. Wallenburg, P. L'Archeveque, G. Bellemare, K. Saeki, D. An, H. Fu, Q. Li, Z. Wang, R. Wang, A. L. Holden, L. D. Brooks, J. E. McEwen, M. S. Guyer, V. O. Wang, J. L. Peterson, M. Shi, J. Spiegel, L. M. Sung, L. F. Zacharia, F. S. Collins, K. Kennedy, R. Jamieson, J. Stewart, Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913-918 (2007); published online EpubOct 18 (10.1038/nature06250).

30. A. Fujimoto, J. Ohashi, N. Nishida, T. Miyagawa, Y. Morishita, T. Tsunoda, R. Kimura, K. Tokunaga, A replication study confirmed the EDAR gene to be a major contributor to population differentiation regarding head hair thickness in Asia. *Human genetics* **124**, 179-185 (2008).
31. J. Ohashi, I. Naka, N. Tsuchiya, The impact of natural selection on an ABCC11 SNP determining earwax type. *Mol Biol Evol* **28**, 849-857 (2011); published online EpubJan (10.1093/molbev/msq264).
32. Y. Peng, H. Shi, X. B. Qi, C. J. Xiao, H. Zhong, R. L. Ma, B. Su, The ADH1B Arg47His polymorphism in east Asian populations and expansion of rice domestication in history. *BMC Evol Biol* **10**, 15 (2010)10.1186/1471-2148-10-15).

Supplementary Information 8

Analysis of segmental duplications and copy number variants

Peter H. Sudmant* and Evan E. Eichler

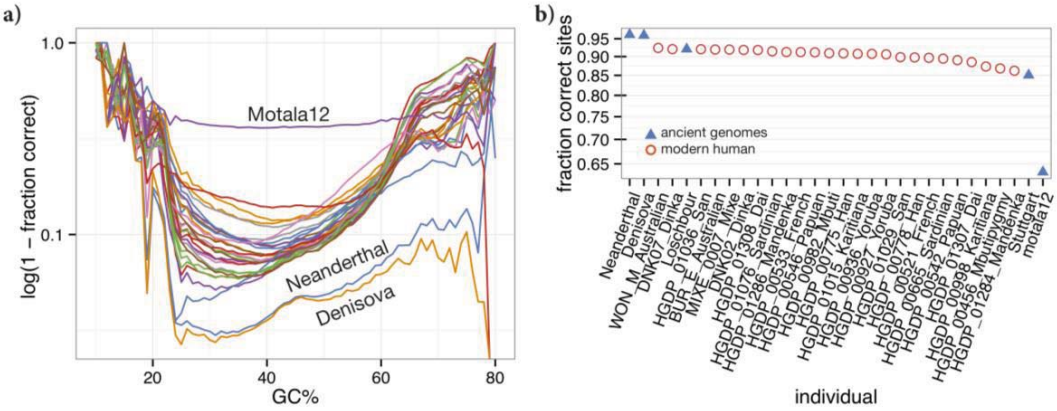
* To whom correspondence should be addressed (psudmant@gmail.com)

We constructed read-depth based copy number maps for the Loschbour, Stuttgart and Motala12 individuals, and co-analyzed them with whole genome sequence data from the archaic Denisova[1] and the archaic Altai Neandertal genome[2], as well as 25 deeply sequenced present-day human genomes that we have described previously[3].

These maps consist of copy number estimates across the genome in windows that are 500 unmasked base-pairs wide, which we slide by intervals of 100 unmasked base-pairs.

We first assessed the quality of each genome in regions putatively free of copy number variation[3], which allowed us to quantify our ability to accurately determine a diploid copy number state for each 500 bp window encompassed in these loci. As read-depth based copy number estimates are often affected by GC-associated sequencing biases we assessed our accuracy as a function of genomic GC% (Figure S8.1) and cumulatively across all regions examined. This is a fairly strict test as to actually call a copy number variant the aggregate signal of many windows is taken into account. All genomes with the exception of the low coverage Motala12 individual demonstrate a high fraction of correctly determined sites (>85%) with higher concordance in individuals sequenced to higher coverage, such as the Neandertal and Denisova genomes.

Figure S8.1: Quality control and copy number calling. *(a) The fraction of incorrectly correctly determined diploid loci is plotted as a function of GC content. (b) The total fraction of correctly determined diploid loci in each individual assessed in this study.*



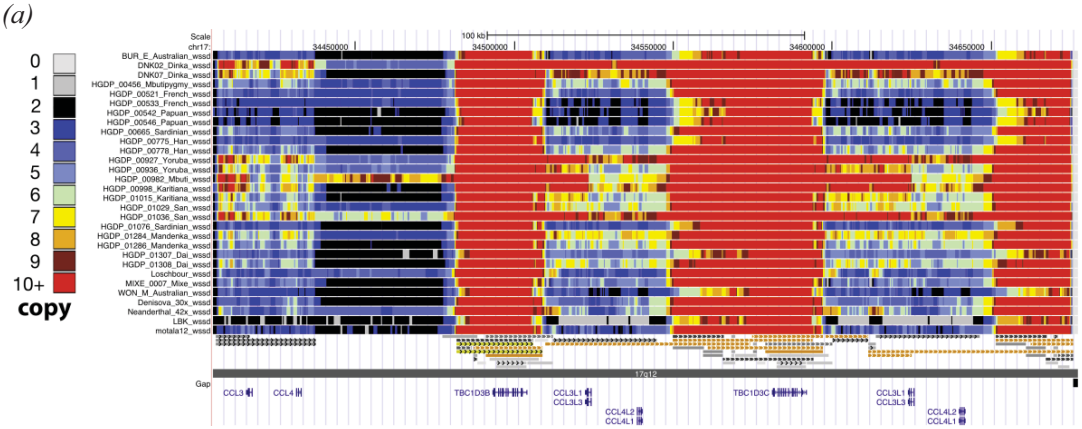
We next performed a genome-wide scan for copy number variants using digital array comparative genomic hybridization (dCGH)[4]. Briefly, for all-pairwise combinations of genomes, we calculate the $\log_2$-ratio across copy number windows. We then segment each of these $\log_2$-ratio maps using a scale-space filtering based technique[4]. We compute the significance of the putative copy number variants determined by the segmentation using a modified T-statistic to account for the autocorrelation of the underlying data. Putative CNV calls amongst individual pairs of genomes are merged by calculating the reciprocal overlap between overlapping calls and merging overlapping calls with cophenetic distance ≤ 0.85. We restricted to calls with a log-likelihood of $\leq$-6. We excluded the Motala12 individual from our initial scan due to its lower coverage.

We identified 3,846 putative copy number variants, 2,094 of which intersected segmental duplications. For segmentally duplicated CNVs in the Stuttgart and Loschbour individuals, the copy and position of segmental duplications is within the range of present-day humans.

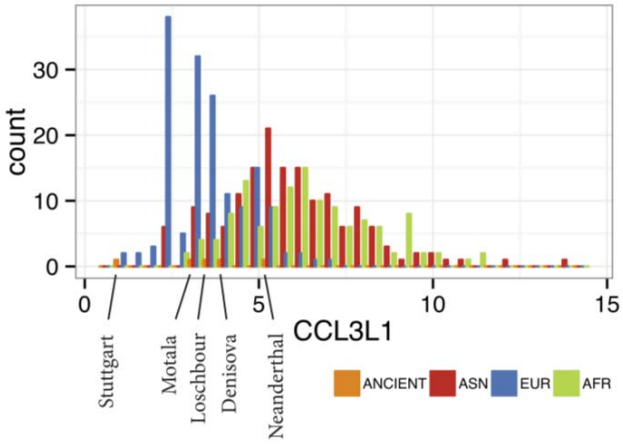
We focused on two loci for in-depth discussion. The first region is the *CCL3L1* locus on 17q12 (Figure S8.2). The *CCL3L1* gene encodes for a chemokine involved in immune and inflammatory processes. The copy number of *CCL3L1* varies widely among humans and is stratified between European and non-European populations (Figure S8.3). While European populations exhibit fewer copies of *CCL3L1* (a median of 2 copies), the Stuttgart individual has only a single copy of the locus, a state shared by only ~1.5% of individuals (as assessed from 1000 Genomes Project populations).

Figure S8.2. A copy number heat-map of the 17q13 locus. (a) *The Stuttgart individual exhibits a deletion of the locus encompassing the chemokine genes CCL3L1, CCL3L3, CCL4L1 and CCL4L2. Deletions of these genes occurs in ~1.5% of 1000 genomes individuals. (b) Distribution of CCL3L1 copy number in the 1000 genomes phase I data and the analyzed archaic genomes. The Stuttgart genome exhibits a single copy of CCL3L1, an outlier relative to the distribution.*

(a)

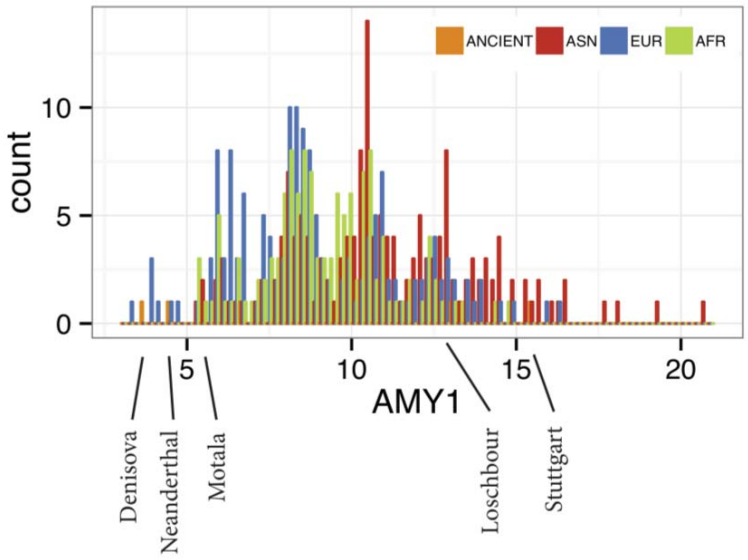


(b)



We also highlight the patterns at the amylase gene (*AMY1*), which has also recently expanded in human populations, potentially as a result of adaptations to start rich diets [5]. We recently reported that the Denisova and Neandertal genomes have the ancestral state of two copies of amylase. We find that Motala12, Loschbour and Stuttgart have 6, 13, and 16 copies of *AMY1* respectively, well within the range of current European populations. This suggests that amylase copy number expanded in *Homo sapiens* before the advent of agriculture. Further sequencing of early modern humans should help to refine the picture of the emergence of extra *AMY1* copies in humans.

Figure S8.3. Distribution of AMY1 copy number. *Both Stuttgart and Loschbour have high copy numbers of AMY1, while Motala12 has a low copy number.*



Overall, we identified 1,556 non-segmentally duplicated CNVs among the individuals assessed and reported genotypes for these in Supplementary Online Table 1**.** These include 76 deletions and 168 duplications in the Stuttgart individual and 68 deletions and 104 duplications in the Loschbour individual. These loci include loss of 4 olfactory genes and exon intersecting homozygous deletions of the *LCE3C* and *LCE3B* genes in Stuttgart. In the Loschbour individual, we identify the loss of 2 olfactory genes, the same homozygous *LCE3C* and *LCE3B* intersecting deletion, and a heterozygous deletion of the first exon of one isoform of *SLC25A24.* No Loschbour- or Stuttgart-specific events were identified, consistent with these individuals having variation within the range of present-day humans.

References

1. Meyer, M. *et al.* A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science* **338,** 222–226 (2012).
2. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* 1–13 (2013). doi:10.1038/nature12886
3. Sudmant, P. H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330,** 641–646 (2010).
4. Sudmant, P. H. *et al.* Evolution and diversity of copy number variation in the great ape lineage. *Genome Res.* **23,** 1373–1382 (2013).
5. Perry, G. H. *et al.* Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.* **39,** 1256–1260 (2007).

Supplementary Information 9

Affymetrix Human Origins genotyping dataset and ADMIXTURE analysis

Iosif Lazaridis, Nick Patterson, Susanne Nordenfelt, George Ayodo, Hamza A. Babiker, Elena Balanovska, Oleg Balanovsky, Haim Ben-Ami, Judit Bene, Fouad Berrada, Francesca Brisighelli, George B.J. Busby, Francesco Cali, Mikhail Churnosov, David E.C. Cole, Larissa Damba, George van Driem, Stanislav Dryomov, Sardana A. Fedorova, Irene Gallego Romero, Marina Gubina, Michael Hammer, Brenna Henn, Tor Hervig, Ugur Hodoglugil, Aashish R. Jha, Rick Kittles, Elza Khusnutdinova, Toomas Kivisild, Vaidutis Kučinskas, Rita Khusainova, Alena Kushniarevich, Leila Laredj, Sergey Litvinov, Robert W. Mahley, Béla Melegh, Ene Metspalu, Joanna Mountain, Thomas Nyambo, Ludmila Osipova, Jüri Parik, Fedor Platonov, Olga L. Posukh, Valentino Romano, Igor Rudan, Ruslan Ruizbakiev, Hovhannes Sahakyan, Antonio Salas, Elena B. Starikovskaya, Ayele Tarekegn, Draga Toncheva, Shahlo Turdikulova, Ingrida Uktveryte, Olga Utevska, Mikhail Voevoda, Pierre Zalloua, Levon Yepiskoposyan, Tatijana Zemunik, Cristian Capelli, Mark G. Thomas, Sarah A. Tishkoff, Lalji Singh, Kumarasamy Thangaraj, Richard Villems, David Comas, Rem Sukernik, Mait Metspalu, Svante Pääbo, David Reich*

* To whom correspondence should be addressed (reich@genetics.med.harvard.edu)

Overview of the Human Origins dataset

Here we describe the Affymetrix Human Origins dataset of single nucleotide polymorphism (SNP) genotyped in diverse present-day humans.

The array consists of 14 panels of SNPs for which the ascertainment is well understood[1, 2]. The array includes oligonucleotide probes that target of 627,421 SNPs of which 620,744 are on the autosomes, 4,331 are on chromosome X, 2,089 are on chromosome Y and 257 are on the mitochondrial DNA.

Genotypes of present-day humans the array have already been reported in two studies:

- Patterson et al. 2012[2] is the original technical description of the array (File S1 of that study). The study also released genotyping data from 934 samples from the CEPH / Human Genome Diversity Panel from diverse worldwide populations (ftp://ftp.cephb.fr/hgdp_supp10/).
- Pickrell et al. 2013[3] presented genotyping data of an additional 187 individuals from southern and eastern African populations.

Here we report data from many additional populations, filling in sampling gaps in the dataset especially in West Eurasia, and also adding in sampling from other world regions.

We took the genotypes from all samples combined them into a single file. We then carried out a comprehensive curation of the data (described in what follos) to identify a list of SNPs that appeared to perform reliably in genotyping, and to identify a list of samples that were not close relatives of others in the dataset or outliers relative to others from their own populations.

SNP filtering of the Human Origins Dataset

The genotyping was performed in seven batches over the course of several years. We were concerned that differences in the experimental or bioinformatic processing across batches might cause systematic differences in genotyping results for each batch that have nothing to do with population history. Moreover, a subset of samples were from whole genome amplified (WGA) material rather than from genomic DNA extracted from blood and saliva, and we were concerned that these samples might have systematic differences from the other samples. These populations comprised of samples derived from WGA material are identified with the suffix "_WGA" in the datasets

To curate the data, we began by computing the following statistics on each SNP:

(1) Genotyping concordance over 69 samples from the West African Yoruba (YRI) that overlapped between this study and low coverage sequencing data from the 1000 Genomes Project[4].

(2) Genotyping concordance rate over 25 samples from diverse populations that overlapped between this study and high coverage genome sequences reported in ref. 5.

(3) Completeness of genotyping (restricting to males on chromosome Y). This was performed for:
(a) All samples except WGA,
(b) Just WGA
(c) By genotyping batch excluding WGA (1, 2, 3, 4, 5, 6, 7)

(4) Alternate allele frequency of a pool of West Africans and a pool of West Eurasians in each batch.

(5) Homozygous, Heterozygous, and Variant genotype counts for a pool of West Africans and a pool of Europeans over all batches but excluding WGA samples. We include only females from chromosome X SNPs so that all genotypes are diploid.

(6) Male and female frequencies for a pool of West Africans and a pool of Europeans over all batches excluding WGA samples.

We first computed the following derived statistics to filter out potentially problematic SNPs.

• "Maxconcordance" – Maximum concordance with either the 69 1000 Genomes Project or 25 deeply sequenced samples. If a site has missing data in one of the sequencing datasets the concordance is reported as 0.

• "Completeness" – Completeness percentage of genotyping across the WGA samples

• "Mincompleteness" – Minimum completeness percentage for the SNP across 7 genotyping batches.

• "WGAcompleteness" – Completeness for SNP in the WGA data.

The results in Table S9.1 give the fraction of autosomal SNPs kept after applying different thresholds (for all but the WGAcompleteness metric).

Table S9.1: Fraction of SNPs retained with different concordance and completeness

Threshold:	50%	80%	85%	90%	95%	96%	97%	97.5%	98%	98.5%	99%	99.5%
Maxconcordance	0.997	0.996	0.996	0.996	0.994	0.992	0.991	**0.990**	0.976	0.974	0.963	0.919
Completeness	1.000	1.000	1.000	0.999	**0.995**	0.990	0.978	0.968	0.954	0.931	0.881	0.639
Mincompleteness	1.000	0.992	0.987	**0.975**	0.933	0.907	0.872	0.853	0.772	0.619	0.308	0.060

Note: We highlight in bold the thresholds we use for our main dataset.

We also pooled all non-whole genome amplified (non-WGA) West Eurasians and all non-WGA West Africans. This gave us counts of the three possible genotypes for each SNP: homozygous reference, heterozygous, and homozygous derived in a relatively homogeneous population. Using these counts, we computed Hardy-Weinberg-like statistics for all SNPs, looking for a deficiency of heterozygous sites as might be expected from poor allele calling:

$$HW = \sum_{i=1}^{3} \frac{(obs_i - exp_i)^2}{exp_i + 1} \quad \text{(S9.1)}$$

We added 1 to the expected allele counts in the denominator to deflate the statistics in the context of low expected values. This resulted in a West African HW statistic and a West Eurasian HW statistic. We imposed thresholds for significance based on a χ^2 distribution with 1 degree of freedom.

We next computed empirical derived allele frequencies for many different sample sets for each SNP *i*. We performed 21 different pairwise comparisons:

15 All West African pairwise comparisons for batches 1-6

3 All West Eurasian pairwise comparisons for batches 1, 6 and 7

1 All non-WGA male West Africans vs. all non-WGA female West Africans

1 All non-WGA male West Eurasians vs. all non-WGA female West Eurasians

1 All non-WGA West Eurasians vs. allele frequencies from randomly drawn reads from 107 YRI West Africans from the 1000 Genomes Project[4], computed as in Prufer et al.[6].

Consider two allele frequencies a_i and b_i for sample sets *A* and *B* respectively, in a subset of the genome (either all the autosomes, or just chromosome X) with *n* SNPs. Further define $\mu_i=(a_i+b_i)/2$ as the mean of these frequencies. Then we can compute the following statistic that is approximately χ^2 distributed with 1 degree of freedom. In the denominator, we normalize by the mean of the numerator over all SNPs. This is a form of "genomic control" that normalizes by the mean of this over-dispersed chi-square distribution, so that the statistic is well described by a χ^2 distribution with 1 d.f.

$$Stat_i = \left[\frac{(a_i-b_i)^2}{\mu_i(1-\mu_i)}\right] \Big/ \left[\frac{1}{n}\sum_{i=1}^{n}\frac{(a_i-b_i)^2}{\mu_i(1-\mu_i)}\right] \quad \text{S9.1}$$

In practice, we carried out our filtering by computing the maximum statistic "MaxStat$_i$" over 23 of the approximately χ^2 statistics that we analyzed (2 Hardy-Weinberg and 21 frequency comparisons). We then only accepted SNPs with "MaxStat$_i$" less than a specified threshold.

The threshold we use for our main analysis is 20. For this threshold, we empirically found that we removed almost no SNPs from the bulk of the distribution that was symmetrically spread around the $y=x$ axis (as might be expected from the fact that it corresponds to a P-value of ~10^{-5}, on the order of 1 divided by the number of SNPs in the dataset). However, this thresh-holding did remove a population of SNPs that had frequency of 0% in one population and were polymorphic in the other, which are clear genotyping failures suggesting that the filter is valuable.

Table S9.2 shows the filters we chose for the dataset. For the analyses reported in this study, we restrict to non-WGA samples and use thresholds that strike a balance between retaining a large fraction of SNPs while removing extreme outliers. For users who wish to analyze the data from WGA samples which have a higher missing data rate than the other samples and where the missing data is concentrated disproportionally at particular SNPs, we recommend imposing stronger thresholds.

Table S9.2. Summary of SNP filters used

	Maxconc-ordance*	Comp-leteness	Mincomp-leteness	WGA comp-leteness	Max of 23 χ^2 stats	SNPs killed	SNPs retained
Main dataset	>0.975	>0.95	>0.9	None	>20	25,131	602,290
With WGA data	>0.995	>0.99	>0.95	>0.99	>9	185,132	442,289

Note: We only remove SNPs on chromosomes 1-23. Users who wish to analyze the Y chromosome and mtDNA data should do so at their own discretion and need to design their own customized filtering.

In our paper we use 594,924 SNPs for all analyses; these are autosomal SNPs from the 602,290 SNPs indicated in Table S9.2, from which we further removed 1,449 when merging with the ancient samples and requiring a homologous chimpanzee allele, diallelic SNPs and a valid genetic distance. For genetic distance, we used the linkage disequilibrium-based map that includes chromosome X and which is available on the 1000 Genomes Project website at
http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20110106_recombination_hotspots/.

Filtering of samples
A total of 2,589 samples were successfully genotyped.

We manually curated the data using ADMIXTURE[7] and EIGENSOFT[8, 9] to identify samples that were visual outliers compared with samples from their own populations. We also identified samples that were apparently closely relatives of others in the dataset. In the dataset that we release, the population IDs for these individuals are prefixed by the string "Ignore_", but users who wish to analyze these samples are still able to access the data. A total of 2,303 samples remain after this outlier removal. Our paper focuses on a subset of 2,196 individuals in this dataset, omitting samples that were genotyped from whole genome amplified (WGA) material as their genotypes appear to be less reliable and applying the more stringent filtering criteria of Table S9.2 to allow us to use these samples would substantially reduce the number of SNPs available for joint analysis of all the data.

Table S9.3. Breakdown of Human Origins dataset by population

World Region	Populations	Samples before curation	Samples after curation
Africa	66	740	624
America	16	186	169
Central Asia / Siberia	17	190	163
EastAsia	22	251	243
Oceania	3	39	27
SouthAsia	22	329	280
WestEurasia	84	854	797
Total	230	2589	2303

Table S9.3 summarizes the geographic distribution of the dataset, while Table S9.4 presents detailed information on each of the populations. A total of 2,166 samples (of the 2,589) are in a version of the dataset we have made available at http://genetics.med.harvard.edu/reichlab/Reich_Lab/Datasets.html. The remaining samples have more restrictive procedures for data access due to the nature of the informed consent, and users who wish to access the data will need to send the corresponding author (DR) a signed letter containing the text shown in Box S9.1.

Box S9.1. Text that needs to be included in a letter to access the data not posted publicly

I affirm that
(a) I will not distribute the data outside my collaboration,
(b) I will not post it publicly,
(c) I will make no attempt to connect the genetic data to personal identifiers for the samples,
(d) I will use the data only for studies of population history,
(e) I will not use the data for any selection studies,
(f) I will not use the data for any medically or disease related analyses.
(g) I will not use the data for any commercial purposes

Note: Please send a PDF of a signed letter with this text to David Reich (reich@genetics.med.harvard.edu)

In summary, with this paper we are releasing genotyping data corresponding to two sets of samples

2,166 samples (1,946 after curation) that are fully publicly available

2,589 samples (2,303 after curation) for researchers who send a signed letter

For both of these sample sets, we include a "verbose.ind" list of populations that includes verbose sample identifiers which correspond to the 230 populations listed in Table S9.4 and plotted in Figure 1A. We also release a "simple.ind" list of sample identifier which corresponds to the merged groups of 185 populations and simpler names used for most of the analyses in the study.

We note that in practice, for each of these two sets of samples we are releasing 14 genotyping datasets – 14 genotype files and SNP lists. The file that the majority of researchers are likely to wish to use is the "allsnps" dataset that includes all SNPs. However, we also release separately called SNP datasets for each of the 13 panels in the Affymetrix Human Origins array for researchers who wish to take full advantage of the uniform ascertainment in the array.

Table S9.3. List of populations genotyped on the Human Origins array and record of curation
"n" is the sample size before curation and "m" after curation.

Verbose population ID	Simplified ID	n	m	Country	Region	Lat.	Long.	Source of samples
AA_Denver	AA	12	12	USA	Africa	39.7	-105	Rick Kittles
AA_Seattle	AA	14	14	USA	Africa	47.6	-122.4	Mark Shriver
AA_Gullah_WGA	AA_Gullah_WGA	13	13	USA	Africa	33.1	18	Mark Shriver
AA_Houston_WGA	AA_WGA	20	20	USA	Africa	29.7	-95.3	Mark Shriver
AA_New_Orleans_WGA	AA_WGA	12	12	USA	Africa	30	90	Mark Shriver
Abkhasian	Abkhasian	9	9	Abkhazia	WestEurasia	43	41	Mait Metspalu / Elza Khusnutdinova
Adygei	Adygei	25	17	Russia	WestEurasia	44	39	Coriell Cell Repositories
Afar	Afar	5	5	Ethiopia	Africa	11.8	41.4	Mark G. Thomas, Ayele Tarekegn
Ain_Touta_WGA	Ain_Touta_WGA	3	3	Algeria	Africa	35.4	5.9	Mark G. Thomas, Leila Laredj
Albanian	Albanian	6	6	Albania	WestEurasia	41.3	19.8	David Comas
Algerian	Algerian	7	7	Algeria	Africa	36.8	3	David Comas
Algonquin	Algonquin	9	9	Canada	America	48.4	-71.1	Damian Labuda
Altaian	Altaian	7	7	Russia	C.Asia/Sib	51.9	86	Mait Metspalu
Ami_Coriell	Ami	10	10	Taiwan	EastAsia	22.8	121.2	Coriell Cell Repositories
Armenian	Armenian	10	10	Armenia	WestEurasia	40.2	44.6	Mait Metspalu/ Levon Yepiskoposyan / Havhannes
Armenian_WGA	Armenian_WGA	3	3	Armenia	WestEurasia	40.2	44.5	Mark G. Thomas / Levon Yepiskoposyan
Ashkenazi_Jew	Ashkenazi_Jew	9	7	Poland	WestEurasia	52.2	21	Tel Aviv Cell Line repository
Assyrian_WGA	Assyrian_WGA	5	5	Armenia	WestEurasia	40.3	44.6	Mark G. Thomas / Levon Yepiskoposyan
Atayal_Coriell	Atayal	10	9	Taiwan	EastAsia	24.6	121.3	Coriell Cell Repositories
Australian_ECCAC	Australian	9	3	Australia	Oceania	-13	143	European Collection of Cell Cultures
Baku_WGA	Baku_WGA	3	3	Azerbaijan	WestEurasia	40.4	49.9	Mark G. Thomas, Ruslan Ruzibakiev (deceased)
Balkar	Balkar	10	10	Russia	WestEurasia	43.5	43.6	Mait Metspalu / Elza Khusnutdinova
Balochi	Balochi	24	20	Pakistan	SouthAsia	30.5	66.5	CEPH / Human Genome Diversity Project Cell Lines
BantuKenya	BantuKenya	11	6	Kenya	Africa	-3	37	CEPH / Human Genome Diversity Project Cell Lines
Bantu_SA_Herero	BantuSA	2	2	Bot./Namibia	Africa	-22	19	CEPH / Human Genome Diversity Project Cell Lines
Bantu_SA_Ovambo	BantuSA	1	1	Angola	Africa	-19	18	CEPH / Human Genome Diversity Project Cell Lines
Bantu_SA_Pedi	BantuSA	1	1	SouthAfrica	Africa	-29	30	CEPH / Human Genome Diversity Project Cell Lines
Bantu_SA_S_Sotho	BantuSA	1	1	Lesotho	Africa	-29	29	CEPH / Human Genome Diversity Project Cell Lines
Bantu_SA_Tswana	BantuSA	2	2	Bot./Namibia	Africa	-28	24	CEPH / Human Genome Diversity Project Cell Lines
Bantu_SA_Zulu	BantuSA	1	1	SouthAfrica	Africa	-28	31	CEPH / Human Genome Diversity Project Cell Lines
Basque_French	Basque	22	20	France	WestEurasia	43	0	CEPH / Human Genome Diversity Project Cell Lines
Basque_Spanish	Basque	10	9	Spain	WestEurasia	43.1	-2.1	David Comas
Bedouin2	Bedouin2	21	19	Israel(Negev)	WestEurasia	31	35	CEPH / Human Genome Diversity Project Cell Lines
Bedouin1	BedouinA	25	25	Israel(Negev)	WestEurasia	31	35	CEPH / Human Genome Diversity Project Cell Lines
Belorussian	Belorussian	10	10	Belarus	WestEurasia	53.9	28	Mait Metspalu / Alena Kushniarevich
Bengali_Bangladesh_BEB	Bengali	8	7	Bangladesh	SouthAsia	23.7	90.4	Coriell Cell Repositories
Italian_Bergamo	Bergamo	13	12	Italy(Bergamo)	WestEurasia	46	10	David Comas
BiakaPygmy	Biaka	23	20	C.Afr.Repub	Africa	4	17	CEPH / Human Genome Diversity Project Cell Lines
Bolivian_Cochabamba	Bolivian	1	1	Bolivia	America	-17.4	-66.2	Antonio Salas
Bolivian_LaPaz	Bolivian	3	3	Bolivia	America	-16.5	-68.2	Antonio Salas
Bolivian_Pando	Bolivian	3	3	Bolivia	America	-11.2	-67.2	Antonio Salas
Bougainville	Bougainville	12	10	PNG	Oceania	-6	155	CEPH / Human Genome Diversity Project Cell Lines
Brahui	Brahui	24	21	Pakistan	SouthAsia	30.5	66.5	CEPH / Human Genome Diversity Project Cell Lines
Bulgarian	Bulgarian	10	10	Bulgaria	WestEurasia	42.2	24.7	Mait Metspalu / Draga Toncheva / Mari Nelis
Burbur_WGA	Burbur_WGA	5	5	Morocco	Africa	33.5	5.1	Mark G. Thomas, Fouad Berrada
Burusho	Burusho	25	23	Pakistan	SouthAsia	36.5	74	CEPH / Human Genome Diversity Project Cell Lines
Cambodian	Cambodian	10	8	Cambodia	EastAsia	12	105	CEPH / Human Genome Diversity Project Cell Lines
Spanish_Canarias_IBS	Canary_Islanders	2	2	Spain	WestEurasia	28.1	-15.4	Coriell Cell Repositories
Chechen	Chechen	9	9	Russia	WestEurasia	43.3	45.7	Mait Metspalu / Elza Khusnutdinova
Chipewyan	Chipewyan	32	30	Canada	America	59.6	-107.3	Damian Labuda
Chukchi	Chukchi	24	20	Russia	C.Asia/Sib	69.5	168.8	DiRienzo_Sukernik
Chuvash	Chuvash	10	10	Russia	WestEurasia	56.1	47.3	Mait Metspalu / Elza Khusnutdinova
Cree	Cree	13	13	Canada	America	50.3	-102.5	Damian Labuda
Croatian	Croatian	10	10	Croatia	WestEurasia	43.5	16.5	Cristian Capelli / Igor Rudan / Tatijana Zemunik / George
Cypriot	Cypriot	8	8	Cyprus	WestEurasia	35.1	33.4	David Comas / Pierre Zalloua
Czech	Czech	10	10	Czech	WestEurasia	50.1	14.4	Coriell Cell Repositories
Dai	Dai	10	10	China	EastAsia	21	100	CEPH / Human Genome Diversity Project Cell Lines
Damara	Damara	13	12	Namibia	Africa	-19.8	16.2	Mark Stoneking / Brigitte Pakendorf
Datog	Datog	3	3	Tanzania	Africa	-3.3	35.7	Brenna Henn / Joanna Mountain
Daur	Daur	9	9	China	EastAsia	48.5	124	CEPH / Human Genome Diversity Project Cell Lines
Dinka	Dinka	9	7	Sudan	Africa	8.8	27.4	Michael Hammer

Dolgan	Dolgan	5	4	Russia	C.Asia/Sib	73	115.4	Mait Metspalu / Sardana Fedorova
Druze	Druze	42	39	Israel(Carmel)	WestEurasia	32	35	CEPH / Human Genome Diversity Project Cell Lines
Egyptian_Comas	Egyptian	14	11	Egypt	Africa	31	31.2	David Comas
Egyptian_Metspalu	Egyptian	8	7	Egypt	Africa	30.2	31.2	Mait Metspalu
English_Cornwall_GBR	English	5	5	England	WestEurasia	50.3	-4.9	Coriell Cell Repositories
English_Kent_GBR	English	5	5	England	WestEurasia	51.2	0.7	Coriell Cell Repositories
Esan_Nigeria_ESN	Esan	8	8	Nigeria	Africa	6.5	6	Coriell Cell Repositories
Eskimo_Naukan	Eskimo	20	13	Russia	C.Asia/Sib	66	169.7	Rem Sukernik
Estonian	Estonian	10	10	Estonia	WestEurasia	58.5	24.9	Mait Metspalu / Meie Pank
Ethiopian_Jew	Ethiopian_Jew	7	7	Ethiopia	Africa	9	38.7	Tel Aviv Cell Line repository
Finnish_FIN	Finnish	8	8	Finland	WestEurasia	60.2	24.9	Coriell Cell Repositories
French	French	29	25	France	WestEurasia	46	2	CEPH / Human Genome Diversity Project Cell Lines
French_South	French_South	7	7	France	WestEurasia	43.4	-0.6	David Comas
Gambian_GWD	Gambian	6	6	Gambia	Africa	13.4	16.7	Coriell Cell Repositories
Gana	Gana	9	8	Botswana	Africa	-21.7	23.4	Mark Stoneking / Brigitte Pakendorf
Georgian_Megrels	Georgian	10	10	Georgia	WestEurasia	42.5	41.9	Mait Metspalu
Georgian_Jew	Georgian_Jew	9	7	Georgia	WestEurasia	41.7	44.8	Tel Aviv Cell Line repository
Georgian_WGA	Georgian_WGA	2	2	Georgia	WestEurasia	41.7	44.8	Mark G. Thomas, Haim Ben-Ami
Greek_Comas	Greek	14	14	Greece	WestEurasia	40.6	22.9	David Comas
Greek_Coriell	Greek	8	6	Greece	WestEurasia	38	23.7	Coriell Cell Repositories
Greek_WGA	Greek_WGA	18	18	Greece	WestEurasia	37.9	23.7	Mark G. Thomas, Theologos Loukidis
Gui	Gui	11	7	Botswana	Africa	-21.5	23.3	Mark Stoneking / Brigitte Pakendorf
Gujarati1_GIH	Gujarati1	5	5	India	SouthAsia	23.2	72.7	Coriell Cell Repositories
Gujarati2_GIH	Gujarati2	5	5	India	SouthAsia	23.2	72.7	Coriell Cell Repositories
Gujarati3_GIH	Gujarati3	5	5	India	SouthAsia	23.2	72.7	Coriell Cell Repositories
Gujarati4_GIH	Gujarati4	5	5	India	SouthAsia	23.2	72.7	Coriell Cell Repositories
Hadza	Hadza	20	17	Tanzania	Africa	-3.8	35.3	Sarah Tishkoff / Dr. Thomas Nyambo
Hadza_Henn	Hadza	8	5	Tanzania	Africa	-3.6	35.1	Brenna Henn
Haiom	Haiom	9	7	Namibia	Africa	-19.4	17	Mark Stoneking / Brigitte Pakendorf
Han	Han	35	33	China	EastAsia	32.3	114	CEPH / Human Genome Diversity Project Cell Lines
Han_NChina	Han_NChina	10	10	China	EastAsia	32.3	114	CEPH / Human Genome Diversity Project Cell Lines
Hazara	Hazara	22	14	Pakistan	SouthAsia	33.5	70	CEPH / Human Genome Diversity Project Cell Lines
Hezhen	Hezhen	9	8	China	EastAsia	47.5	133.5	CEPH / Human Genome Diversity Project Cell Lines
Himba	Himba	5	4	Namibia	Africa	-19.1	14.1	Mark Stoneking / Brigitte Pakendorf
Hoan	Hoan	7	7	Botswana	Africa	-24	23.4	Mark Stoneking / Brigitte Pakendorf
Hungarian_Coriell	Hungarian	10	10	Hungary	WestEurasia	47.5	19.1	Coriell Cell Repositories
Hungarian_Metspalu	Hungarian	10	10	Hungary	WestEurasia	47.5	19.1	Mait Metspalu / Bela Melegh / Judit Bene
Icelandic	Icelandic	12	12	Iceland	WestEurasia	64.1	-21.9	Coriell Cell Repositories
Iranian	Iranian	9	8	Iran	WestEurasia	35.6	51.5	Mait Metspalu / EBC
Iranian_Jew	Iranian_Jew	10	9	Iran	WestEurasia	35.7	51.4	Tel Aviv Cell Line repository
Iraqi_Jew	Iraqi_Jew	9	6	Iraq	WestEurasia	33.3	44.4	Tel Aviv Cell Line repository
Japanese	Japanese	29	29	Japan	EastAsia	38	138	CEPH / Human Genome Diversity Project Cell Lines
Jordanian	Jordanian	10	9	Jordan	WestEurasia	32.1	35.9	Mait Metspalu / EBC
Ju_hoan_North	Ju_hoan_North	24	22	Namibia	Africa	-18.9	21.5	CEPH / Human Genome Diversity Project Cell Lines
Ju_hoan_South	Ju_hoan_South	9	6	Botswana	Africa	-21.2	20.7	Mark Stoneking / Brigitte Pakendorf
Kalash	Kalash	19	18	Pakistan	SouthAsia	36	71.5	CEPH / Human Genome Diversity Project Cell Lines
Kalmyk	Kalmyk	10	10	Russia	C.Asia/Sib	46.2	45.3	Mait Metspalu / Elza Khusnutdinova
Karitiana	Karitiana	14	12	Brazil	America	-10	-63	CEPH / Human Genome Diversity Project Cell Lines
Kgalagadi	Kgalagadi	5	5	Botswana	Africa	-24.8	21.8	Mark Stoneking / Brigitte Pakendorf
Kharia	Kharia	15	12	India	SouthAsia	25.8	82.7	K. Thangaraj / Lalji Singh
Khomani	Khomani	12	11	SouthAfrica	Africa	-27.8	21.1	Brenna Henn
Khwe	Khwe	10	8	Botswana	Africa	-18.4	21.9	Mark Stoneking / Brigitte Pakendorf
Kikuyu	Kikuyu	4	4	Kenya	Africa	-0.4	36.9	Geoge Ayodo
Kinh_Vietnam_KHV	Kinh	8	8	Vietnam	EastAsia	21	105.9	Coriell Cell Repositories
Korean	Korean	6	6	Korea	EastAsia	37.6	127	Coriell Cell Repositories
Kuchin_Jew	Kuchin_Jew	5	5	India	SouthAsia	10	76.3	Tel Aviv Cell Line repository
Kumyk	Kumyk	9	8	Russia	C.Asia/Sib	43.3	46.6	Mait Metspalu / Elza Khusnutdinova
Kurd_WGA	Kurd_WGA	2	2	Armenia	WestEurasia	40.7	44.4	Mark G. Thomas / Levon Yepiskoposyan
Kusunda	Kusunda	10	10	Nepal	SouthAsia	28.1	82.5	CEPH / Human Genome Diversity Project Cell Lines
Kyrgyz	Kyrgyz	10	9	Kyrgyzystan	C.Asia/Sib	42.9	74.6	Robert Mahley and Ugur Hodoglugil
Lahu	Lahu	8	8	China	EastAsia	22	100	CEPH / Human Genome Diversity Project Cell Lines
Lebanese	Lebanese	8	8	Lebanon	WestEurasia	33.8	35.6	Mait Metspalu / EBC
Lezgin	Lezgin	10	9	Russia	WestEurasia	42.1	48.2	Mait Metspalu / Elza Khusnutdinova
Libyan_Jew	Libyan_Jew	9	9	Libya	Africa	32.9	13.2	Tel Aviv Cell Line repository
Lithuanian	Lithuanian	10	10	Lithuania	WestEurasia	54.9	23.9	Mait Metspalu / Vaidutis Kucinskas / Mari Nelis
Lodhi	Lodhi	14	13	India	SouthAsia	25.5	78.6	K. Thangaraj / Lalji Singh
Luhya_Kenya_LWK	Luhya	8	8	Kenya	Africa	1.3	36.8	Coriell Cell Repositories
Luo	Luo	9	8	Kenya	Africa	-0.1	34.3	Geoge Ayodo
Makrani	Makrani	25	20	Pakistan	SouthAsia	26	64	CEPH / Human Genome Diversity Project Cell Lines
Mala	Mala	15	13	India	SouthAsia	18.7	78.2	K. Thangaraj / Lalji Singh
Maltese	Maltese	8	8	Malta	WestEurasia	35.9	14.4	David Comas / Pierre Zalloua
Mandenka	Mandenka	22	17	Senegal	Africa	12	-12	CEPH / Human Genome Diversity Project Cell Lines
Masai_Ayodo	Masai	3	2	Kenya	Africa	-1.1	35.9	David Reich / George Ayodo
Masai_Kinyawa_MKK	Masai	10	10	Kenya	Africa	-1.5	35.2	Coriell Cell Repositories
Mayan	Mayan	21	18	Mexico	America	19	-91	CEPH / Human Genome Diversity Project Cell Lines
MbutiPygmy	Mbuti	14	10	Congo	Africa	1	29	CEPH / Human Genome Diversity Project Cell Lines
Mende_MSL	Mende	8	8	SierraLeone	Africa	8.5	-13.2	Coriell Cell Repositories
Miao	Miao	10	10	China	EastAsia	28	109	CEPH / Human Genome Diversity Project Cell Lines
Mixe	Mixe	10	10	Mexico	America	17	96.6	William Klitz / Cheryl Winkler
Mixtec	Mixtec	10	10	Mexico	America	16.7	-97.2	William Klitz / Cheryl Winkler
Mongola	Mongola	11	6	China	C.Asia/Sib	45	111	CEPH / Human Genome Diversity Project Cell Lines
Mordovian	Mordovian	10	10	Russia	WestEurasia	54.2	45.2	Mait Metspalu / Elza Khusnutdinova
Moroccan_Jew	Moroccan_Jew	7	6	Morocco	Africa	34	-6.8	Tel Aviv Cell Line repository
Mozabite	Mozabite	27	21	Algeria	Africa	32	3	CEPH / Human Genome Diversity Project Cell Lines
Nama	Nama	18	16	Namibia	Africa	-24.3	17.3	Mark Stoneking / Brigitte Pakendorf
Naro	Naro	10	8	Botswana	Africa	-22	21.6	Mark Stoneking / Brigitte Pakendorf
Naxi	Naxi	9	9	China	EastAsia	26	100	CEPH / Human Genome Diversity Project Cell Lines
Nganasan	Nganasan	14	11	Russia	C.Asia/Sib	71.1	96.1	DiRienzo_Sukernik
Nogai	Nogai	9	9	Russia	C.Asia/Sib	44.4	41.9	Mait Metspalu / Elza Khusnutdinova
North_Ossetian	North_Ossetian	10	10	Russia	WestEurasia	43	44.7	Mait Metspalu / Elza Khusnutdinova
Norwegian	Norwegian	11	11	Norway	WestEurasia	60.4	5.4	Cristian Capelli
Ojibwa	Ojibwa	28	19	Canada	America	46.5	-81	Damian Labuda / David E.C. Cole
Onge	Onge	17	11	India	SouthAsia	10.8	92.5	K. Thangaraj / Lalji Singh
Orcadian	Orcadian	13	13	OrkneyIslands	WestEurasia	59	-3	CEPH / Human Genome Diversity Project Cell Lines
Oromo	Oromo	7	6	Ethiopia	Africa	9	36.5	Anna DiRienzo / Beall / Gebremedhin
Oroqen	Oroqen	9	9	China	EastAsia	50.4	126.5	CEPH / Human Genome Diversity Project Cell Lines

Spanish_Pais_Vasco_IBS	Pais_Vasco	5	5	Spain	WestEurasia	42.8	-2.7	Coriell Cell Repositories
Palestinian	Palestinian	45	38	Israel(Central)	WestEurasia	32	35	CEPH / Human Genome Diversity Project Cell Lines
Papuan	Papuan	18	14	PNG	Oceania	-4	143	CEPH / Human Genome Diversity Project Cell Lines
Pathan	Pathan	24	19	Pakistan	SouthAsia	33.5	70.5	CEPH / Human Genome Diversity Project Cell Lines
Piapoco	Piapoco	5	4	Colombia	America	3	-68	CEPH / Human Genome Diversity Project Cell Lines
Pima	Pima	14	14	Mexico	America	29	-108	CEPH / Human Genome Diversity Project Cell Lines
Punjabi_Lahore_PJL	Punjabi	8	8	Pakistan	SouthAsia	31.5	74.3	Coriell Cell Repositories
Quechua_Coriell	Quechua	5	5	Peru	America	-13.5	-72	Coriell Cell Repositories
Russian	Russian	23	22	Russia	WestEurasia	61	40	CEPH / Human Genome Diversity Project Cell Lines
Saami	Saami	1	1	n/a	WestEurasia	n/a	n/a	Svante Paabo
Saharawi	Saharawi	7	6	Algeria	Africa	24.2	-12.9	David Comas
Sandawe	Sandawe	28	22	Tanzania	Africa	-5.5	35.5	Sarah Tishkoff / Dr. Thomas Nyambo
Sardinian	Sardinian	29	27	Italy(Sardinia)	WestEurasia	40	9	CEPH / Human Genome Diversity Project Cell Lines
Saudi	Saudi	10	8	Saudi_Arabia	WestEurasia	18.5	42.5	Mait Metspalu / EBC
Scottish_Argyll_Bute_GBR	Scottish	4	4	England	WestEurasia	56	-3.9	Coriell Cell Repositories
Selkup	Selkup	10	10	Russia	C.Asia/Sib	65.5	82.3	Mait Metspalu / Ludmila Osipova
Shaigi_WGA	Shaigi_WGA	3	3	Sudan	Africa	15.6	32.5	Mark G. Thomas, Hiba MA Babiker
She	She	10	10	China	EastAsia	27	119	CEPH / Human Genome Diversity Project Cell Lines
Shua	Shua	10	9	Botswana	Africa	-20.6	25.3	Mark Stoneking / Brigitte Pakendorf
Italian_EastSicilian	Sicilian	5	5	Italy	WestEurasia	37.1	15.3	Cristian Capelli
Italian_WestSicilian	Sicilian	6	6	Italy	WestEurasia	38	12.5	Cristian Capelli
Sindhi	Sindhi	24	18	Pakistan	SouthAsia	25.5	69	CEPH / Human Genome Diversity Project Cell Lines
Somali	Somali	13	13	Kenya	Africa	5.6	48.3	Geoge Ayodo
Spanish_ Leon_IBS	Spanish	5	5	Spain	WestEurasia	41.4	-4.5	Coriell Cell Repositories
Spanish_Andalucia_IBS	Spanish	4	4	Spain	WestEurasia	37.4	-6	Coriell Cell Repositories
Spanish_Aragon_IBS	Spanish	6	6	Spain	WestEurasia	41	-1	Coriell Cell Repositories
Spanish_Baleares_IBS	Spanish	4	4	Spain	WestEurasia	39.5	3	Coriell Cell Repositories
Spanish_Cantabria_IBS	Spanish	5	5	Spain	WestEurasia	43.3	-4	Coriell Cell Repositories
Spanish_Cataluna_IBS	Spanish	5	5	Spain	WestEurasia	41.8	1.5	Coriell Cell Repositories
Spanish_Extremadura_IBS	Spanish	5	5	Spain	WestEurasia	39	-6	Coriell Cell Repositories
Spanish_Galicia_IBS	Spanish	5	5	Spain	WestEurasia	42.5	-8.1	Coriell Cell Repositories
Spanish_Mancha_IBS	Spanish	5	5	Spain	WestEurasia	39.9	-4	Coriell Cell Repositories
Spanish_Murcia_IBS	Spanish	5	4	Spain	WestEurasia	38	-1.1	Coriell Cell Repositories
Spanish_Valencia_IBS	Spanish	5	5	Spain	WestEurasia	39.5	-0.4	Coriell Cell Repositories
Surui	Surui	8	8	Brazil	America	-11	-62	CEPH / Human Genome Diversity Project Cell Lines
Syrian	Syrian	8	8	Syrian	WestEurasia	35.1	36.9	Mait Metspalu / EBC
Taa_East	Taa_East	8	7	Botswana	Africa	-24.2	22.8	Mark Stoneking / Brigitte Pakendorf
Taa_North	Taa_North	11	9	Botswana	Africa	-23	22.4	Mark Stoneking / Brigitte Pakendorf
Taa_West	Taa_West	17	16	Botswana	Africa	-23.6	20.3	Mark Stoneking / Brigitte Pakendorf
Tajik	Tajik	8	8	Tadjikistan	C.Asia/Sib	37.4	71.6	Mait Metspalu / Oleg Balanovsky
Thai	Thai	10	10	Thailand	EastAsia	13.8	100.5	European Collection of Cell Cultures
Tiwari	Tiwari	15	15	India	SouthAsia	21.9	83.4	K. Thangaraj / Lalji Singh
Tshwa	Tshwa	9	5	Botswana	Africa	-21	25.9	Mark Stoneking / Brigitte Pakendorf
Tswana	Tswana	5	5	Botswana	Africa	-24.1	25.4	Mark Stoneking / Brigitte Pakendorf
Tu	Tu	10	10	China	EastAsia	36	101	CEPH / Human Genome Diversity Project Cell Lines
Tujia	Tujia	10	10	China	EastAsia	29	109	CEPH / Human Genome Diversity Project Cell Lines
Tunisian	Tunisian	8	8	Tunisia	Africa	36.8	10.2	David Comas
Tunisian_Jew	Tunisian_Jew	7	7	Tunisia	Africa	36.8	10.2	Tel Aviv Cell Line repository
Turkish	Turkish	4	4	Turkey	WestEurasia	39.6	28.5	David Comas / Pierre Zalloua
Turkish_Adana	Turkish	10	10	Turkey	WestEurasia	37	35.3	David Comas / Pierre Zalloua
Turkish_Aydin	Turkish	10	7	Turkey	WestEurasia	37.9	27.8	David Comas / Pierre Zalloua
Turkish_Balikesir	Turkish	10	6	Turkey	WestEurasia	39.4	27.5	David Comas / Pierre Zalloua
Turkish_Istanbul	Turkish	10	10	Turkey	WestEurasia	41	29	David Comas / Pierre Zalloua
Turkish_Kayseri	Turkish	10	10	Turkey	WestEurasia	38.7	35.5	David Comas / Pierre Zalloua
Turkish_Trabzon	Turkish	10	9	Turkey	WestEurasia	41	39.7	David Comas / Pierre Zalloua
Turkish_Jew	Turkish_Jew	9	8	Turkey	WestEurasia	41	29	Tel Aviv Cell Line repository
Turkmen	Turkmen	7	7	Uzbekistan	C.Asia/Sib	42.5	59.6	Mait Metspalu / Oleg Balanovsky
Italian_Tuscan	Tuscan	8	8	Italy(Tuscany)	WestEurasia	43	11	CEPH / Human Genome Diversity Project Cell Lines
Tuvinian	Tuvinian	10	10	Russia	C.Asia/Sib	50.3	95.2	Mait Metspalu / Larissa Damba / Mikhail Voevoda
Ukrainian	Ukrainian	9	9	Ukraine	WestEurasia	50.3	31.6	Mait Metspalu / Oleg Balanovsky
Uygur	Uygur	10	10	China	EastAsia	44	81	CEPH / Human Genome Diversity Project Cell Lines
Uzbek	Uzbek	10	10	Uzbekistan	C.Asia/Sib	41.3	69.2	Mait Metspalu / Elza Khusnutdinova
Uzbek_WGA	Uzbek_WGA	1	1	Uzbekistan	C.Asia/Sib	41.3	69.3	Mark G. Thomas, Ruslan Ruzibakiev (deceased)
Vishwabrahmin	Vishwabrahmin	15	13	India	SouthAsia	16.3	80.5	K. Thangaraj / Lalji Singh
Wambo	Wambo	5	5	Namibia	Africa	-17.7	18.1	Mark Stoneking / Brigitte Pakendorf
Xibo	Xibo	9	7	China	EastAsia	43.5	81.5	CEPH / Human Genome Diversity Project Cell Lines
Xuun	Xuun	15	13	Namibia	Africa	-18.7	19.7	Mark Stoneking / Brigitte Pakendorf
Yakut	Yakut	25	20	Russia	C.Asia/Sib	63	129.5	CEPH / Human Genome Diversity Project Cell Lines
Yemen	Yemen	7	6	Yemen	WestEurasia	14	44.6	Mait Metspalu, Richard Villems, Ene Metspalu
Yemenite_Jew	Yemenite_Jew	8	8	Yemen	WestEurasia	15.4	44.2	Tel Aviv Cell Line repository
Yi	Yi	10	10	China	EastAsia	28	103	CEPH / Human Genome Diversity Project Cell Lines
Yoruba	Yoruba	108	70	Nigeria	Africa	7.4	3.9	Coriell Cell Repositories
Zapotec	Zapotec	10	10	Mexico	America	17	-96.5	William Klitz / Cheryl Winkler

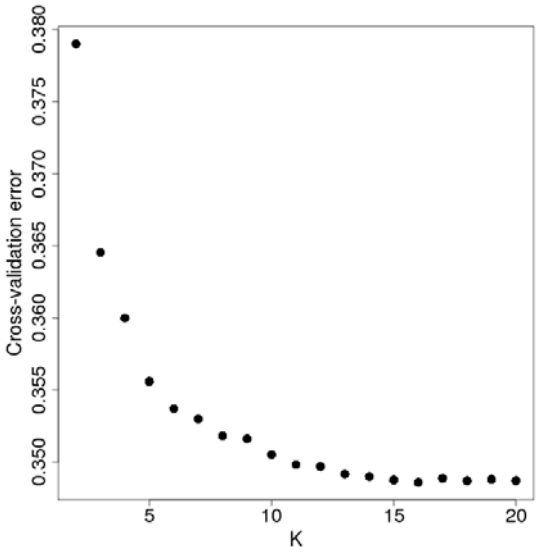
ADMIXTURE analysis

We carried out model-based clustering analysis using ADMIXTURE [7] 1.23 on the genome-wide dataset, combining the present-day humans with Loschbour, Stuttgart, Motala12 and Motala_merge.

ADMIXTURE is a commonly used method for investigating admixture proportions in human populations, although its interpretation in terms of history is not straightforward. In the context of the present paper we use it only to (i) identify a set of West Eurasian populations for further analysis, and (ii) to identify a set non-West Eurasian populations from the rest of the world to be used as references for our methods of ancestry estimation. This analysis also serves as an exploration of populations included in the Affymetrix Human Origins Array dataset made available with this paper.

We used PLINK[10] 1.07 to thin the original dataset of 594,924 autosomal SNPS to remove SNPs in strong linkage disequlibrium, employing a window of 200 SNPs advanced by 25 SNPs and an r^2 threshold of 0.4 (--indep-pairwise 200 25 0.4). A total of 293,832 SNPs remained for analysis after this procedure. We ran ADMIXTURE with 10-fold cross-validation (--cv=10), varying the number of ancestral populations between K=2 and K=20 (Figure S9.1).

Figure S9.1: Cross-validation error of ADMIXTURE analysis. *We observe a plateau as K increases, with the minimum (0.34858) attained at K=16 within the range of K=2 to 20.*



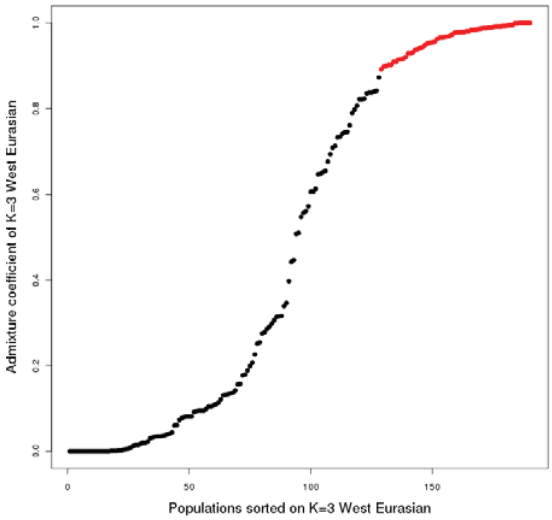
The results of the ADMIXTURE analysis can be found in Extended Data Figure 3.

We have the following observations:

K=2 separates African from non-African populations.

K=3 reveals a West Eurasian ancestry component. We used this to identify a set of 58 populations from West Eurasia (Figure 1, Table 1) to be used in our paper, applying a >88% membership threshold for inclusion in this set (Figure S9.2).

Figure S9.2: A subset of populations (red) was selected for analyses involving West Eurasia.



The ancient samples appear to be mostly West Eurasian in their ancestry, although the hunter-gatherers are also inferred to have greater or lesser extents of the eastern non-African (ANE) component which is lacking in Stuttgart. This is consistent with the positive f_4*(ENA, Chimp; Hunter Gatherer, Stuttgart)* statistic reported in SI12, which we interpret there as showing that ENA populations are closer to European Hunter-Gatherers than to Stuttgart.

K=4 breaks the African component into an African hunter-gatherer ancestry maximized in Bushmen such as the Ju_hoan_North and an African farmer component maximized in the Yoruba.

K=5 breaks the ENA component down into one maximized in the Karitiana from the Americas and one maximized in the Ami from Taiwan. This analysis further suggests that the ENA affinity of Hunter-Gatherers is related to the Karitiana component.

K=6 reveals a south Eurasian component maximized in Papuans, which is also represented in South Asians. MA1 shows some affinity to this component, in contrast to more recent Eurasian hunter-gatherers who continue to mainly show ties to Native Americans.

K=7 reveals a component maximized in the East African Hadza.

K=8 reveals a South Asian component maximized in the Mala. This is separated from the earlier south Eurasian component, and MA1 shows some affinity to this new component, rather than to the Papuan-maximized Oceanian component that also results from this split.

K=9 shows the first appearance of a Mbuti-maximized African Pygmy-related component that reappears at K=13, K=15, and persists thereafter.

K=10 reveals a split within West Eurasia, with one component maximized in Loschbour and one maximized in BedouinB. From the ancient samples, only Stuttgart shows mixed membership in these two components, consistent with the hypothesis that Early European Farmers represented a mixture of West Eurasian Hunter Gatherers and Near Eastern farmers. Membership in the Near Eastern component is prevalent in Europe, consistent with the hypothesis that Europeans have inherited some Near Eastern ancestry via early farmers. At K=10, an Onge-maximized component also appears; this re-appears at K=12 and persists thereafter.

K=11 reveals a Siberian component, which is maximized in the Nganasan. Siberians previously showed mixed membership in the Native American and East Asian components, consistent with the idea that present-day Siberians have been formed by admixture with East Asians. The near absence of the Siberian component in ancient hunter-gatherers contrasts with its presence at low levels in present-day Northeastern Europeans, consistent with more recent Siberian influences in that part of Europe. A Chipewyan-maximized North American component also appears; this re-appears at K=13 and persists thereafter.

K=12 reveals a Somali-maximized East African component; this re-appears at K=14 and persists thereafter.

K=13 shows the reappearance of the African Pygmy-related and North American-related components from lower K.

K=14 shows the appearance of a component that is maximized in the Kalash and that is widely distributed in South Asia, the Caucasus, the Near East, and in diminishing strength in Europe. It is absent in Sardinians, Basques, and all ancient Europeans, although it is present in MA1. This component also does not appear in North and East Africa where other West Eurasian admixture is observed. This is consistent with MA1 having contributed some ancestry to present-day Europeans not accounted for by West Eurasian Hunter Gatherers and Early European Farmers. The presence of this component in the Near East contrasts with its absence in Stuttgart, consistent with the widely

shared negative f_3*(Near East; Stuttgart, MA1)* statistics (Table 1) indicating that present-day Near Easterners have been affected by gene flow not present in early Near Eastern migrants into Europe.

K=15 shows the re-appearance of the Mbuti-maximized component which persists thereafter.

K=16 is the value which minimizes the cross-validation error. It reveals a Pima-maximized Central American component. In the tail of the distribution (Figure S9.1) the cross-validation error plateaus, and we report the results of our analysis up to K=20 (Extended Data Figure 3), showing the appearance of several additional geographically circumscribed components.

We wish to avoid over-interpretation of the admixture proportions, but nonetheless highlight some patterns each of which is validated by *f*-statistic analyses reported in this study and previous studies:

1. The absence of a Near Eastern relatedness in all European hunter-gatherer groups but its presence in Stuttgart.

2. The clear affinity of MA1 to Native American populations but not to East Asian or present-day Siberian populations.

3. The occurrence of low levels of additional gene flows in west Eurasia from Africa (in parts of the Near East or southern Europe) or recent Siberia (in parts of Northeastern Europe or the Near East and Caucasus).

4. Evidence tying MA1 to Europe, the northern Near East and Caucasus, and south/central Asia.

We identified 31 populations maximizing ancestral components across all 20 runs, breaking ties using populations with the highest sample size. These populations represent a set of groups encompassing different aspects of modern human variation, which we find to be useful in analyses involving the relationships of West Eurasians to other present-day populations. This list is:

Ami, Atayal, Basque, BedouinB, Bougainville, Brahui, Chipewyan, Dinka, Esan, Georgian, Gujarati4, Hadza, Han, Ju_hoan_North, Kalash, Karitiana, Loschbour, Mala, Masai, Mbuti, Mozabite, Naxi, Nganasan, Onge, Papuan, Pima, She, Somali, Stuttgart, Vishwabrahmin, Yoruba

We also identified a list of 13 of these populations which show no evidence of either European or Near Eastern ancestry at K=10 (which is the lowest K in which Europe/Near Eastern-centric components emerge). This set of outgroup populations is as follows:

Ami, Atayal, Bougainville, Esan, Han, Ju_hoan_North, Karitiana, Mbuti, Naxi, Onge, Papuan, She, Yoruba

The Mala also show evidence of West Eurasian-related ancestry at K=10, but do so at earlier K and appear to have West Eurasian-related “Ancestral North Indian” ancestry within the last few thousand years[11], so we exclude them from the set of outgroup populations.

References

1 Alon Keinan, James C. Mullikin, Nick Patterson, and David Reich, 'Measurement of the Human Allele Frequency Spectrum Demonstrates Greater Genetic Drift in East Asians Than in Europeans', *Nat Genet,* 39 (2007), 1251-55.

2 N. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, and D. Reich, 'Ancient Admixture in Human History', *Genetics,* 192 (2012), 1065-93.

3 Joseph K. Pickrell, Nick Patterson, Po-Ru Loh, Mark Lipson, Bonnie Berger, Mark Stoneking, Brigitte Pakendorf, and David Reich, 'Ancient West Eurasian Ancestry in Southern and Eastern Africa', *arXiv:1307.8014v1* (2013).
4 'An Integrated Map of Genetic Variation from 1,092 Human Genomes', *Nature,* 491 (2012), 56-65.
5 Kay Prufer, Fernando Racimo, Nick Patterson, Flora Jay, Sriram Sankararaman, Susanna Sawyer, Anja Heinze, Gabriel Renaud, Peter H. Sudmant, Cesare de Filippo, Heng Li, Swapan Mallick, Michael Dannemann, Qiaomei Fu, Martin Kircher, Martin Kuhlwilm, Michael Lachmann, Matthias Meyer, Matthias Ongyerth, Michael Siebauer, Christoph Theunert, Arti Tandon, Priya Moorjani, Joseph Pickrell, James C. Mullikin, Samuel H. Vohr, Richard E. Green, Ines Hellmann, Philip L. F. Johnson, Helene Blanche, Howard Cann, Jacob O. Kitzman, Jay Shendure, Evan E. Eichler, Ed S. Lein, Trygve E. Bakken, Liubov V. Golovanova, Vladimir B. Doronichev, Michael V. Shunkov, Anatoli P. Derevianko, Bence Viola, Montgomery Slatkin, David Reich, Janet Kelso, and Svante Paabo, 'The Complete Genome Sequence of a Neanderthal from the Altai Mountains', *Nature,* advance online publication (2013).
6 Kay Prüfer et al., 'The Genome Sequence of a Neandertal from the Altai Mountains', *Nature (in press)*.
7 D. H. Alexander, J. Novembre, and K. Lange, 'Fast Model-Based Estimation of Ancestry in Unrelated Individuals', *Genome Res,* 19 (2009), 1655-64.
8 N. Patterson, A. L. Price, and D. Reich, 'Population Structure and Eigenanalysis', *PLoS Genet,* 2 (2006), e190.
9 Alkes L. Price, Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, and David Reich, 'Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies', *Nat Genet,* 38 (2006), 904-09.
10 S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham, 'Plink: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses', *Am J Hum Genet,* 81 (2007), 559-75.
11 Priya Moorjani, Kumarasamy Thangaraj, Nick Patterson, Mark Lipson, Po-Ru Loh, Periyasamy Govindaraj, Bonnie Berger, David Reich, and Lalji Singh, 'Genetic Evidence for Recent Population Mixture in India', *American journal of human genetics* (2013).

Supplementary Information 10

Admixture proportions for Stuttgart

Iosif Lazaridis*, Nick Patterson and David Reich

* To whom correspondence should be addressed (lazaridis@genetics.med.harvard.edu)

A few lines of evidence suggest that the Stuttgart female harbors ancestry not only from Near Eastern farmers but also from pre-Neolithic European hunter-gatherers:

1. Her position in Fig. 1B, intermediate between the Near East and European hunter-gatherers.

2. The fact that the statistic f_4*(Stuttgart, Near East; Loschbour, Chimp)* is positive (Table S10.1).

3. The results of ADMIXTURE analysis (SI 9) which show that when the West Eurasian ancestral population is split into European/Near Eastern sub-populations centered in Loschbour and southern Near Easterners respectively (K=10), Stuttgart shows mixed ancestry from both.

Table S10.1: Loschbour shares more genetic drift with Stuttgart than with Near Easterners. *This pattern is consistent with European hunter-gatherer admixture in Stuttgart.*

Population *X*	***f*$_4$*(Stuttgart, X; Loschbour, Chimp)***	**Z**
Kumyk	0.00153	3.094
Turkish_Jew	0.00169	3.563
Turkish	0.00179	3.837
Cypriot	0.00191	3.904
Abkhasian	0.00199	4.151
Georgian	0.00200	4.155
Moroccan_Jew	0.00214	4.309
Georgian_Jew	0.00216	4.284
Armenian	0.00218	4.490
Tunisian_Jew	0.00257	5.169
Iranian_Jew	0.00276	5.672
Druze	0.00277	5.924
Libyan_Jew	0.00297	6.214
Iraqi_Jew	0.00305	6.066
Iranian	0.00311	6.290
Lebanese	0.00377	7.741
Saudi	0.00423	8.575
Syrian	0.00437	8.618
Yemenite_Jew	0.00458	9.100
BedouinB	0.00464	9.331
Palestinian	0.00474	10.183
Jordanian	0.00480	9.603
BedouinA	0.00618	12.951

Note: Only significant Z>3 statistics with X being any West Eurasian are shown (the complete set of these statistics for all West Eurasian populations is given in Extended Data Table 1).

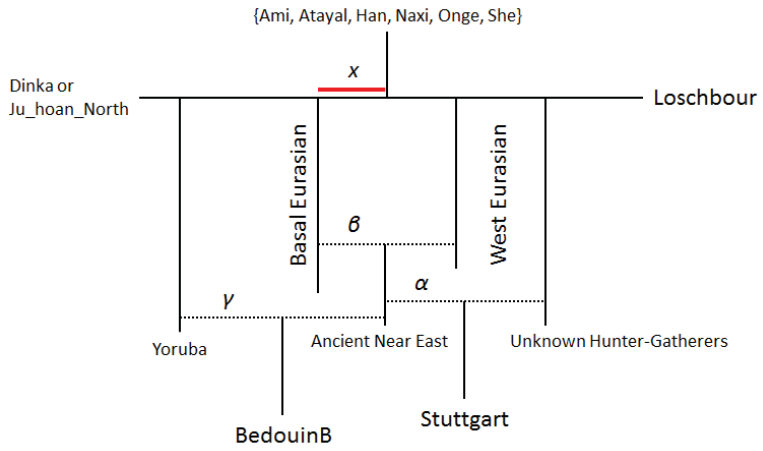
The existence of such admixture is also plausible archaeologically, as the Linearbandkeramik postdates the earliest Neolithic of southeastern Europe, and there may have been opportunity for Near Eastern Neolithic farmers to acquire a portion of European hunter-gatherer ancestry prior to the

establishment of the central European Neolithic, either en route to central Europe (e.g., in the Balkans) or by mixing with the indigenous central European hunter-gatherers they encountered.

A challenge in estimating mixture proportions for Stuttgart is that the two constituent elements contributing to it may not be represented in our data. The present-day Near East has plausibly been affected by events postdating the migration of Neolithic migrants into Europe, showing negative $f_3(Near\ East;\ Stuttgart,\ X)$ where X is MA1, Native American, South Asian, or African (Table 1, Extended Data Table 1). As a result, it is risky to treat any individual Near Eastern population as an unmixed descendant of early Near Eastern farmers. Similarly, the ancient European hunter gatherer samples that we have sequenced (Loschbour and Motala12) are very informative, but it is not clear how they relate to the pre-Neolithic inhabitants of the Balkans and central Europe.

Recognizing the challenge posed by the lack of accurate surrogates for the ancestral populations, we hypothesized that Stuttgart is a mixture of an unknown hunter-gatherer population that forms a clade with Loschbour and an unknown Near Eastern population (NE) in proportions 1-α and α. We do not know the exact NE population contributing ancestry to Stuttgart. However, we explored using BedouinB as a surrogate, as this is the population that appears at the southern end of the Near Eastern cline (Fig. 1B) and appears to have no Asian ancestry (SI 9). A complication of using the BedouinB population is that it has some African admixture, as indicated by the ADMIXTURE analysis (SI 9). We estimated a lower bound (4.2 ± 0.3%) on this admixture proportion using ALDER[1] using the Yoruba as a reference population. The advantage of this linkage-disequilibrium based method is that, unlike f_4-ratio estimation[2] no explicit model of population relationships is needed. We can also use the 5.1% estimate from ADMIXTURE K=3, or 7.2% from ADMIXTURE K=4 (SI 9). The two estimates differ because the Yoruba are inferred to have low levels of West Eurasian admixture at K=3, but to belong 100% to their own ancestral component at K=4.

Figure S10.1: f_4-ratio estimation of Near Eastern admixture in Stuttgart



Consider Fig. S10.1 in which we show Stuttgart as a mixture of an unknown hunter-gatherer (UHG) population and (NE) in proportions (1-α, α). From our modeling note (SI 11), we infer that NE is plausibly a mixture of a West Eurasian element plus a basal Eurasian one, so let 1-β, β be the mixture proportions of these two elements. We also assume the phylogenetic position of eastern non-African population *X*, alternatively using Ami, Atayal, Han, Naxi, Onge, She, from the set of 13 populations identified in SI 9. (We cannot use Karitiana because of its ANE ancestry, or Oceanians because of their Denisovan ancestry which does not conform to the model of Fig. S10.1.)

We can then write:

$$f_4(Outgroup,\ X;\ Loschbour,\ Stuttgart) = -\alpha\beta x \qquad \text{(S10.1)}$$

where x is the drift shared by most Eurasians but not basal Eurasians. We can also write:

$$f_4(Outgroup, X; Loschbour, NE) = -\beta x \quad \text{(S10.2)}$$

The ratio of the two yields the Near Eastern admixture of Stuttgart, α. While $f_4(Outgroup, X; Loschbour, NE)$ is unknown, we can estimate it via ancestry subtraction as follows:

$$f_4(\text{Outgroup, X; Loschbour, BedouinB}) = \quad \text{(S10.3)}$$
$$= \gamma f_4(\text{Outgroup, X; Loschbour, Yoruba}) + (1-\gamma) f_4(\text{Outgroup, X; Loschbour, NE})$$

or, equivalently:

$$f_4(\text{Outgroup, X; Loschbour, NE}) = \quad \text{(S10.4)}$$
$$= [f_4(\text{Outgroup, X; Loschbour, BedouinB}) - \gamma f_4(\text{Outgroup, X; Loschbour, Yoruba})]/(1-\gamma)$$

We choose Yoruba as a source of the African admixture in Stuttgart, as the source of the admixture in BedouinB appears to be African-farmer related (K=4, SI 9), and Yoruba are the population of African farmers with the highest sample size in the Human Origins dataset.

Shared common drift between "Outgroup" and Yoruba in the above equation complicates analysis, so we choose the "Outgroup" to be Dinka and Ju_hoan_North, two populations that do not appear to have recent common ancestry with West Africans.

We estimate γ=4.2%, or 5.1%, or 7.2%, as mentioned previously; these differ by only a few percent, but because they are used to subtract a portion of African ancestry from the BedouinB that is quite divergent from Eurasians, these small differences have substantial effects.

The amount of Near Eastern admixture estimated for Stuttgart can be seen in Table S10.2 and range between 61-98% with estimates increasing as the amount of estimated African admixture in BedouinB increases. Estimates using Dinka or Ju_hoan_North as an African outgroup are similar. There are reasons to doubt both the lower estimates (near 60%), since ALDER provides only a lower bound on African ancestry, but also the higher estimates (near 100%) since there is direct evidence that Stuttgart has European hunter-gatherer ancestry (Fig. 1B and Table S10.1). Determining the precise levels of Near Eastern admixture in Stuttgart must await further ancient DNA studies from both Europe and the Near East, but we can at least reasonably claim that most of the sample's ancestry was Near Eastern, consistent with the mtDNA evidence for the Linearbandkeramik, which demonstrated a strong Near Eastern influence[3-5].

Table S10.2: Near Eastern admixture estimates for Stuttgart

	Outgroup=Dinka			**Outgroup=Ju_hoan_North**		
African ancestry assumed in BedouinB	**4.20%**	**5.10%**	**7.20%**	**4.20%**	**5.10%**	**7.20%**
Ami	0.667	0.727	0.927	0.662	0.729	0.965
Atayal	0.625	0.686	0.899	0.617	0.685	0.938
Han	0.632	0.689	0.880	0.625	0.689	0.912
Naxi	0.616	0.670	0.853	0.608	0.669	0.883
Onge	0.665	0.717	0.885	0.660	0.718	0.914
She	0.684	0.744	0.945	0.680	0.748	0.984

References

1 Po-Ru Loh, Mark Lipson, Nick Patterson, Priya Moorjani, Joseph K. Pickrell, David Reich, and Bonnie Berger, 'Inferring Admixture Histories of Human Populations Using Linkage Disequilibrium', *Genetics,* 193 (2013), 1233-54.

2 N. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, and D. Reich, 'Ancient Admixture in Human History', *Genetics,* 192 (2012), 1065-93.

3 Wolfgang Haak, Peter Forster, Barbara Bramanti, Shuichi Matsumura, Guido Brandt, Marc Tänzer, Richard Villems, Colin Renfrew, Detlef Gronenborn, Kurt Werner Alt, and Joachim Burger, 'Ancient DNA from the First European Farmers in 7500-Year-Old Neolithic Sites', *Science,* 310 (2005), 1016-18.

4 Wolfgang Haak, Oleg Balanovsky, Juan J. Sanchez, Sergey Koshel, Valery Zaporozhchenko, Christina J. Adler, Clio S. I. Der Sarkissian, Guido Brandt, Carolin Schwarz, Nicole Nicklisch, Veit Dresely, Barbara Fritsch, Elena Balanovska, Richard Villems, Harald Meller, Kurt W. Alt, Alan Cooper, and Consortium the Genographic, 'Ancient DNA from European Early Neolithic Farmers Reveals Their near Eastern Affinities', *PLoS Biol,* 8 (2010), e1000536.

5 Guido Brandt, Wolfgang Haak, Christina J. Adler, Christina Roth, Anna Szécsényi-Nagy, Sarah Karimnia, Sabine Möller-Rieker, Harald Meller, Robert Ganslmeier, Susanne Friederich, Veit Dresely, Nicole Nicklisch, Joseph K. Pickrell, Frank Sirocko, David Reich, Alan Cooper, Kurt W. Alt, and Consortium The Genographic, 'Ancient DNA Reveals Key Stages in the Formation of Central European Mitochondrial Genetic Diversity', *Science,* 342 (2013), 257-61.

Supplementary Information 11

Statistical evidence for at least three source populations for present-day Europeans

Nick Patterson*, Iosif Lazaridis and David Reich

* To whom correspondence should be addressed (nickp@broadinstitute.org)

Overview

In a previous study on Native American population history, we showed that it is possible to provide formal evidence for a minimum number of migrations into the ancestors of a test set of populations[1].

The method involves studying a matrix of f_4-statistics relating a set of test populations to a set of proposed outgroups.

To infer the minimum number of ancestral populations that must have mixed to form the test set of populations, the method exploits the fact that each of these ancestral mixing populations must have had a vector of f_4-statistics relating them to the outgroup populations.

Thus, the test populations today must be linear combinations of these ancestral f_4-statistic vectors.

By using linear algebra techniques to infer the minimum number of ancestral f_4-statistic vectors that are necessary (in linear combination) to explain the f_4-statistic vectors in all the test populations, we can infer a minimum on the number of migration events that must have occurred.

Concretely, we have a scenario where we have a set of "left" populations L (proposed outgroups) and a set of "right" populations R (test populations from a geographic region of interest, like Europe or the Americas) (Note S6 of ref. 1). We define:

$$X(l, r) = f_4(l_0, l; r_0, r) \quad \text{(S11.1)}$$

Here, l_0, r_0 are arbitrarily chosen "base" populations within the sets L and R, and l, r range over all choices of other populations in L and R. The choice of "base" populations does not matter statistically (we obtain mathematically identical results for any choice of base population).

We showed in[1] that if $X(l, r)$ has rank r and there were n waves of immigration into R with no back-migration from R to L, then:

$$r+1 \leq n \quad \text{(S11.2)}$$

We used this to show that there were at least 3 waves of immigration into pre-Colombian America.

Evidence for at least three source populations for most present-day Europeans

To investigate whether a subset of European populations could be derived from n waves of immigration, or equivalently that $X(l,r)$ has rank $n+1$, we with the following sets L and R:

L = *{Stuttgart, Loschbour, MA1, Onge, Karitiana, Mbuti}*

R = *{Albanian, Basque, Belorussian, Bulgarian, Croatian, Czech, English, Estonian, French, French_South, Greek, Hungarian, Icelandic, Italian, Lithuanian, Norwegian, Orcadian, Pais_Vasco, Sardinian, Scottish, Spanish, Tuscan, Ukrainian}*

The set *L* is chosen to match the populations used in SI 11 for modeling, and includes a Sub-Saharan African group (Mbuti), two eastern non-Africans (Onge and Karitiana) that are differentially related to West Eurasians and MA1, and the three representatives of the ancestral populations inferred by our study (the Stuttgart individual representing EEF, the Loschbour individual representing WHG, and the MA1 individual representing ANE). The set *R* includes all populations identified in both SI 11 and SI 13 as compatible with being derived from the same 3 ancestral populations, and excludes Sicilians, Maltese, Ashkenazi Jews, Finnish, Russians and Mordovians as suggested in the analysis of that note which showed that these populations have evidence of additional complex history.

From the f_4 statistics we can empirically estimate the matrix *X* and test its consistency with a specified rank as described in ref. 1. For each possible rank *r* we assume that *X* has that rank (a null hypothesis) and test *X* for rank *r+1*. In our previous study[1], we published a likelihood ratio test that yields a χ^2 statistic to evaluate the consistency of this null hypothesis with the data[1]. In the tables below we present *r*, the number of degrees of freedom (d.o.f), the χ^2 statistic value, and a P-value.

For the chosen *L* and *R* lists, we find that rank 2 is excluded, and hence at least 4 ancestral populations have contributed to the populations of *R* (Table S11.1).

Table S11.1: At least 4 ancestral populations for 23 European groups. *Rank 2 is excluded ($p<10^{-12}$), so rank 3, or at least 4 ancestral populations are inferred for European populations.*

R	d.o.f.	χ^2	P-value
0	26	2088.9	$<10^{-12}$
1	24	740.8	$<10^{-12}$
2	22	149.4	$<10^{-12}$
3	20	30.4	0.063
4	18	15.1	0.654

The finding of at least 4 ancestral populations is seemingly at odds with our modeling approach which assumes 3 populations, so we sought to determine the cause of the added complexity.

We removed each of the populations of *R* in turn and repeated the analysis over all 23 subsets. If the 4th ancestral population has largely affected only one of the populations in *R*, the evidence for four populations should disappear or greatly weaken when one of the affected population is removed.

We find that the P-value for rank 2 remains $<10^{-12}$ for 22 subsets, but for the subset *R*-{Spanish} it becomes 0.019, which is not significant after correcting for multiple hypothesis testing.

We next repeated the analysis of 253 subsets, removing all pairs of populations in turn. Again, for the vast majority of subsets the *P*-value for rank 2 remains $<10^{-12}$ but for all 22 pairs involving Spanish and another population, the P-value increases, ranging from 0.013-0.104, all non-significant.

We conclude that additional complexity exists in the Spanish population. It is possibly that this is due to the presence of low levels of Sub-Saharan ancestry in the Mediterranean[2] or of North African[3] admixture as has been reported previously. Such ancestry has also been suggested to occur at low levels in other European populations, and perhaps the Spanish stand out in our analysis because of their large sample size.

To shed more light on the additional source of ancestry that we detected in the Spanish we used ALDER[4], a method that uses admixture linkage disequilibrium to infer the time and extent of admixture. We used Mbuti, Yoruba, and Mozabite as African reference populations (Table S11.2).

This analysis confirm that gene flow from Sub-Saharan or North African populations has occurred in the Spanish sample.

Table S11.2: Estimates of African admixture in Spanish population. *The Spanish population may harbor some African-related admixture representing a fourth wave of migration into Europe, but affecting Spain much more than the other groups.*

	African admixture (%)		Time of African admixture (%)	
African reference	**Lower bound**	**Std. error**	**Generations**	**Std. Error**
Mbuti	0.7	0.1	66.2	9.7
Yoruba	1.5	0.2	65.5	9.7
Mozabite	12.6	2.0	73.7	10.4

Adding outgroups to a minimal set of European populations
A different approach is not to start with the full set of populations, but to choose a "small" *R* as:

R = {Belorussian, Bulgarian, Croatian, Czech, English, French, Hungarian, Icelandic, Norwegian, Orcadian, Sardinian, Scottish}

This set of populations includes members of the main south-north European cline (Fig. 1B), and avoids most Mediterranean and Baltic populations where there may be more complex history involving Near Eastern, African, or East Eurasian ancestry.

We want to investigate whether this "simpler" subset of populations could be the result of admixture between only two ancestral populations. We had to "guess" a smaller set because of the combinatorial explosion of possible subsets of 23 populations (e.g., 1,352,078 possible subsets of 12 populations).

We first used a minimal set of proposed outgroup populations *L*:

L = {MA1, Karitiana, Stuttgart, Loschbour}

We find that rank 1 is excluded ($P < 10^{-12}$), and thus there must be at least 3 source populations related to the outgroups even for this restricted set of European populations (Table S11.3)

Table S11.3. Test for L = {MA1, Karitiana, Stuttgart, Loschbour}

R	d.o.f.	χ^2	P-value
0	13	1067	$<10^{-12}$
1	11	121	$<10^{-12}$
2	9	10.5	0.312

We next added Onge and Yoruba to *L* (the Onge are an indigenous group from the Andaman Islands who have been genetically isolated for tens of thousands of years[5]). Again the data indicate at least 3 source populations, without significant evidence for more (Table S11.4).

Table S11.4. Test for L = {MA1, Karitiana, Stuttgart, Loschbour, Onge, Yoruba}

R	d.o.f.	χ^2	P-value
0	15	1504	$<10^{-12}$
1	13	145	$<10^{-12}$
2	11	17	0.114

A limitation of these methods is that they only work when there has been no back-migration from the populations related to the test set *R* into the ancestors of the outgroups *L*. In Native Americans, this seemed like a reasonable assumption, although even here there is evidence of back-migration from Native Americans into far northeastern Siberians (Naukan and Chukchi)[1].

For West Eurasians, the situation is potentially more problematic, as Europe and the Near East (and the Caucasus) have been far from isolated. Thus if enough Near East populations are introduced into *L* we can expect that the rank of *X* will increase if we have enough statistical power. In practice, however, such effects are mild. Specifically, we added each population *P* from the following list to the outgroup set *L* consisting of four populations.

> *P = {Abkhasian, Armenian, Ashkenazi_Jew, BedouinA, BedouinB, Chechen, Cypriot, Dinka, Druze, Georgian, Georgian_Jew, Han, Iranian, Iranian_Jew, Iraqi_Jew, Jordanian, Kalmyk, Lebanese, Libyan_Jew, Moroccan_Jew, Onge, Palestinian, Saudi, Syrian, Tunisian_Jew, Turkish, Turkish_Jew, Turkmen, Vishwabrahmin, Yemenite_Jew, Yoruba}*

For each population *P* in turn we computed the χ^2 statistic (here with 12 d.o.f.) for the null that the rank of *X* is 2.

The smallest P-value that we obtained was 0.024 for the 7 samples from *Turkish_Jew* population. On further exploration we obtained a P-value of 0.000048 (which is likely significant even after correcting for multiple hypothesis testing) by adding both *Yoruba* and *Turkish_Jew* to the 4 population *L* set and testing for consistency with rank 2. The underlying genetic history here is not clear to us. We conclude that the set *R* of European populations specified above cannot have arisen from a mixture of as few as 2 ancestral populations, but there is no strong evidence for more than 3 even when we add additional outgroup populations.

Conclusion

The strength of the approach in this section is that it formally tests for the number of ancestral components for all populations in *R* without assuming a model of population relationships.

Our results confirm that a large number of European populations cannot be derived from a mixture of just two ancestral populations. However, large subsets of populations are formally consistent with a mixture of at least three ancestral populations, without substantial evidence for a fourth ancestral population if the added complexity in the Spanish population is removed.

Finally we find that even for a much reduced set of European populations, at least three ancestral populations are inferred, and that this result is robust to addition of many non-European populations into the outgroup panel.

We anticipate that with larger population sample sizes additional minor inputs into Europe may be identified, further refining the history of European populations beyond the three ancestral populations identified by our study. However, these results increase our confidence that a model of three ancestral inputs can explain important features of the data.

References

1 D. Reich, N. Patterson, D. Campbell, A. Tandon, S. Mazieres, N. Ray, M. V. Parra, W. Rojas, C. Duque, N. Mesa, L. F. Garcia, O. Triana, S. Blair, A. Maestre, J. C. Dib, C. M. Bravi, G. Bailliet, D. Corach, T. Hunemeier, M. C. Bortolini, F. M. Salzano, M. L. Petzl-Erler, V. Acuna-Alonzo, C. Aguilar-Salinas, S. Canizales-Quinteros, T. Tusie-Luna, L. Riba, M. Rodriguez-Cruz, M. Lopez-Alarcon, R. Coral-Vazquez, T. Canto-Cetina, I. Silva-Zolezzi, J. C. Fernandez-Lopez, A. V. Contreras, G. Jimenez-Sanchez, M. J. Gomez-Vazquez, J. Molina, A. Carracedo, A. Salas, C. Gallo, G. Poletti, D. B. Witonsky, G. Alkorta-Aranburu, R. I. Sukernik, L. Osipova, S. A. Fedorova, R. Vasquez, M. Villena, C. Moreau, R. Barrantes, D. Pauls, L. Excoffier, G. Bedoya, F. Rothhammer, J. M. Dugoujon, G. Larrouy, W. Klitz, D. Labuda, J. Kidd, K. Kidd, A. Di Rienzo, N. B. Freimer, A. L. Price, and A. Ruiz-Linares, 'Reconstructing Native American Population History', *Nature,* 488 (2012), 370-4.

2 Priya Moorjani, Nick Patterson, Joel N. Hirschhorn, Alon Keinan, Li Hao, Gil Atzmon, Edward Burns, Harry Ostrer, Alkes L. Price, and David Reich, 'The History of African Gene Flow into Southern Europeans, Levantines, and Jews', *PLoS Genet,* 7 (2011), e1001373.

3 Laura R. Botigué, Brenna M. Henn, Simon Gravel, Brian K. Maples, Christopher R. Gignoux, Erik Corona, Gil Atzmon, Edward Burns, Harry Ostrer, Carlos Flores, Jaume Bertranpetit, David Comas, and Carlos D. Bustamante, 'Gene Flow from North Africa Contributes to Differential Human Genetic Diversity in Southern Europe', *Proceedings of the National Academy of Sciences* (2013).

4 Po-Ru Loh, Mark Lipson, Nick Patterson, Priya Moorjani, Joseph K. Pickrell, David Reich, and Bonnie Berger, 'Inferring Admixture Histories of Human Populations Using Linkage Disequilibrium', *Genetics,* 193 (2013), 1233-54.

5 Kumarasamy Thangaraj, Gyaneshwer Chaubey, Toomas Kivisild, Alla G. Reddy, Vijay Kumar Singh, Avinash A. Rasalkar, and Lalji Singh, 'Reconstructing the Origin of Andaman Islanders', *Science,* 308 (2005), 996-96.

Supplementary Information 12
Admixture Graph Modeling

Iosif Lazaridis*, Nick Patterson and David Reich

* To whom correspondence should be addressed (lazaridis@genetics.med.harvard.edu)

We use ADMIXTUREGRAPH methodology[1] as implemented in the *qpGraph* software of ADMIXTOOLS[2] in order to investigate the relationships of Stuttgart and Loschbour to present-day human populations from Eurasia, Oceania, and the Americas. This allows us to test models relating a number of populations that may also contain admixture edges. Our purpose is not so much to uncover the deep prehistoric relationships of present-day humans –which may be fairly complex– but rather to show how several simple models can be falsified. We also propose alternatives that are consistent with the available data, and identify a parsimonious model that fits the data successfully and makes predictions that are consistent with those of a model-free methodology described in SI 13.

We begin by investigating some simple relationships using f_4-statistics which will inform the more detailed models we will later investigate. We will use a set of populations identified by ADMIXTURE analysis (SI 9) which encompass different aspects of human variation. For each non-West Eurasian geographical region we will show statistics of the form $f_4(Ancient_1, Ancient_2; Non\ West\ Eurasian, Chimp)$ and $f_4(Ancient, Chimp; Non\ West\ Eurasian_1, Non\ West\ Eurasian_2)$ that test, respectively, whether two ancient individuals form a clade with respect to Non West Eurasians and whether two Non West Eurasian groups form a clade with respect to an ancient Eurasian.

We report only statistics with $|Z|\geq 2$ in the tables that follow.

Relationship of ancient samples to Onge

We first consider the relationship of ancient samples to Onge (indigenous Little Andaman Islanders[3]), an island population from the Bay of Bengal without very close relatives and that is distantly related to Ancestral South Indians[1].

Table S12.1: Onge are closer to Eurasian hunter-gatherers than to Stuttgart.

$Ancient_1$	**$Ancient_2$**	***$f_4(Onge, Chimp; Ancient_1, Ancient_2)$***	**Z**
Loschbour	Stuttgart	0.00191	3.452
MA1	Stuttgart	0.001842	2.987
Motala12	Stuttgart	0.002043	3.512

The results of Table S12.1 provide suggestive evidence that Onge share more common ancestry with hunter-gatherers than with Stuttgart. All statistics involving two hunter-gatherer populations have $|Z|<0.9$, so ancient Eurasian hunter-gatherers are approximately symmetrically related to Onge, and they are all more closely related to them than is Stuttgart.

Relationship of ancient samples to East Asia

We next consider the relationship of ancient samples to East Asia using the set (Ami, Atayal, Han, Naxi, She). East Asians are more closely related to all hunter-gatherers than to Stuttgart, but there are no significant differences between hunter-gatherers (all such statistics have $|Z|<1.1$) (Table S12.2).

We also found no significant statistics of the form $f_4(Ancient, Chimp; East\ Asian_1, East\ Asian_2)$ (all $|Z|<2$). Thus, there is no evidence of differential relatedness of East Asians to ancient west Eurasians.

Table S12.2: East Asians are more closely related to ancient hunter-gatherers than to Stuttgart.

East Asian	Ancient$_1$	Ancient$_2$	f_4(East Asian, Chimp; Ancient$_1$, Ancient$_2$)	Z
Ami	Loschbour	Stuttgart	0.001745	3.424
Ami	MA1	Stuttgart	0.001751	2.884
Ami	Motala12	Stuttgart	0.001357	2.414
Atayal	Loschbour	Stuttgart	0.001463	2.722
Atayal	MA1	Stuttgart	0.001599	2.518
Atayal	Motala12	Stuttgart	0.00146	2.443
Han	Loschbour	Stuttgart	0.001634	3.275
Han	MA1	Stuttgart	0.001548	2.634
Han	Motala12	Stuttgart	0.001494	2.729
Naxi	Loschbour	Stuttgart	0.001592	3.097
Naxi	MA1	Stuttgart	0.001729	2.891
Naxi	Motala12	Stuttgart	0.001592	2.82
She	Loschbour	Stuttgart	0.001814	3.538
She	MA1	Stuttgart	0.001719	2.824
She	Motala12	Stuttgart	0.001561	2.771

Relationship of ancient samples to Oceania

We consider the relationship of ancient samples to Oceania using the set (Papuan, Bougainville). The statistics in Table S12.3 border on |Z|=3 and are suggestive that hunter-gatherer groups share more genetic drift with Oceanian populations than with Stuttgart. All statistics involving two ancient hunter-gatherers are non-significant with |Z|<0.8.

Statistics of the form f_4*(Ancient, Chimp; Bougainville, Papuan)* (not shown) are all positive (|Z|>2.3) but do not suggest gene flow between Bougainville and west Eurasia, as they are affected by differential Denisovan admixture into the two Oceanian groups[4]. We conclude that Oceanian populations are genetically closer to Eurasian hunter-gatherers than to Stuttgart.

Table S12.3: Oceanian populations are genetically closer to hunter-gatherers than to Stuttgart.

Oceanian	Ancient$_1$	Ancient$_2$	f_4(Oceanian, Chimp; Ancient$_1$, Ancient$_2$)	Z
Bougainville	Loschbour	Stuttgart	0.001566	2.951
Bougainville	MA1	Stuttgart	0.001491	2.337
Bougainville	Motala12	Stuttgart	0.001686	3.012
Papuan	Loschbour	Stuttgart	0.001364	2.599
Papuan	MA1	Stuttgart	0.00141	2.165
Papuan	Motala12	Stuttgart	0.001609	2.854

Relationship of ancient samples to the Americas

We explore the relationship of ancient samples to the Americas using Native Americans without post-Colombian European admixture (Karitiana, Mixe, Piapoco, Surui) (Extended Data Fig. 3, K=10).

The pattern of Table S12.4 is different from that in the previous sections: Native American populations are more closely related to hunter-gatherers than to Stuttgart, but also more closely related to MA1 than to the European hunter-gatherers. This recapitulates the recently reported evidence of gene flow involving MA1 and the ancestors of Native Americans[5]. In this paper we use the Karitiana as a recently unadmixed population[6] with the largest sample size in the Human Origins dataset to investigate more ancient gene flow between the Americas and Eurasia.

Table S12.4: Native American populations more closely related to ancient hunter-gatherers than to Stuttgart, and more closely related to MA1 than to European hunter-gatherers.

Nat.Am.	$Ancient_1$	$Ancient_2$	f_4(Nat.Am., Chimp; $Ancient_1$, $Ancient_2$)	Z
Karitiana	Loschbour	Stuttgart	0.002813	4.861
Karitiana	MA1	Loschbour	0.004746	7.056
Karitiana	MA1	Motala12	0.003438	4.763
Karitiana	MA1	Stuttgart	0.007701	11.423
Karitiana	Motala12	Loschbour	0.001377	2.220
Karitiana	Motala12	Stuttgart	0.00421	6.728
Mixe	Loschbour	Stuttgart	0.002497	4.507
Mixe	MA1	Loschbour	0.004386	6.718
Mixe	MA1	Motala12	0.002768	4.122
Mixe	MA1	Stuttgart	0.006886	10.869
Mixe	Motala12	Loschbour	0.001478	2.534
Mixe	Motala12	Stuttgart	0.004071	6.743
Piapoco	Loschbour	Stuttgart	0.002976	5.129
Piapoco	MA1	Loschbour	0.004136	6.286
Piapoco	MA1	Motala12	0.002865	4.193
Piapoco	MA1	Stuttgart	0.007275	11.246
Piapoco	Motala12	Stuttgart	0.004208	6.868
Surui	Loschbour	Stuttgart	0.002905	4.763
Surui	MA1	Loschbour	0.00385	5.559
Surui	MA1	Motala12	0.00303	4.190
Surui	MA1	Stuttgart	0.006936	10.041
Surui	Motala12	Stuttgart	0.003693	5.697

Relationship of ancient samples to eastern non-Africans

We finally explore the relationship of ancient samples to all eastern non-Africans (ENA) together using the set (Onge, Papuan, Atayal, Karitiana). Table S12.5 shows that that Papuans universally appear to be most distant from ancient Eurasians than any other ENA population, consistent with their additional admixture from archaic Denisovans. Comparisons involving (Onge, Atayal) slightly favor Atayal, but barely reach significance and we do not view this evidence as compelling. Karitiana, on the other hand appear generally closer to present-day west Eurasians than all other ENA populations. We will thus develop models for West Eurasia that take into account Karitiana and Onge, to account for both the specific link between MA1 and Native Americans and the more general link between eastern non-Africans and ancient Eurasian hunter-gatherers.

Summary of lessons from f_4-statistics

Our systematic survey of f_4-statistics serves to identify features of the relationships between different populations that must be accounted for in a successful model. We itemize the most pertinent observations from our survey:

1. Ancient Eurasians (Europeans and MA1) are genetically closest to Karitiana, intermediately related to Onge/Atayal, and least related to Papuans
2. Hunter-gatherers do not differ in their relationships to eastern non-Africans, except for Karitiana where MA1 is clearly more related to them than are the European hunter-gatherers.
3. Eastern non-Africans are all more closely related to ancient hunter gatherers than to Stuttgart

We confirm these findings on subsets of all SNPs ascertained in a Yoruba and a San individual (Extended Data Table 3). We refer to these items in what follows as we begin to explore the space of possible models.

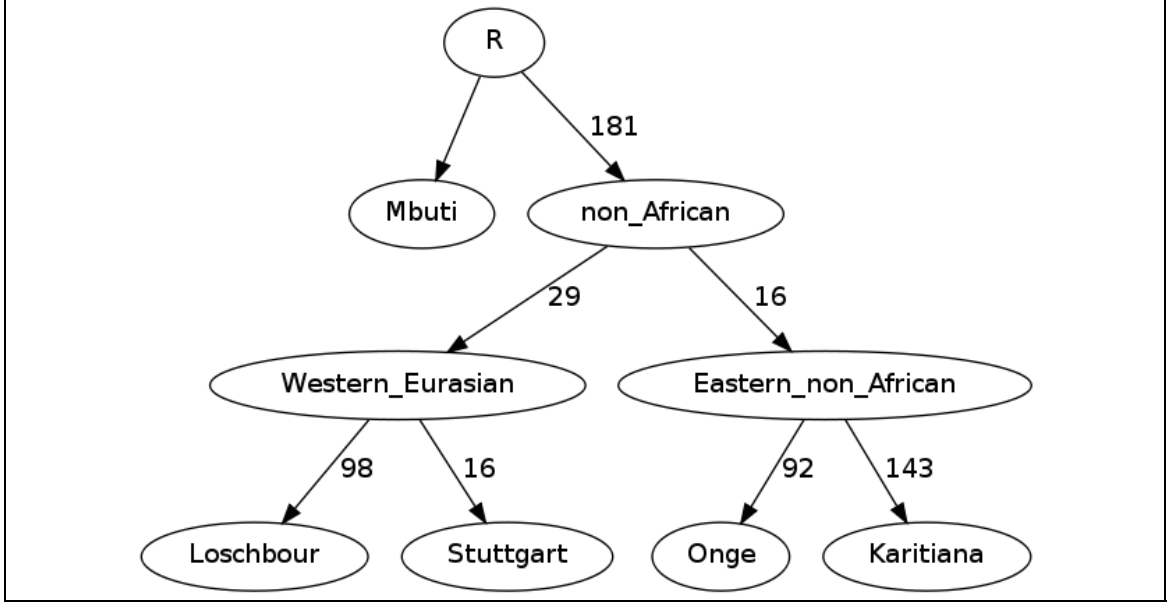
Table S12.5: Ancient Eurasians are closest to Karitiana and most distant to Papuans.

Ancient	ENA_1	ENA_2	f_4(Ancient, Chimp; ENA_1, ENA_2)	Z
Loschbour	Atayal	Papuan	0.004245	8.215
Loschbour	Karitiana	Atayal	0.002944	6.678
Loschbour	Karitiana	Onge	0.003188	6.406
Loschbour	Karitiana	Papuan	0.007189	12.361
Loschbour	Onge	Papuan	0.004001	7.762
MA1	Atayal	Papuan	0.00414	8.192
MA1	Karitiana	Atayal	0.007888	17.122
MA1	Karitiana	Onge	0.008267	15.464
MA1	Karitiana	Papuan	0.012028	20.574
MA1	Onge	Papuan	0.00376	7.506
Motala12	Atayal	Papuan	0.003802	7.304
Motala12	Karitiana	Atayal	0.004391	10.034
Motala12	Karitiana	Onge	0.004565	9.078
Motala12	Karitiana	Papuan	0.008194	14.258
Motala12	Onge	Papuan	0.003629	7.087
Stuttgart	Atayal	Onge	0.000784	2.013
Stuttgart	Atayal	Papuan	0.004173	8.407
Stuttgart	Karitiana	Atayal	0.001592	3.838
Stuttgart	Karitiana	Onge	0.002376	5.113
Stuttgart	Karitiana	Papuan	0.005765	10.838
Stuttgart	Onge	Papuan	0.003388	6.843

A tree model fails

We begin with a simple model fitted unsuccessfully with ADMIXTUREGRAPH (Fig. S12.1). For example, it predicts that Stuttgart is equally related to Onge and Karitiana (contradicting item #1), and it predicts that Stuttgart and Loschbour are equally related to Karitiana (contradicting item #3). Note that drifts along edges are multiplied by 1000 in this and following figures.

Figure S12.1: A (failed) model with no admixture.



Models with a single admixture edge fail

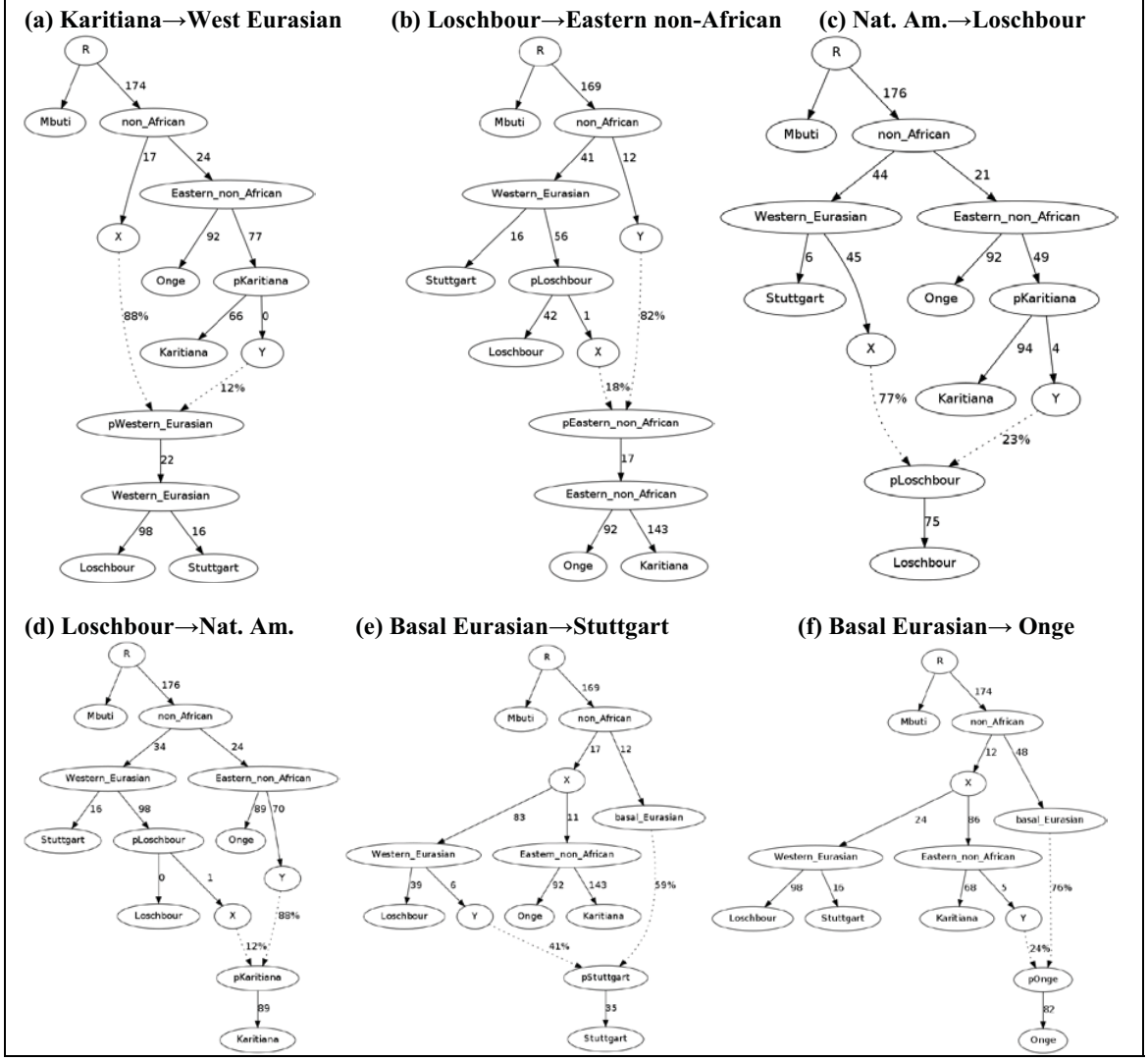
We exhaustively search for amendments to the model of Fig. S12.1 involving a single admixture edge, but find that they all fail to account for the observed f_4-statistics and the asymmetries between both Stuttgart/Loschbour and Onge/Karitiana.

A single admixture event between Eastern non-Africa and West Eurasia fails

We attempted to amend the model by adding one admixture event between the West Eurasian and Eastern non-African subtrees, but this fails:

1. Admixture into Western_Eurasian from the Karitiana branch (Fig. S12.2a) fails, because it predicts that Stuttgart and Loschbour are equally related to Onge (contradicting item #3)
2. Admixture into Eastern_non_African from Loschbour (Fig. S12.2b) fails, because it predicts that Stuttgart are equally related to Karitiana and Onge (contradicting item #1).
3. Admixture into Loschbour from the Karitiana branch (Fig. S12.2c) fails, because it predicts that Stuttgart is equally related to Onge and Karitiana (contradicting item #1)
4. Admixture into Karitiana from Loschbour (Fig. S12.2d) fails, because it predicts that Onge are equally related to Stuttgart and Loschbour (contradicting item #3)

Figure S12.2: Failed models with one admixture edge.



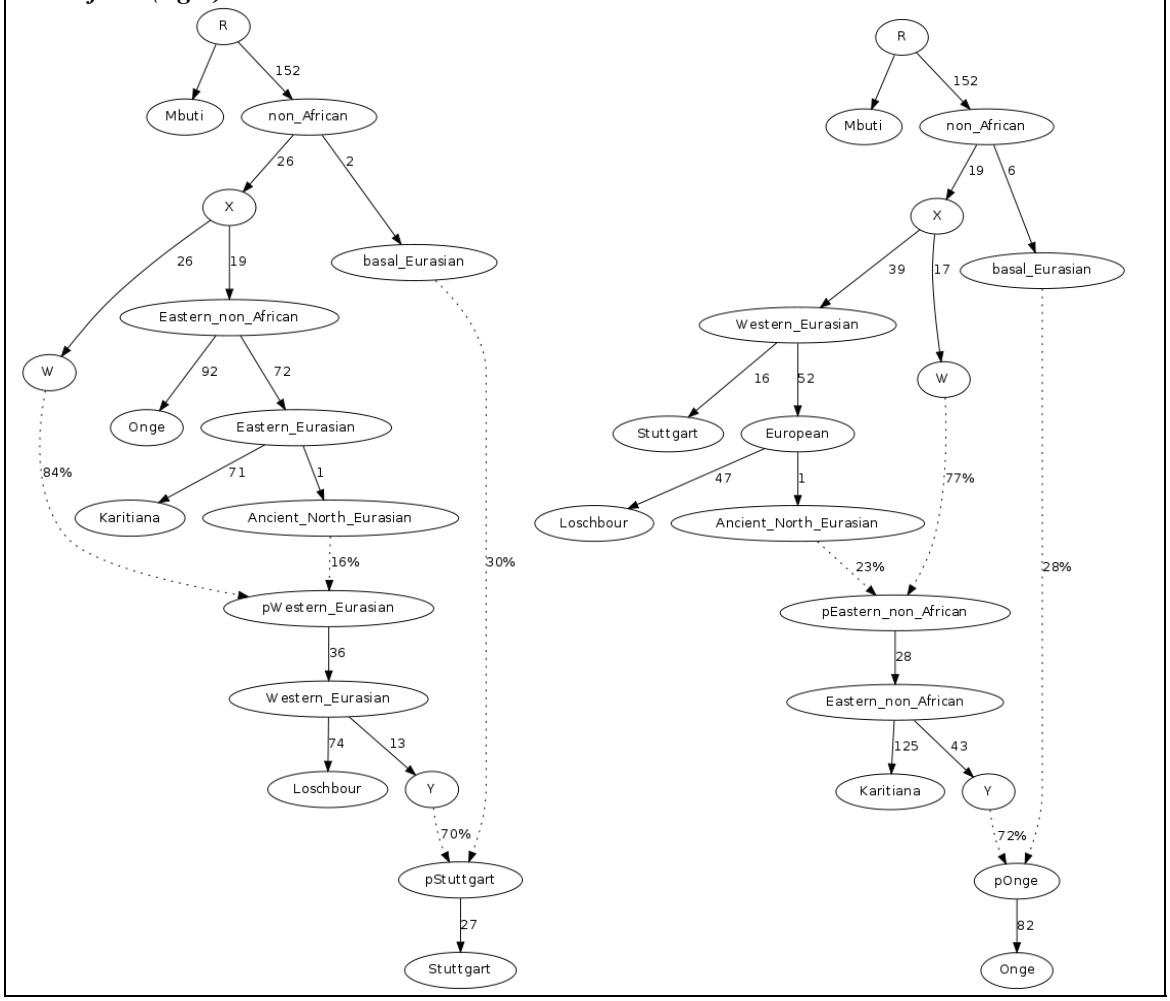
We further considered scenarios of early admixture between West Eurasians and Eastern non-Africans but found that this does not help as it preserves the topological form of Fig. S12.1.

We also considered Eastern non-African admixture into Stuttgart or conversely West Eurasian admixture into Onge, but these break the symmetry in the wrong direction, making fits worse. Thus, a single admixture between Eastern non-Africa and Western Eurasia is insufficient to explain the data.

We also considered a scenario of "Basal Eurasian" admixture into either Stuttgart or Onge (Fig. S12.3e and f respectively). This is admixture from a source that branched off before the divergence of West Eurasians and eastern non-Africans. By adding this type of admixture into Stuttgart we explain greater Loschbour proximity to eastern non-Africans (#3), but not greater proximity of Stuttgart to Karitiana than to Onge (#1). Conversely, by adding this admixture into Onge we explain greater Karitiana proximity to west Eurasia (#1), but not greater proximity of eastern non-Africans to Loschbour than to Stuttgart (#3). Basal Eurasian admixture to either Loschbour or Karitiana break the symmetry in the wrong direction, implying that Karitiana should be closer to Stuttgart than to Loschbour or that Loschbour should be closer to Onge than to Karitiana respectively.

To summarize, models with one admixture edge cannot resolve the observed asymmetries, motivating a search for a model with at least two admixture edges that can fit.

Figure S12.3: Successful models with all admixture taking place in West Eurasia (left) or eastern non-Africa (right)



Successful models with two admixture edges

The idea of basal Eurasian is nonetheless attractive, so we pursued it further.

A single such admixture event into Stuttgart (as in Fig. S12.2e) would fully explain #3, i.e., that all eastern non-Africans are more closely related to hunter-gatherers than to Stuttgart. Such an idea is also archaeologically plausible on account of the Near Eastern related admixture that we have detected in Stuttgart. The Near East was the staging point for the peopling of Eurasia by anatomically modern humans. As a result, it is entirely plausible that it harbored deep Eurasian ancestry which did not initially participate in the northward colonization of Europe, but was later brought into Europe by Near Eastern farmers. More speculatively, some basal Eurasian admixture in the Near East may reflect the early presence of anatomically modern humans[7] in the Levant, or the populations responsible for the appearance of the Nubian Complex in Arabia[8], both of which date much earlier than the widespread dissemination of modern humans across Eurasia. Finally, it could reflect continuing more recent gene flows between the Near East and nearby Africa after the initial out-of-Africa dispersal, perhaps associated with the spread of Y-chromosome haplogroup E subclades from eastern Africa[9, 10] into the Near East, which appeared at least 7,000 years ago into Neolithic Europe[11].

Equally archeologically plausible is basal Eurasian admixture in Onge (Fig. S12.2f) which would partially explain #1. The Onge are a southern Eurasian population, and a scenario of a "southern route" colonization of Eurasia (of which the Onge are plausible partial descendants) might have resulted in them having deep Eurasian ancestry, similar to a model proposed for the early colonization of Australia by anatomically modern humans[12]. Such ancestry would cause them to share less drift with West Eurasians than the Karitiana.

As shown in Fig. S12.2, basal Eurasian admixture into either Stuttgart or Onge fails to explain the data. However, we can combine it with gene flow between west Eurasia and eastern non-Africa in a successful model.

Fig. S12.3 shows scenarios that fit the data involving basal Eurasian admixture. If Stuttgart harbors basal Eurasian admixture (left), then the affinity of Loschbour to eastern non-Africans is maintained, but the greater proximity of Karitiana than Onge to west Eurasians is not. We can amend our model by proposing gene flow from Karitiana into the ancestors of west Eurasians. Note that this admixture must go to the ancestor of west Eurasians, because both Stuttgart and Loschbour are genetically closer to Karitiana than to Onge (#2). The situation is symmetrical if Onge has basal Eurasian admixture (right), in which case the affinity of west Eurasians to Karitiana is maintained, but the greater proximity of Loschbour to eastern non-Africans (#3) is not; this can be fixed by proposing admixture from relatives of Loschbour into the ancestor of eastern non-Africans. In both the models of Fig. S12.3, all admixture takes place either in west Eurasia (left), or eastern non-Africa (right), with the other populations not being admixed.

Fig. S12.4 proposes a second set of possibilities, also involving basal Eurasian admixture. If Stuttgart has basal Eurasian admixture (left), then the greater proximity of Karitiana than Onge to West Eurasians could be explained by gene flow from West Eurasians into Native American ancestors; this could originate either in the Loschbour branch (top row), or from a basal West Eurasian lineage (bottom). Symmetrically, basal Eurasian admixture into Onge (right) can be combined with eastern non-African gene flow into Loschbour from Native Americans (top) or eastern non-Africans (bottom).

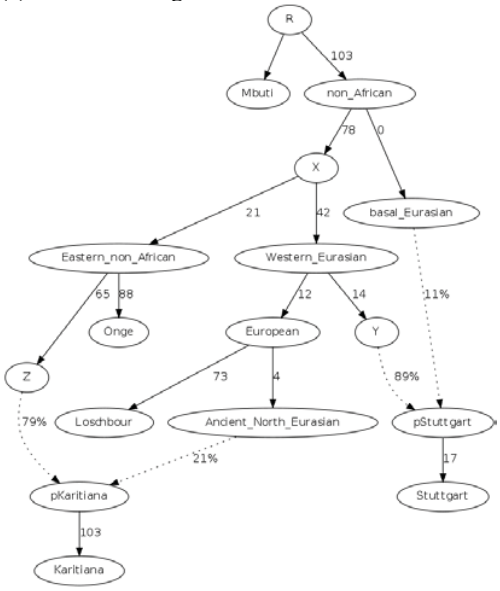
In Fig. S12.5 we propose two successful models without basal Eurasian admixture but that instead invoke variable admixture in either direction across Eurasia. These models propose two admixture events for the set of considered populations, but make Karitiana and Onge (left) and Loschbour and Stuttgart (right) be composed of the same ancestral elements but in different proportions.

We have thus identified a total of eight models (Figs S12.3, S12.4 and S12.5) each with two admixture events that are all consistent with the *f*-statistics for the four populations and yet make quite different predictions about the prehistory of Eurasia. We note that even more complex models could

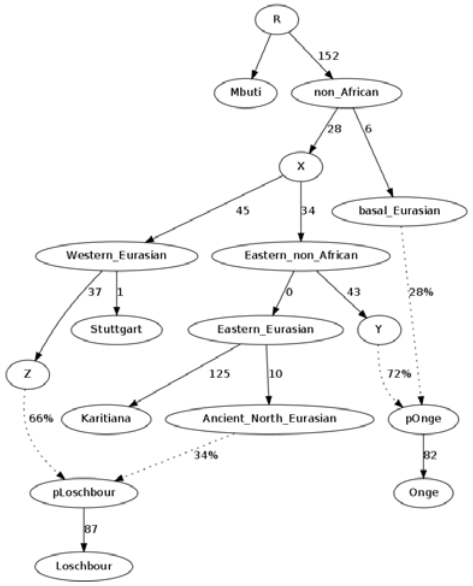
be devised (with more than two admixture events) that would be equally consistent, but may be unparsimonious for a set of only four populations. For the time being, we conclude that very simple models (with one admixture event) fail, while a plethora of available choices exist for slightly more complex models (with two admixture events).

Figure S12.4: Successful models combining basal admixture with a second gene flow event

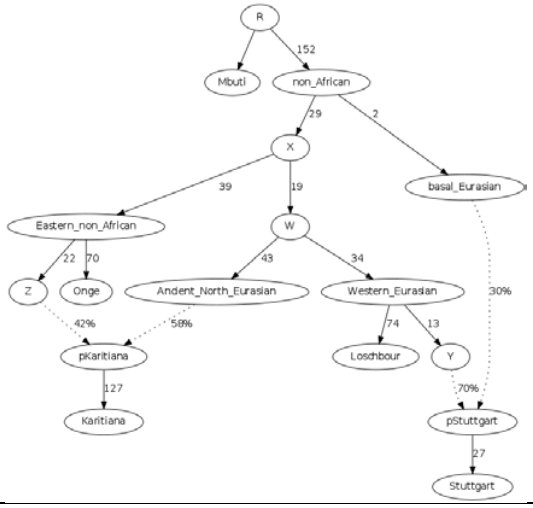
(a) Basal→Stuttgart / Loschbour→Karitiana



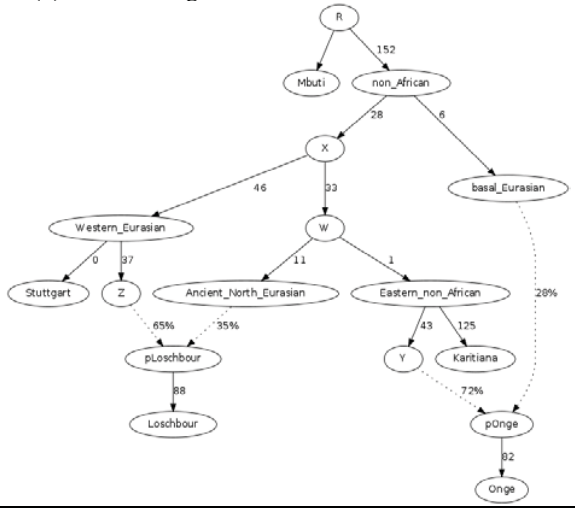
(b) Basal→Onge / Karitiana→Loschbour



(c) Basal→Stuttgart / West Eurasian→Karitiana



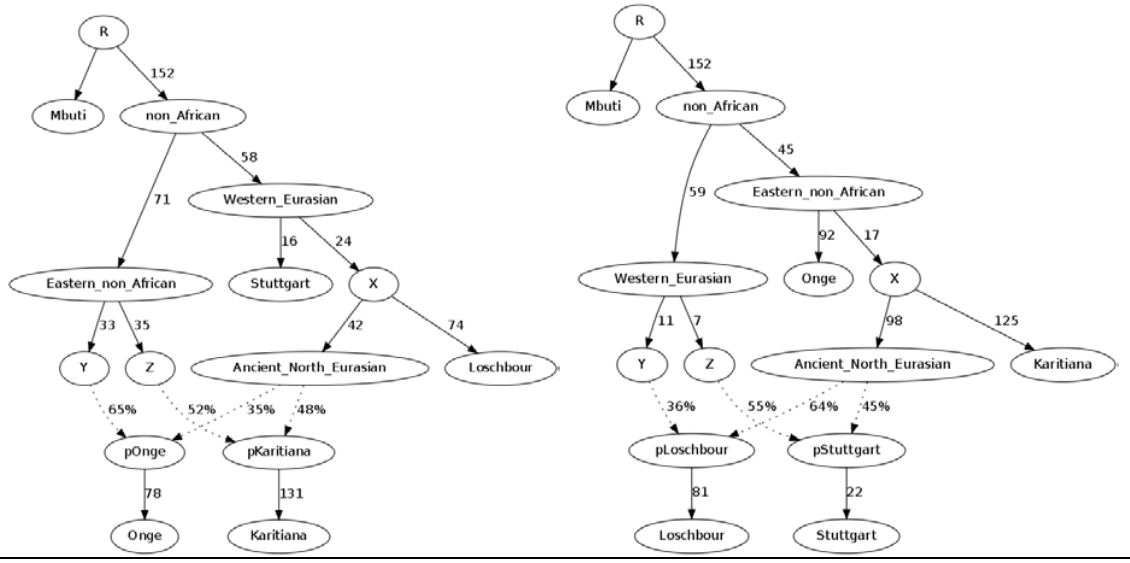
(d) Basal→Onge / Eastern non-African→Loschbour



MA1 as representative of Ancient North Eurasians

A possible way to constrain the choice of model is to attempt to fit additional populations into their structure. MA1 is an Upper Paleolithic Siberian with demonstrated genetic links to both Europe and Native Americans[5] and thus is a powerful sample for constraining possible historical scenarios. It is potentially a "missing link": a representative of a population mediating gene flow between east and west across Eurasia, so we consider whether it could be incorporated into the models of Figs S12.3 to S12.5 without breaking them. We summarize the results for the eight models in Table S12.6.

Figure S12.5: Two successful models with variable Karitiana-related admixture into Loschbour and Stuttgart (left), or variable Loschbour-related admixture into Karitiana and Onge (right).



Only model 4LB is consistent with MA1 being a descendent of the Ancient_North_Eurasian node mediating gene flow between West Eurasia and Eastern non-Africans. For the remaining models we list the *f*-statistics that are most discrepant between model and data together with their Z value; for example model 4RT makes Karitiana and MA1 sister clades, so we fit zero for the violating statistic, but we in fact observe a positive value with Z=10.5. We conclude that a model in which (i) gene flow into the Karitiana originated from a basal West Eurasian population and (ii) Neolithic farmers such as Stuttgart had admixture from a Basal Eurasian population is consistent with the evidence, and we will explore this model further. We show the fitted model 4LB with MA1 in Fig. S12.6.

Table S12.6: Attempting to fit MA1 into the structure of models of Figures S12.3, S12.4 and S12.5.

ID	Admixture event 1	Admixture event 2	Violation	Z
3L	Basal→Stuttgart	Karitiana→West Eurasian	f_2(Onge, MA1)	5.1
3R	Basal→Onge	Loschbour→East non-African	f_2(Loschbour, Stuttgart)	-6.2
4LT	Basal→Stuttgart	Loschbour→Karitiana	f_3(MA1; Loschbour, Stuttgart)	8.0
4RT	Basal→Onge	Karitiana→Loschbour	f_4(Onge, Stuttgart; Kar., MA1)	10.5
4LB	Basal→Stuttgart	West Eurasian→Karitiana	✓	
4RB	Basal→Onge	East non-African→ Losch.	f_4(Onge, MA1; Losch., Stutt.)	3.8
5L	Loschbour→Karitiana	Loschbour→Onge	f_2(Loschbour, Stuttgart)	-6.0
5R	Karitiana→Loschbour	Karitiana→Stuttgart	f_2(Onge, MA1)	5.2

No evidence of Basal East Asian admixture in MA1

Model 4LB proposes that MA1 is unadmixed, but it was argued[5] that MA1 may have basal East Asian (basal eastern non-African in our terminology) admixture on the evidence that MA1 shares more drift than Sardinians with both Oceanians and East Asians. This was a reasonable suggestion because of the sample's provenance, but statistics of the form f_4(ENA, Chimp; Loschbour, MA1) appear symmetric for any eastern-non African (ENA) population from the set (Ami, Atayal, Han, Naxi, She, Papuan, Bougainville, Onge) with $|Z|<0.3$. If MA1 had more basal East Asian admixture than Loschbour, these statistics should be negative.

Our model provides an alternative explanation for the asymmetry between MA1/Sardinians with respect to ENA, not in terms of admixture into MA1 but with basal Eurasian admixture into Neolithic farmers. This scenario accounts for both the fact that ENA share more drift with MA1 than with

Stuttgart (because Stuttgart has basal Eurasian admixture), and for the fact that Loschbour and MA1 are symmetrically related to ENA (because they both lack Neolithic Near Eastern ancestry).

Figure S12.6: A successful model involving Stuttgart, Loschbour, MA1, Onge, and Karitiana. *The high genetic drift in the MA1-specific branch is an artifact of the low coverage (about 1x) of this sample, which means that many sites that are in fact heterozygous appear as homozygous. However, this is not expected to affect inferences of the relationships between MA1 and the other samples.*

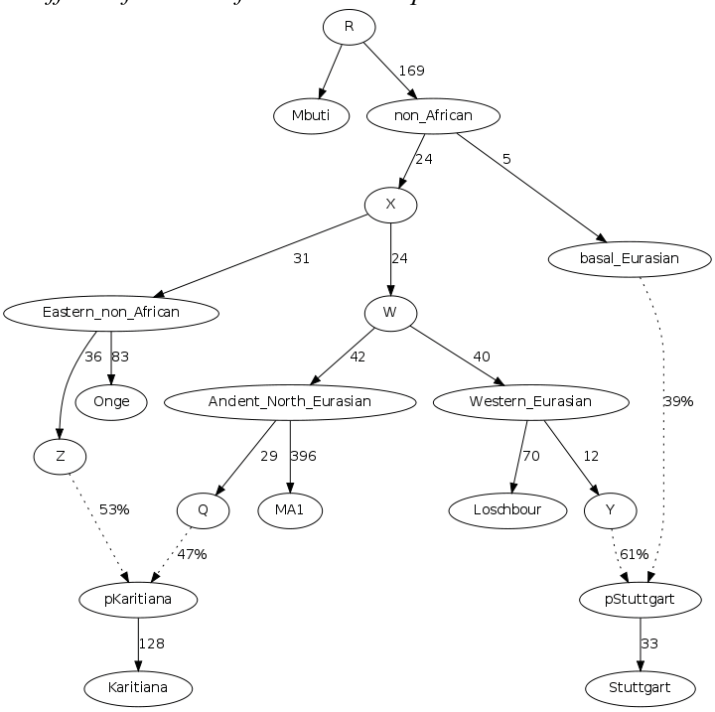
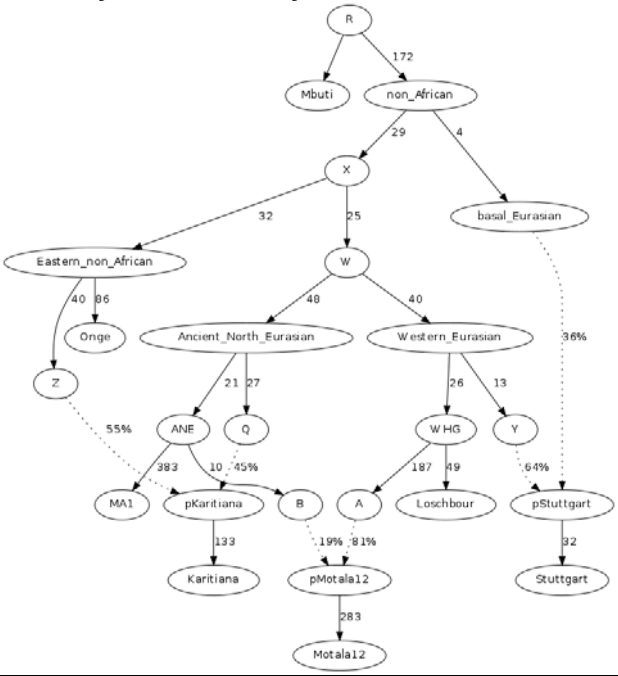


Figure S12.7: Motala12 can be fit as a mixture of Loschbour and MA1



Motala12 is not a clade with Loschbour as it has MA1-related admixture

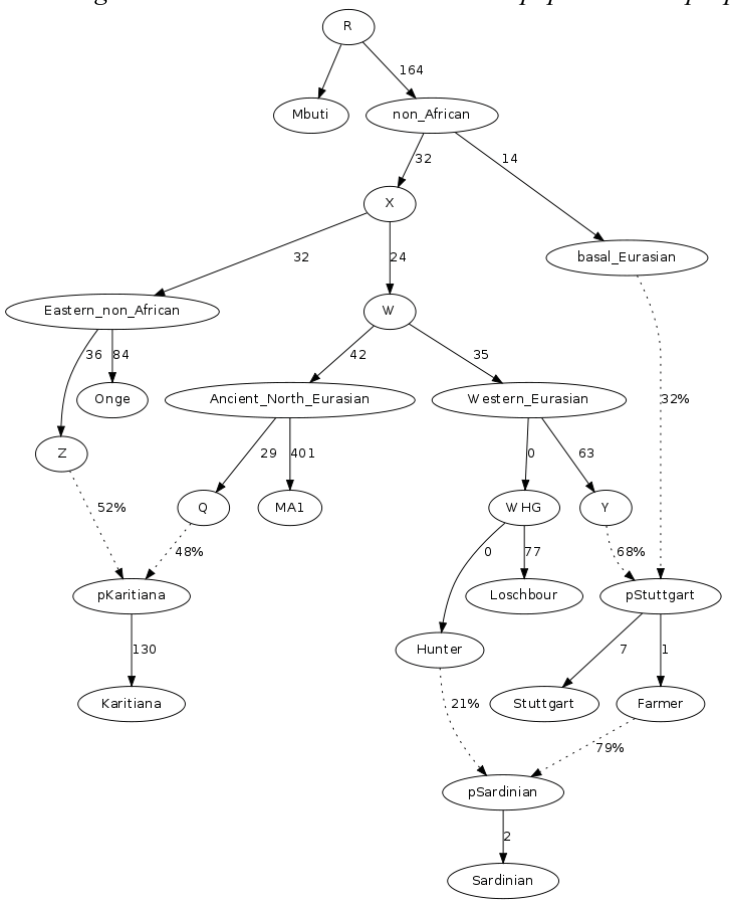
We next attempted to fit Motala12 as a clade with Loschbour in the topology of Fig. S12.6, but were unable to do so, because f_4*(Loschbour, Motala12; Stuttgart, MA1)* is significantly positive (Z=5.6). A possible explanation for this is that the European hunter-gatherers admixing with Near Eastern farmers to form Stuttgart were more like Loschbour than Motala12. However, the statistic f_4*(Motala12; Loschbour; MA1, Mbuti)* is also significantly positive (main text), and this suggests that MA1 and Motala12 share more common drift than MA1 and Loschbour. Scandinavian hunter-gatherers could indeed be fit if they were modeled as a mixture of Loschbour and MA1. This scenario is consistent with the above statistics, Motala12's intermediate geographical position between Western Europe and Siberia, and their intermediate position between West European hunter-gatherers and Ancient North Eurasians (Fig. 1B). The successful fit is shown in Fig. S12.7.

Most Europeans are not a 2-way mixture of Loschbour and Stuttgart

We next attempted to fit individual West Eurasian populations as a mixture of Loschbour and Stuttgart, as representatives of Early European farmers and West European Hunter Gatherers.

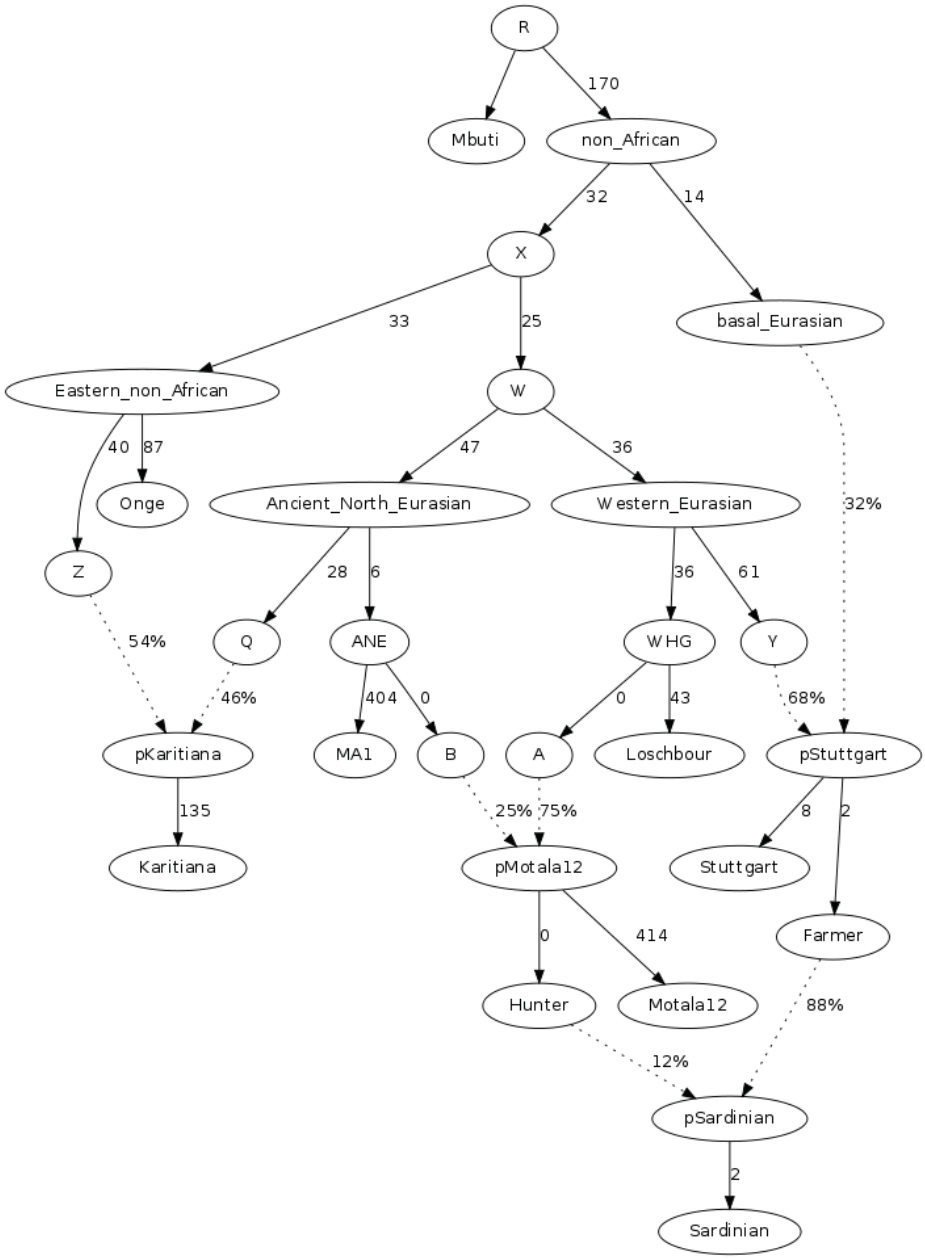
Fig. 1B suggests that this is not possible, as most Europeans form a cline that cannot be reconciled with such a mixture. Nonetheless, for Sardinians (Extended Data Table 1), the most negative f_3-statistic is of the form f_3*(Test; Loschbour, Stuttgart)*, which suggests that at least some Europeans may be consistent with having been formed by such a mixture. We thus fit each European population into the topology of Fig. S12.6. Only Basques, Pais_Vasco, and Sardinians, can be fit successfully with this model. Fig. S12.8 shows a successful fit.

Figure S12.8: A successful 2-way mixture for Sardinians on the Fig. S12.6 scaffold. *They fit as a mix of Loschbour and Stuttgart-related "Hunter" and "Farmer" populations in proportions 21/79%.*



Most European populations cannot be fit as this type of 2-way mixture and, intuitively, this is due to their tendency (Fig. 1B) towards Ancient North Eurasians that is not modeled by such a mixture. Indeed, when we examined the set of f_4-statistics exceeding |Z|>3 for European populations, MA1 was involved for all populations who did not fit the model structure of Fig. S12.8, ranging from Bergamo (fitted f_4*(Loschbour, MA1; Stuttgart, Bergamo)* = -0.002162, Z=3.04 standard errors lower than the estimated value of 0.003951) to Mordovians (fitted *f4(Stuttgart, Mordovian; MA1, Mordovian)* = 0.000886, Z=7.4 standard errors higher than the estimated value of -0.010302).

Figure S12.9: A successful 2-way mixture for Sardinians on the Fig. S12.7 scaffold. *They fit as a mix of Motala12 and Stuttgart-related "Hunter" and "Farmer" populations in proportions 12/88%.*

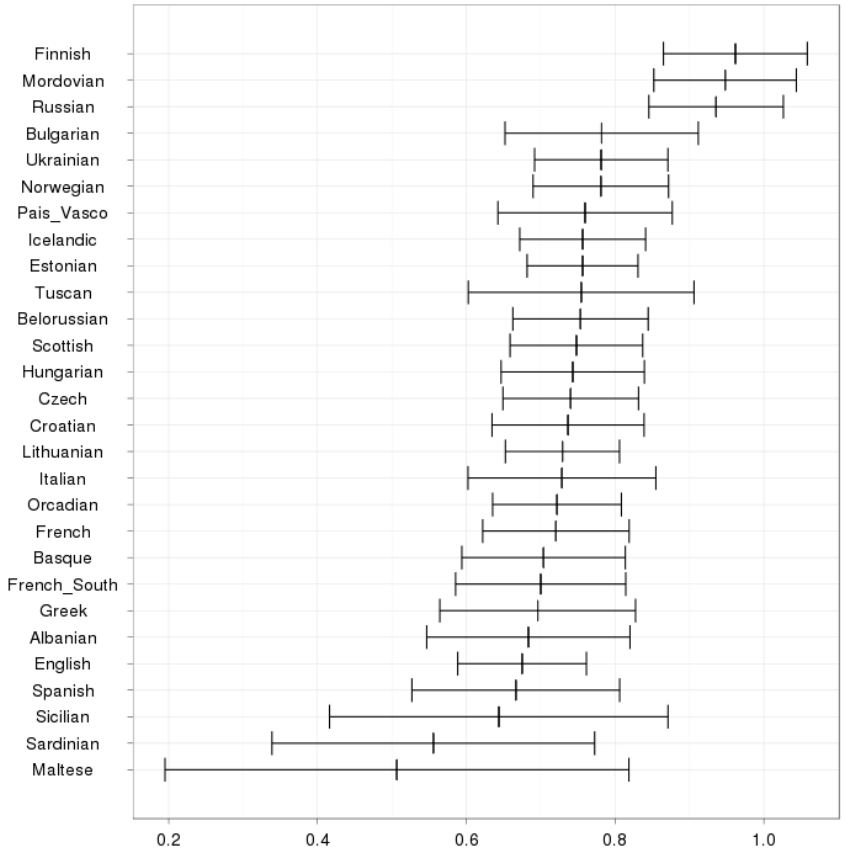


Most Europeans are not a 2-way mixture of Motala12 and Stuttgart

The fact that a Stuttgart/Loschbour mixture did not preserve the relationship of European populations to MA1 motivated us to model them as a Stuttgart/Motala12 mixture, given the evidence that

Motala12 has some MA1-related admixture. Fig. 1B suggests that this may not be enough to explain the data, since, despite being intermediate between Loschbour and MA1, Scandinavian hunter gatherers are still fairly close to Western European ones. We thus fit individual European populations into the topology of Fig. S12.7, but, only Basque, French_South, and Sardinian could be accommodated. We show a successful fit for Sardinians in Fig. S12.9. We do not propose that southwestern Europeans were formed by a mixture of Early European Farmers and Scandinavian hunter-gatherers, but the fact that they can be fit as such indicates that Scandinavian hunter-gatherers were close enough to their West European relatives so that they can serve as a proxy for them.

Figure S12.10: The ratio* f4(X, Stuttgart; Karitiana, Chimp) / f4(X, Stuttgart, MA1, Chimp) *is lower than* 1 *for different European populations. *This suggests that MA1 is a better surrogate for Ancient North Eurasians than is Karitiana. The bars indicate ± 1 standard error.*



Europeans can be fit as a 3-way mixture of Loschbour, Stuttgart, and MA1

We inspected the statistics that precluded European populations from fitting both the Loschbour/Stuttgart (Fig. S12.8) and Motala12/Stuttgart (Fig. S12.9) models, and we noticed that these often involved either Karitiana or MA1. We plot the ratio of f_4*(X, Stuttgart; Karitiana, Chimp)* / f_4*(X, Stuttgart, MA1, Chimp)* in Fig. S12.10 for different European populations.

The related statistics f_4*(X, BedouinB; Karitiana, Chimp)* and f_4*(X, BedouinB; MA1, Chimp)* are plotted in Extended Data Fig. 5. By using BedouinB instead of Stuttgart, we can also plot Stuttgart in the space of these statistics. Europeans uniformly share more drift with MA1 than with Karitiana, and form a cline in this space with slope >1. Karitiana, because of its Ancient North Eurasian ancestry was crucial in detecting the presence of such ancestry in Europeans[2, 13] but can now be replaced in the study of this ancestry by a better proxy of this ancestry (MA1). We hope that in the future additional

representatives of this population may be studied, with either higher sequencing coverage or an even closer genetic relationship to the ANE population admixing into Europe.

Motivated by these observations, we modeled Europeans to be not only a mix of Stuttgart and one of the available ancient samples (Loschbour or Motala12), but also of a "Hunter" population whose amount of MA1-related ancestry was not fixed. Unlike Fig. S12.8 where zero MA1-related ancestry is assumed in Europeans, and Fig. S12.9 where "Hunter" is constrained to be a sister group of Scandinavian hunter-gatherers, we attempted to fit a model in which "Hunter" would only be constrained to be a mixture of Loschbour- and MA1-related ancestry. Fig. S12.11 shows the successful model structure, and Table S12.7 the inferred admixture proportions.

A total of 26 European populations fit this model, and we are encouraged by the fact that none of the Near Eastern populations fit, so the model correctly identified that they could not be derived as a mixture of these three ancestral populations (as they lack European hunter-gatherer that EEF have in part (SI 10) and WHG in full).

Table S12.7: Admixture proportions for West Eurasian populations that can be fit as a 3-way mixture of Early European Farmers (EEF), West European Hunter-Gatherers (WHG) and Ancient North Eurasians (ANE). (These proportions are also included in Extended Data Table 2).

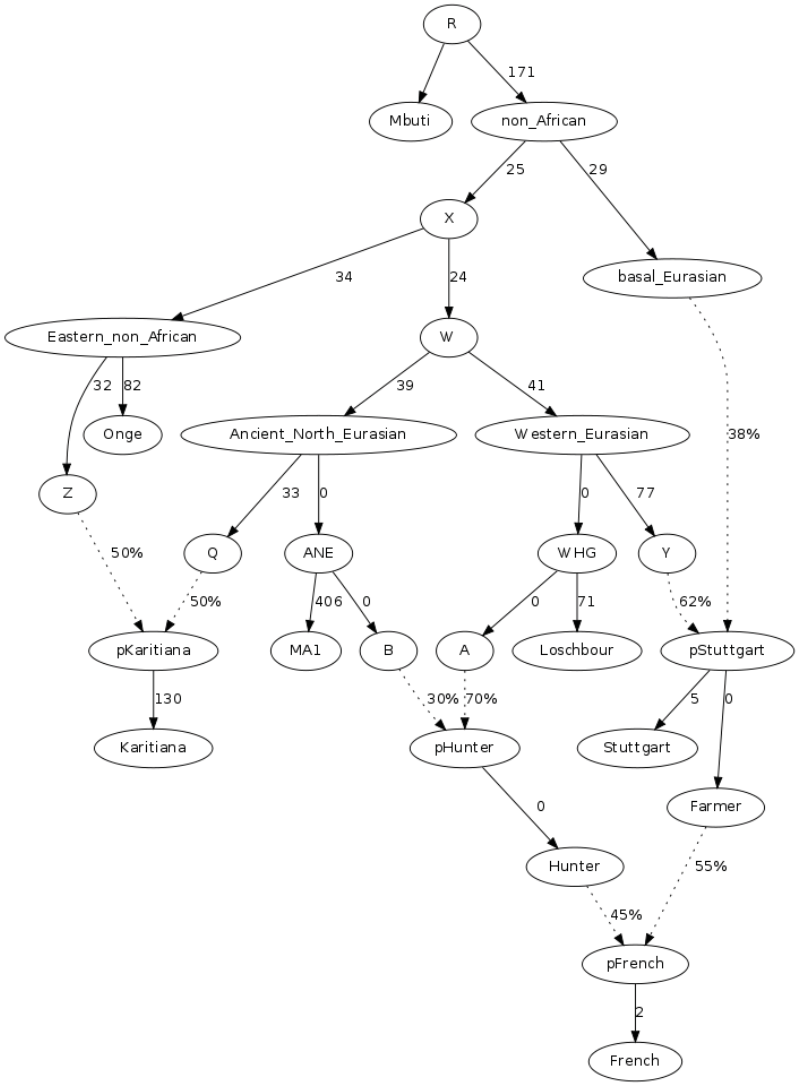
	EEF	WHG	ANE
Albanian	0.781	0.092	0.127
Ashkenazi_Jew	0.931	0.000	0.069
Basque	0.593	0.293	0.114
Belorussian	0.418	0.431	0.151
Bergamo	0.715	0.177	0.108
Bulgarian	0.712	0.147	0.141
Croatian	0.561	0.293	0.145
Czech	0.495	0.338	0.167
English	0.495	0.364	0.141
Estonian	0.322	0.495	0.183
French	0.554	0.311	0.135
French_South	0.675	0.195	0.130
Greek	0.792	0.058	0.151
Hungarian	0.558	0.264	0.179
Icelandic	0.394	0.456	0.150
Lithuanian	0.364	0.464	0.172
Maltese	0.932	0.000	0.068
Norwegian	0.411	0.428	0.161
Orcadian	0.457	0.385	0.158
Pais_Vasco	0.713	0.125	0.163
Sardinian	0.817	0.175	0.008
Scottish	0.390	0.428	0.182
Sicilian	0.903	0.000	0.097
Spanish	0.809	0.068	0.123
Tuscan	0.746	0.136	0.118
Ukrainian	0.462	0.387	0.151

It is evident that southern European populations have a greater affinity to early European farmers, and northern European populations have a greater affinity to Western European hunter gatherers, consistent with the analysis of a Swedish Funnelbeaker farmer[14] (Skoglund_farmer in Fig. 1B) who resembled southern Europeans, and two Iberian Mesolithic hunter-gatherers[15] (LaBrana1 and

LaBrana2 in Fig. 1B) who resembled Northern Europeans. Our analysis supports the view that ancestry from the two groups is variable across Europe, and suggests that a third element related to Upper Paleolithic Siberians, represented by MA1, also contributed to modern Europeans.

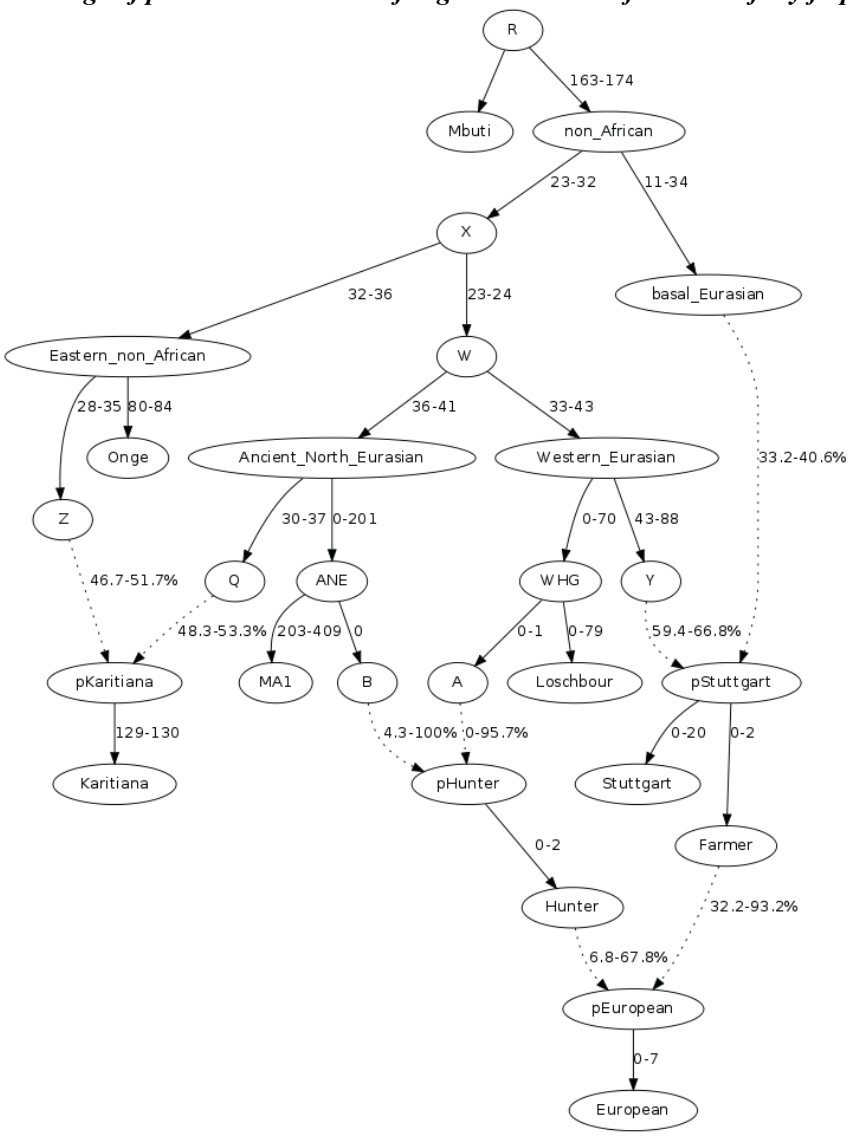
An interesting feature of these proportions is that they contrast the Basques to their Iberian neighbors, with nearly a third of their ancestry coming from WHG; this reflects the same genetic patterns as Fig. 1B which shows the Basques to the left off their Iberian neighbors, and European hunter gatherers projected in the same direction. Basques appear to possess a geographically local maximum of European hunter-gatherer ancestry.

Figure S12.11: A successful 3-way fit for French, a population that cannot fit as a 2-way mixture. *Estimated mixture proportions are 45/55% "Hunter"/"Farmer", or 55/31/14% EEF/WHG/ANE.*



The model fit in Fig. S12.11 is for the French population, but for each of the 26 successfully fit populations, the internal structure of the tree may be different. In Fig. S12.12 we present the range of parameter estimates. Some of these appear quite stable, achieving very similar values regardless of which individual population is fit, while others are less so, with the extreme being the amount of WHG ancestry in "Hunter", ranging from 0 to 95.7%. In that particular case, it was Ashkenazi Jews, Maltese and Sicilians for whom the value was 0, and Sardinians who had the highest 95.7% value.

Figure S12.12: Range of parameter estimates of Fig. S12.10 model for successfully fit populations

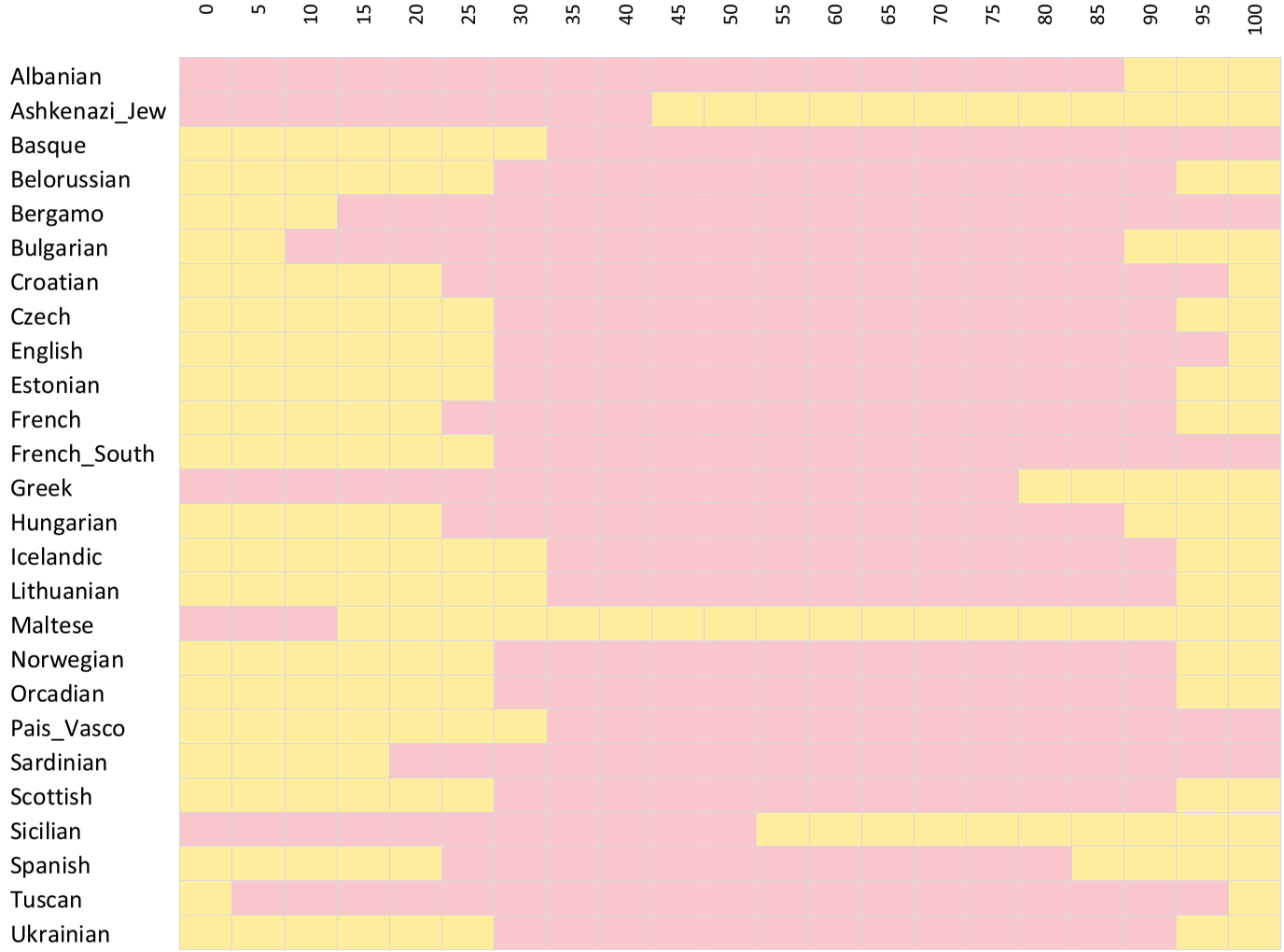


Most European populations have both WHG and ANE ancestry

We wanted to explicitly check which West Eurasian populations could be fit as a mixture of Stuttgart and a "Hunter" population with *x*% WHG and 100-*x*% ANE ancestry. To do this we fit the model of Figures S12.11 and S12.12 again for the populations of Table S12.7, but this time did not allow the proportions of WHG and ANE to vary freely but rather "locked" them in 5% increments, from (0, 100), (5,95), …, (100,0), thus exploring the whole range of possible mixtures for "Hunter".

Fig. S12.13 shows the range of values of *x* that were compatible with each population. While a wide range of possible values is consistent with each population, with the exception of some populations which are consistent with no WHG ancestry (Albanian, Ashkenazi_Jew, Greek, Maltese, Sicilian), and some others consistent with no ANE ancestry (Basque, French_South, Bergamo, Pais_Vasco, Sardinian), most Europeans can only be fit as having both WHG and ANE ancestry. Moreover, even in the case of many populations compatible with no WHG or no ANE ancestry, the best fit (Table S12.7) includes some such ancestry. For example, Basques are compatible with having no ANE ancestry, but according to Table S12.10, the best fit has 0.293 WHG and 0.114 ANE ancestry, for an *x* ratio of 72%, that is, an intermediate value within the range indicated in Fig. S12.1.

Figure S12.13: WHG/(WHG+ANE) ratio (%) feasible ranges (in pink for populations of Table S12.10. *Most European populations have both WHG and ANE ancestry, while a few are compatible with having no WHG ancestry (ratio 0%) and a few are compatible with having no ANE ancestry (ratio 100%).*



Pairs of European populations consistent with descent from the same "Farmers" and "Hunters"
Fig. S12.13 suggests that a large number of European populations can be successfully fit over a wide range of the WHG/(WHG+ANE) ratio. However, this does not necessarily indicate that they are descended from the same "Farmer" and "Hunter" populations, because the internal tree parameters inferred for two populations may differ.

A solution to this problem is to try and fit two European populations *A* and *B* simultaneously as two independent mixtures of "Farmer" and "Hunter". This has the advantage of forcing the tree to accommodate both *A* and *B*, and can thus determine whether a common tree can fit both. However, this simple modeling ignores the post-admixture histories of *A* and *B*, which may be complex and involve gene flow between them. It is unrealistic to model European populations as independent mixtures of "Farmer" and "Hunter" in the context of the major gene flows that must have occurred within Europe over the last few thousand years.

To address this problem, we modified *qpGraph*. As discussed in Patterson et al. (2012)[9] a basis for *f*-statistics involving populations $(A_0, A_1, \ldots, A_n)$ is found from $f_3(A_0; A_i, A_j)$, $f_2(A_0, A_i)$ $0 < i < j$. We think of A_0 as a base population. Suppose *A* and *B* are 2 populations whose descendants have a complex recent history such as two European populations descended from the "Farmer" and "Hunter", above. *qpGraph* calculates an empirical covariance matrix for the *f*-statistics involving the base point A_0. Our modification is simply to add a large constant (we chose 10,000) to the variance term for $f_3(A_0, A, B)$. This has the effect that qpGraph regards *f*-statistics involving both *A* and *B* as essentially uninformative, which has precisely the desired effect. This has the advantage of fitting a tree structure for both *A* and *B* simultaneously while avoiding the interactions between *A* and *B* that might reflect details of their more recent common history.

In Fig. S12.14 we show populations pairs that are consistent with descent from identical "Farmer" and "Hunter" populations.

Sicilians, Ashkenazi Jews, and Maltese are only compatible with each other and not with any other populations, consistent with Fig. S12.13 and Table S12.7 which show them to be have less or even no WHG ancestry in contrast to other populations.

Greeks are compatible with their geographical neighbors in the Balkans (Albanians and Bulgarians) and Italy (Bergamo and Tuscans).

Basques and Pais_Vasco are incompatible with several populations from Mediterranean and Southeastern Europe.

Mediterranean and Southeastern Europeans such as Spanish, Albanians, Bulgarians, Bergamo, Tuscans, Croatians, and Hungarians appear to be compatible with each other

Importantly, this analysis confirms that a large number of European populations are consistent with descent from identical "Farmer" and "Hunter" populations. Overall, 202 of the 325 possible pairs for the 26 populations resulted in graph fits with no outlier f_4-statistics. We conclude that a substantial number of modern European populations are consistent with having inherited ancestry from the three EEF/WHG/ANE groups via only two proximate ancestral populations.

In Fig. S12.15 we plot the WHG/(WHG+ANE) ratio over all 202 compatible pairs. It is clear that the bulk of the distribution is in the 60-80% interval, with a visible peak around 71-74%. This suggests that, for many Europeans, "Hunter" was a population of predominantly WHG-related ancestry but with a substantial ANE-related component.

Figure S12.14: Population pairs marked in pink are consistent with common descent from identical "Farmer" and "Hunter" populations.

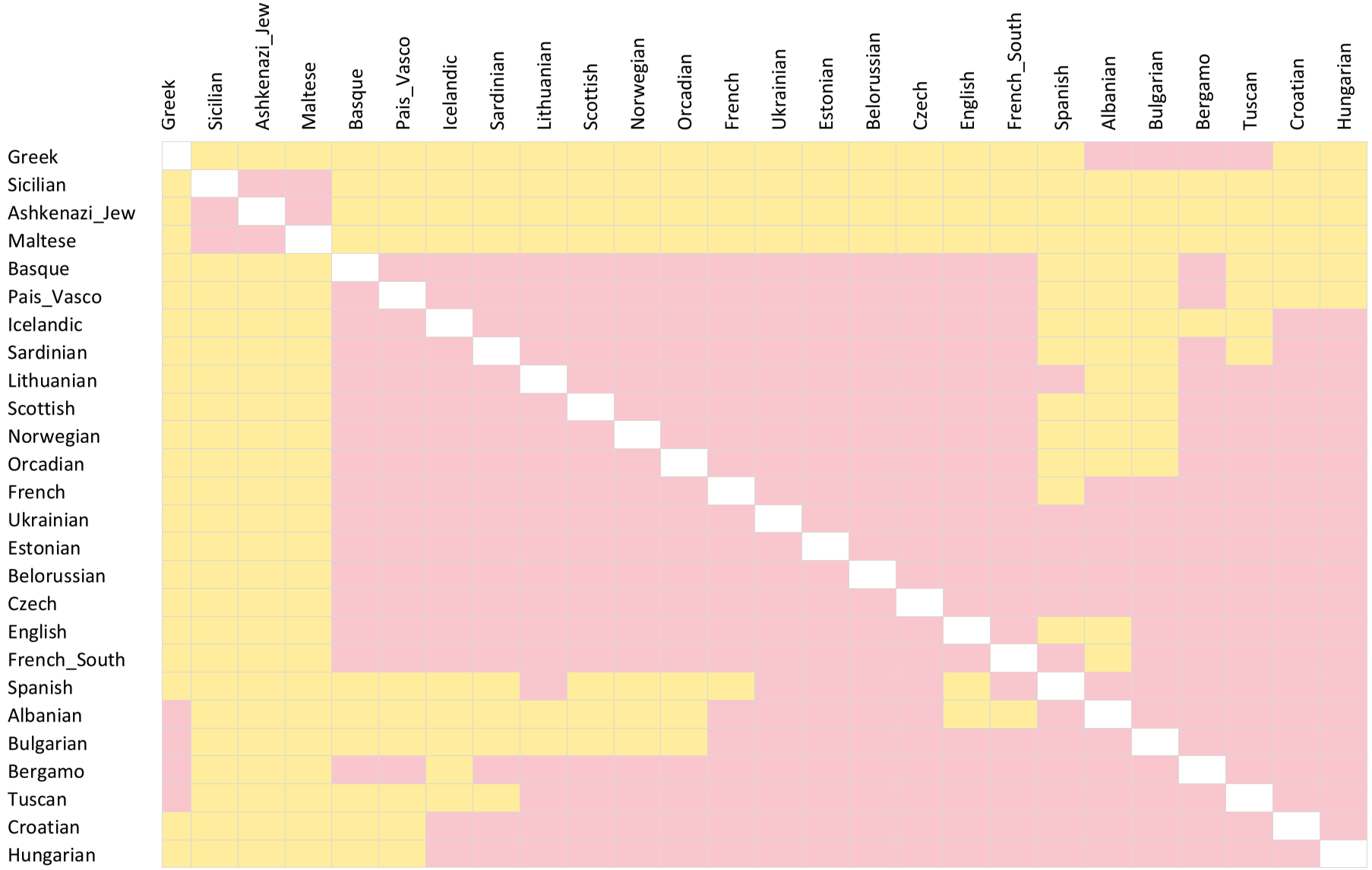
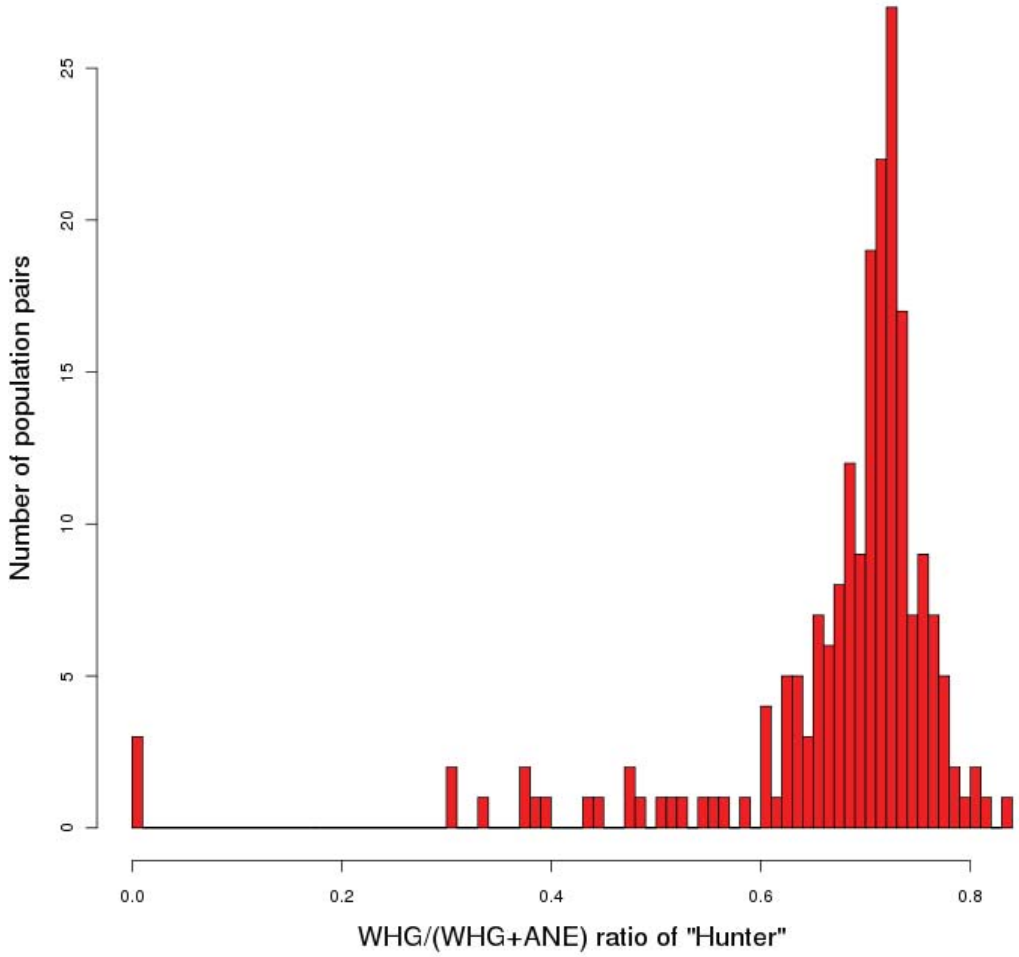


Figure S12.15: Distribution of WHG/(WHG+ANE) ratio for population pairs that can be successfully fit as descendants of identical "Farmer" and "Hunter" populations (Fig. S12.13).



Almost all pairs of European populations consistent with descent from the same ""EEF", "WHG", and "ANE" populations

We repeated the joint fitting of population pairs, but allowed each population in a pair to descend from a different "Hunter" population, i.e., with a variable WHG/(WHG+ANE) ratio. Almost all population pairs were now successful (264 of 325, Fig. S12.16), with the exception of Ashkenazi Jews, Maltese, and Sicilians who could often not be fit with other populations. It appears that these populations have Near Eastern ancestry that is not well-modeled by the 3-population model. This is consistent with their position in Fig. 1B, and the results of analysis of SI 13 which do not explicitly model deep population history.

We estimated averaged admixture proportions for 23 populations (excluding Ashkenazi Jews, Sicilians, and Maltese) who appear in Fig. S12.15 to be descended from identical EEF, WHG, and ANE populations. Whereas the proportions of Table S12.7 were derived from individual fits of the populations, those of Table S12.8 represent the average, for each population, over all compatible population pairs. The proportions of Table S12.8 are the ones plotted in Fig. 2B.

Figure S12.16: Population pairs marked in pink are consistent with common descent from identical "EEF", "WHG", and "ANE" populations.

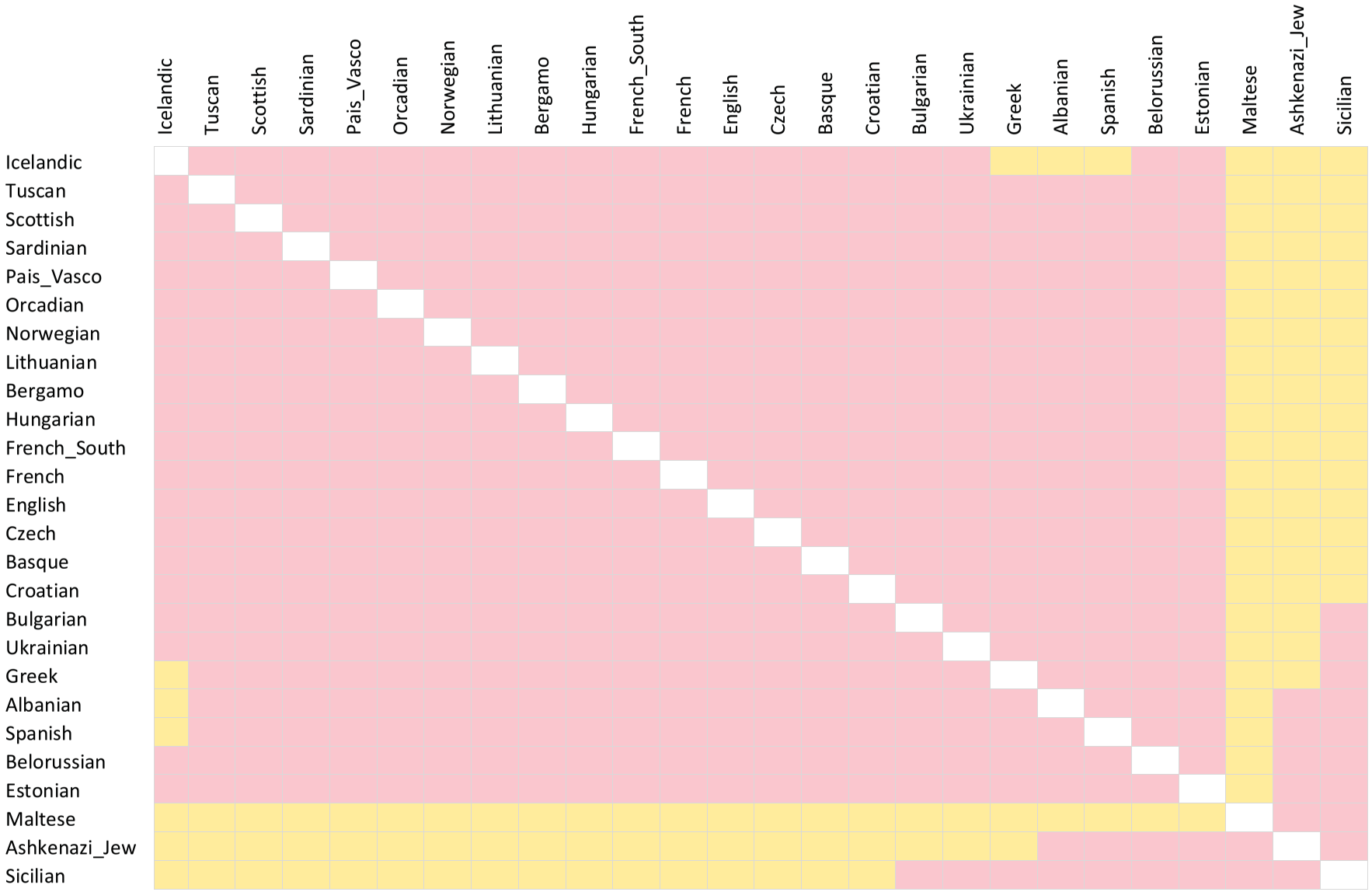


Table S12.8: Averaged admixture proportions for European populations. *Each proportion represents the mean over all fits with compatible populations; the range of the successful fits is also shown. (These proportions are also included with other mixture estimates in Extended Data Table 2).*

	EEF		WHG		ANE	
	Mean	Range	Mean	Range	Mean	Range
Albanian	0.781	0.772-0.819	0.082	0.032-0.098	0.137	0.129-0.158
Basque	0.569	0.527-0.616	0.335	0.255-0.392	0.096	0.076-0.129
Belorussian	0.426	0.397-0.464	0.408	0.338-0.443	0.167	0.150-0.199
Bergamo	0.721	0.704-0.793	0.163	0.061-0.189	0.117	0.104-0.147
Bulgarian	0.718	0.707-0.778	0.132	0.047-0.151	0.151	0.138-0.175
Croatian	0.564	0.548-0.586	0.285	0.242-0.310	0.151	0.137-0.172
Czech	0.489	0.460-0.531	0.348	0.273-0.382	0.163	0.145-0.196
English	0.503	0.476-0.536	0.353	0.296-0.382	0.144	0.130-0.169
Estonian	0.323	0.293-0.345	0.49	0.451-0.520	0.187	0.172-0.205
French	0.563	0.537-0.601	0.297	0.230-0.328	0.140	0.126-0.169
French_South	0.636	0.589-0.738	0.256	0.111-0.323	0.108	0.088-0.151
Greek	0.791	0.780-0.816	0.048	0.019-0.060	0.161	0.150-0.171
Hungarian	0.548	0.520-0.590	0.279	0.199-0.313	0.174	0.156-0.210
Icelandic	0.409	0.386-0.424	0.448	0.409-0.473	0.143	0.126-0.170
Lithuanian	0.352	0.327-0.384	0.488	0.433-0.527	0.160	0.135-0.184
Norwegian	0.417	0.388-0.438	0.423	0.383-0.450	0.160	0.140-0.181
Orcadian	0.465	0.439-0.493	0.378	0.329-0.403	0.157	0.140-0.179
Pais_Vasco	0.612	0.561-0.660	0.292	0.214-0.365	0.096	0.072-0.126
Sardinian	0.818	0.791-0.874	0.141	0.058-0.182	0.041	0.026-0.068
Scottish	0.408	0.387-0.424	0.421	0.384-0.448	0.171	0.149-0.201
Spanish	0.759	0.736-0.804	0.126	0.066-0.170	0.115	0.091-0.151
Tuscan	0.751	0.737-0.806	0.123	0.047-0.145	0.126	0.114-0.150
Ukrainian	0.463	0.445-0.491	0.376	0.322-0.399	0.160	0.148-0.187

f_4-ratio based estimation of Early European Farmer ancestry

The proportions of Table S12.8 are based on model fits using ADMIXTUREGRAPH, which simultaneously optimizes *f*-statistics over several populations. This may make the estimates more robust, but is also based on the accuracy of the entire model we fit. We also confirmed these estimates using a more direct method applied to the proposed graph.

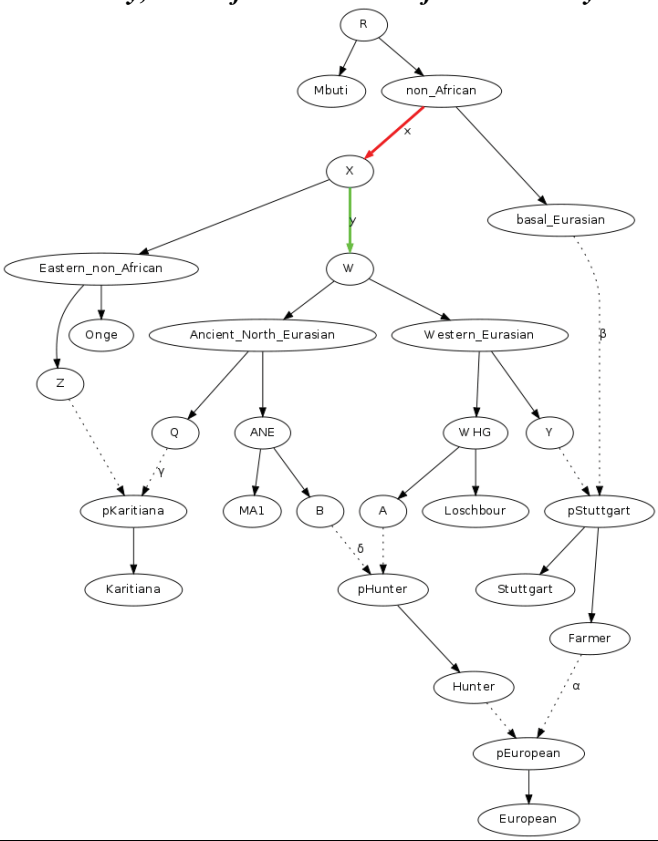
Consider Fig. S12.17. In this model which we have argued above is a fit to the data for many European populations to within the limits of our resolution, a European population has $\alpha\beta$ of its ancestry from Basal_Eurasian and Stuttgart has β of its ancestry from EEF. It is then the case that:

$$f_4(\text{Mbuti, Onge; Loschbour, European}) = -\alpha\beta x$$
$$f_4(\text{Mbuti, Onge; Loschbour, Stuttgart}) = -\beta x$$

This exploits the fact that the paths Mbuti→Onge and Loschbour→European or Loschbour→Stuttgart intersect only over the segment non_African→X whose drift length is x. We can then apply f_4-ratio estimation in a straightforward way by dividing the two[1, 2]. We show in Table S12.9 the estimates we obtain as well as their differences from those of Table S12.8.

The f_4-ratio estimates differ from those of ADMIXTUREGRAPH by no more than 1.3 standard errors. The mean and standard deviation over all populations is 0.047±0.506. Thus, an f_4-ratio estimation of this proportion over the proposed model is consistent with the optimization-based estimate.

Figure S12.17: The fact that a European population has αβ fraction of Basal Eurasian ancestry and Stuttgart has β such ancestry, allows for an estimate of EEF ancestry via an f_4-ratio.



f_4-ratio estimate of Basal Eurasian admixture in Stuttgart

A different parameter that can be estimated via an f_4-ratio is the amount of basal_Eurasian admixture into Stuttgart. Consider the edge $X \rightarrow W$ with drift length y in Fig. S12.17.

We can estimate y directly by the following statistic:

$$y = f_4(Mbuti, MA1; Onge, Loschbour) \quad \text{(S12.1)}$$

But also:

$$\beta y = f_4(Stuttgart, Loschbour; Onge\ MA1) \quad \text{(S12.2)}$$

Taking the ratio we estimate β=0.44 ± 0.10. The fitted values of β are within 1 standard error of this estimate (Fig. S12.12). This finding provides further support for the view that the hypothesized Basal Eurasian ancestry indeed had a major effect on ancient Near Eastern populations. The amount of Basal Eurasian admixture in the ancient Near East is uncertain, as the lack of an unadmixed Near Eastern reference makes the amount of Near Eastern admixture into Stuttgart uncertain (SI 10), but it must have been higher than the estimated value for Stuttgart.

East Eurasian gene flow into far Northeastern European populations

Three European populations failed to successfully fit the model of Fig. S12.11, and we list them in Table S12.10 together with the most significantly differing f-statistics.

These three far northeastern European populations share more genetic drift with Karitiana/Onge than is predicted by the model (both Onge and Karitiana-related statistics are violated for all three). This is consistent with the ADMIXTURE analysis (SI 9) which suggests that they possess a Siberian ancestral component not shared with other Europeans. It is also consistent with the results of Fig. S12.10 which show that these three populations share more drift with Karitiana relative to other Europeans. A possible explanation for this is distinct gene flow from Siberia, perhaps related to the migration of Y-haplogroup N from east Asia into west Eurasia[16, 17], as this lineage is present in the northeast and rare elsewhere in Europe.

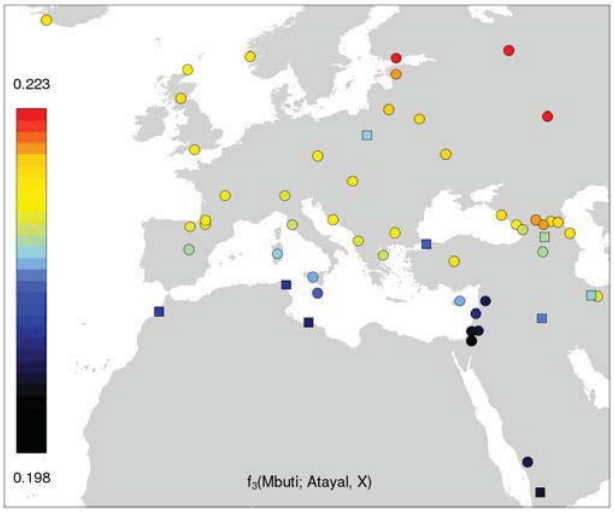
Table S12.10: European populations that cannot be fit as a 3-way mix of EEF, WHG, and ANE

Population	Violated statistic	fitted	estimated	Z
Finnish	Karitiana, MA1; Loschbour, Finnish	0.002025	-0.003984	-3.161
Mordovian	Karitiana, MA1; Loschbour, Mordovian	0.002050	-0.004990	-3.790
Russian	Karitiana, MA1; Loschbour, Russian	0.001947	-0.004214	-3.398

In Extended Data Fig. 6 we plot f_4*(Test, BedouinB; Han, Mbuti)* and f_4*(Test, BedouinB; MA1, Mbuti)* statistics; we use BedouinB so that we can also plot Stuttgart in the same figure. Populations that fit the model of Fig. S12.11 form a clear cline from Stuttgart in the south, to Lithuanians and Estonians in the north, but the three populations violating our model (Table S12.10) are clearly to the right, sharing relatively more common drift with the Han. We also add the single Saami individual from our dataset and the Chuvash on this plot, two additional European groups who deviate from the main European cline even more strongly in the same direction.

While we see no evidence that the Han have West Eurasian admixture (SI 9), it is still possible that they possess some unknown common component with Northeast Europeans. We also calculated the f_3(Mbuti; Atayal, Test) statistic for West Eurasian populations which measures the amount of common drift between Atayal (a Taiwanese aboriginal population that seems extremely unlikely to have historical connections with Northeastern Europeans in particular) and plot it in Figure S12.18. Northeastern Europeans share higher amounts of drift with Atayal as well, consistent with having an east Eurasian influence that is lower (or lacking) in other Europeans.

Figure S12.18: Northeast Europeans share more drift with Atayal than other Europeans



Note that the fact that Europe has higher values of this statistic than the Near East does not indicate East Eurasian admixture across Europe as this statistic is also reduced by the presence of either Basal Eurasian or African admixture.

Finally, we used ALDER[18] to investigate whether a linkage disequilibrium signal of recent admixture exists in Northeastern Europe (Table S12.11) using the Han as a reference and found a significant (Z>3) curve for three populations (for Finnish, Z=1.27, while we could not use this method on a single Saami individual).

Table S11.11: Chuvash, Mordovians & Russians have LD evidence for recent East Asian mixture.

	East Eurasian admixture (%)		Time	
	Lower bound (%)	std. error	Generations	std. error
Chuvash	11.7	1.7	62.5	11.0
Mordovian	6.7	1.3	69.8	16.9
Russian	5.7	0.4	52.3	5.8

The most straightforward explanation for these combined observations is that Northeastern Europeans possess some ancestry from an eastern Eurasian population, although more complicated explanations involving a population that affected both Northeastern Europe and eastern non-Africans are also possible. However, we think that the genetic landscape Siberia has changed since the time of MA1 (~24,000 years ago), as this would explain both the fact that present-day Siberians share less drift with MA1 than both Europeans and Native Americans[11] , that "First Americans" like the Karitiana already possessed east Eurasian admixture, and also that later waves of migration into the Americas share additional common drift with Han Chinese than the wave of "First Americans"[13]. Northeastern Europe may also have received genetic input from a later period of the Siberian gene pool in which (unlike the time of MA1), the eastern Eurasian influence was present. More ancient DNA research in both Northeastern Europe and Siberia may directly validate this proposal.

High levels of Ancient North Eurasian ancestry in the Northeast Caucasus

Finally, we turned to the Near East and Caucasus to explore the implications of our model for admixture events there. We note (Table 1, Extended Data Table 1) that Near Easterners all have their lowest f_3-statistics involving Stuttgart, consistent with the idea that Stuttgart possesses a substantial proportion of ancient Near Eastern ancestry. However, different populations appear to have their strongest signal of admixture involving pairings of Stuttgart with (i) Africans (Esan, Gambian, Kgagaladi), (ii) South Asians (Gujarati 3, Vishwabrahmin), (iii) Piapoco, a native American population, or (iv) MA1. Together with the evidence of Fig. 1B, this points to Near Eastern and Caucasian populations having a common ancestry related to Stuttgart, which is, however, modified by different influences related to many world populations. Unlike Europe, where several ancient DNA samples now exist, including the ones sequenced for our study, no ancient human genomes exist for the Near East, making reconstructions of its past even more difficult.

We intersected the set of Near Eastern populations without substantial (<1%) African admixture as inferred by ADMIXTURE K=10 (SI 9) with those whose most significant f_3-statistic involved the pairing (Stuttgart, MA1) (Table 1). Five populations met these criteria: Abkhasian, Chechen, Cypriot, Druze, Lezgin. We modified the model of Fig. S12.11 to model these populations as a mixture of a Near Eastern population that also contributed to Stuttgart and an MA1-related ANE population (but no WHG ancestry) (Fig. S12.19). All five populations fit successfully, and we report their admixture proportions in Table S12.12.

It is also possible to derive a direct lower bound of ANE ancestry from the model of Fig. S12.19 by the f_4-ratio $f_4(Test, Stuttgart; Karitiana, Onge) / f_4(MA1, Stuttgart; Karitiana, Onge)$.

Table S12.12: Admixture proportions for Near Eastern populations that can be fit as a mixture of Near East and Ancient North Eurasians. *A lower bound that can be obtained via the ratio f_4(Test, Stuttgart; Karitiana, Onge) / f_4(MA1, Stuttgart; Karitiana, Onge) is also indicated and appears only slightly lower than the fitted estimate.*

	Near East	ANE (fitted)	ANE (lower bound)
Abkhasian	0.814	0.186	0.157 ± 0.052
Chechen	0.730	0.270	0.244 ± 0.049
Cypriot	0.867	0.133	0.097 ± 0.056
Druze	0.882	0.118	0.047 ± 0.055
Lezgin	0.712	0.288	0.261 ± 0.049

For the denominator:

$$f_4(MA1, Stuttgart; Karitiana, Onge) = \beta(z+\alpha(1-\gamma)y) \quad \text{(S12.4)}$$

This expectation reflects the fact that the paths MA1→Stuttgart and Karitiana→Onge overlap only when Karitiana descends from Ancient_North_Eurasian (fraction β). The segment with length z is always traversed by both paths in that case, but the segment with length y is only traversed when Stuttgart has Basal_Eurasian ancestry (fraction $\alpha(1-\gamma)$).

For the numerator:

$$f_4(Test, Stuttgart; Karitiana, Onge) = \beta(z+\alpha(1-\gamma)y)(1-\delta) - \beta\delta\gamma\alpha y \quad \text{(S12.5)}$$

The first term in (2) is the same as in (1) multiplied by 1-δ, since 1-δ fraction of the Test population's ancestry descends from ANE. For the portion of Test's ancestry δ that comes from Near East, the path Test→Stuttgart does not overlap with Karitiana→Onge except in the case that Stuttgart descends from UHG (γ fraction) and the Test population descends from Basal Eurasian (δa); in all other cases, the path Test→Stuttgart only passes through the Western_Eurasian subtree and is uncorrelated to the Karitiana→Onge one. By dividing (S12.4) by (S12.3) we thus obtain f_4(Test, Stuttgart; Karitiana, Onge) / f_4(MA1, Stuttgart; Karitiana, Onge) = $1-\delta-\delta\gamma\alpha y/(z+\alpha(1-\gamma)y) \leq 1-\delta$. The lower bound obtained for these five populations is also shown in Table S12.12.

An interesting implication of this analysis is that ANE-related ancestry may be particularly high in the Northeast Caucasus, as both fitted and lower bound values for Lezgins and Chechens exceed inferred ANE values for Europeans (compare Table S12.8 and Table S12.12). The high affinity of the Northeast Caucasus to MA1 is also demonstrated in Extended Data Fig. 7 where the statistic f_4(Test, Chimp; MA1, Loschbour) exhibits highest values in the region. In light of our other results, it is not surprising that these populations would have high ANE-related ancestry. They are at the northern end of the Near Eastern cline (Fig. 1B) and have the highest values of common drift with MA1 among Near Eastern populations (Extended Data Fig. 4), as measured by f_4(Test, Stuttgart; MA1, Chimp). However, the high MA1-related admixture in Northeast Caucasians seemingly contradicts Extended Data Fig. 4 which shows many Europeans to have even higher values of the statistic.

This is not in fact a contradiction, however, because for Europeans the statistic can be written as:

$$\begin{aligned} f_4(European, Stuttgart; MA1, Mbuti) = {} & \alpha_{EEF} f_4(Stuttgart, Stuttgart; MA1, Mbuti) \\ & + \alpha_{WHG} f_4(Loschbour, Stuttgart; MA1, Mbuti) \\ & + \alpha_{ANE} f_4(B, Stuttgart; MA1, Mbuti) \end{aligned} \quad \text{(S12.6)}$$

The first term vanishes, and both other terms are positive, since B and MA1 are sister clades and Loschbour and MA1 share drift that Stuttgart lacks because of its basal Eurasian admixture, with f_4(Loschbour, Stuttgart; MA1, Mbuti) = 0.004573 (Z= 6.799).

By contrast for North Caucasians:

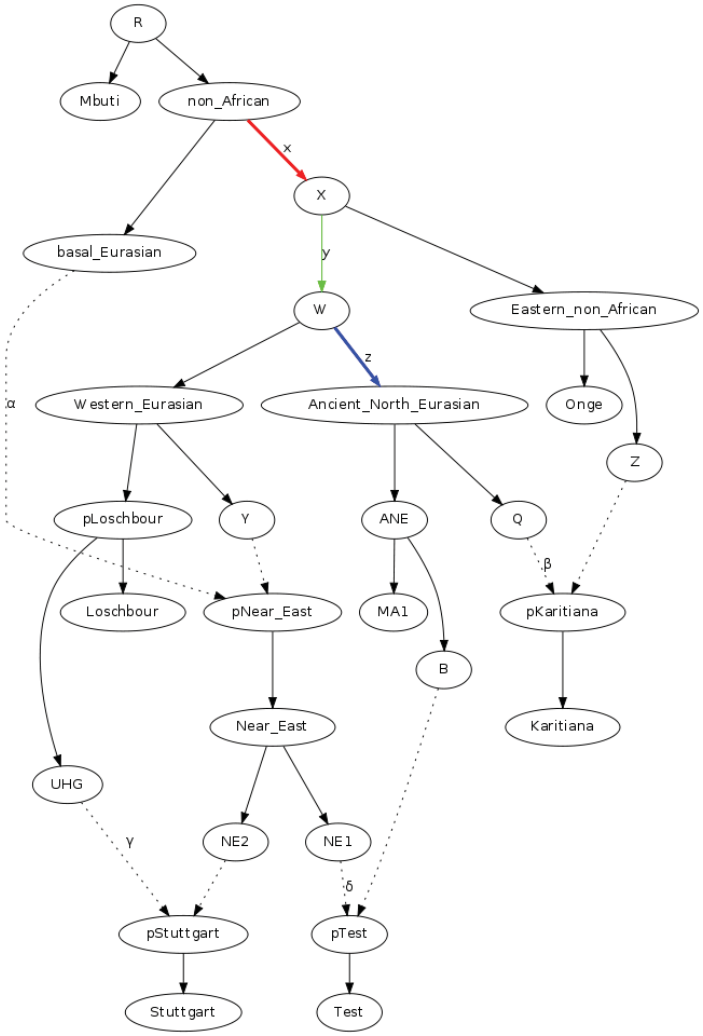
$$f_4(\text{North Caucasian, Stuttgart; MA1, Mbuti}) = \alpha_{ANE} f_4(\text{B, Stuttgart; MA1, Mbuti}) + \alpha_{Near\_East} f_4(\text{NE1, Stuttgart; MA1, Mbuti}) \quad \text{(S12.7)}$$

The second term is negative, because $f_4(\text{NE1, Stuttgart; MA1, Mbuti}) = -\alpha\gamma(x+y)$.

Intuitively, the shared drift shared between a test population and MA1 is diluted by Near Eastern ancestry (because of the Basal Eurasian ancestry in the Near East), and augmented by WHG ancestry (because of the lack of Basal Eurasian ancestry in Loschbour).

We have conveniently labeled MA1-related ancestry "Ancient North Eurasian" because of the provenance of MA1 in Siberia, but at present we cannot be sure whether this type of ancestry originated there or was a recent migrant from some western region.

Figure S12.19: A model for Near Eastern populations with Ancient North Eurasian admixture. Stuttgart is a mixture of Near_East and a sister group of Loschbour (UHG: Unknown Hunter-Gatherers); A Test population (shown here) is a mixture of Near_East and a sister group of MA1.



Conversely, we do not currently know whether the signal of admixture observed in the Near East and Caucasus reflects an arrival of MA1-related ancestry from the east, or alternatively dilution of native MA1-related ancestry by an expansion of a Near Eastern population carrying Basal Eurasian

admixture, associated perhaps with the expansion of Levantine/Mesopotamian early agriculturalists who seem to have influenced the Y-chromosome distribution of the region[19]. Future studies of ancient Central Eurasians may help resolve such questions of migration timing and directionality.

Concluding Remarks

We chose to model the 3-way admixture as taking place in the order (Early European Farmers, (West European Hunter Gatherers, Ancient North Eurasians)), but we should caution that the order is unknown and may become apparent as later samples from Europe and elsewhere provide ancient DNA for study. Different combinations of the three ancestral populations may have contributed to the formation of modern Europeans. Nonetheless, our co-fitting of population pairs (Fig. S12.14 and Fig. S12.15) reveals that the WHG/(WHG+ANE) ratio is fairly narrowly constrained over many European populations, so the chosen order seems reasonable. In addition, the consistency of the estimates with those from SI 13 which do not require a branching order gives further confidence regarding our estimates of ancestry proportion.

A geographically parsimonious hypothesis would be that a major component of present-day European ancestry was formed in eastern Europe or western Siberia where western and eastern hunter-gatherer groups could plausibly have intermixed. Motala12 has an estimated WHG/(WHG+ANE) ratio of 81% (S12.7), higher than that estimated for the population contributing to modern Europeans (Fig. S12.14). Motala and Mal'ta are separated by 5,000km in space and about 17 thousand years in time, leaving ample room for a genetically intermediate population. The lack of WHG ancestry in the Near East (Extended Data Fig. 6, Fig. 1B) together with the presence of ANE ancestry there (Table S12.12) suggests that the population who contributed ANE ancestry there may have lacked substantial amounts of WHG ancestry, and thus have a much lower (or even zero) WHG/(WHG+ANE) ratio.

It is also important to remember that the amount of WHG ancestry indicated in Tables S12.7 and S12.8 is not the total amount of European hunter-gatherer present in these populations, since Early European Farmers also possessed some such ancestry (SI 10). Conversely, we assumed that "Hunter" was composed only of WHG/ANE ancestry, but it is possible that the actual population that admixed with EEF may have already possessed EEF ancestry itself. Our results point to three major ancestral components for most modern Europeans, with many Europeans appearing as a simpler mixture of two components (Fig. S12.14 and Fig. S12.15), but, in the absence of ancient DNA from later periods of European history we cannot determine whether this process of admixture was simple and corresponds to an archaeologically visible event, or was more protracted over time. The fact that late Neolithic farmers still resembled Stuttgart (Fig. 1B) and Early Bronze Age Europeans resembled modern Europeans, at least mitochondrially[20], suggests the hypothesis that at least part of the admixture occurred over a relatively short period of time.

Some of our modeling is surely too simplistic and will need to be modified in some respects as newer ancient DNA samples become available and make it possible to constrain the model even further. Nevertheless, we are encouraged about the robustness of some of our results by the fact that admixture estimates presented in SI 13 that do not require modeling of deep history tend to agree with the ones derived here under an explicit model.

In the spirit of parsimony we chose to limit the number of admixture edges to 2 for the main model (Fig. S12.6), as a model with only as many edges could fit the ancient samples, and modern European populations could be accommodated easily in this scaffold (Fig. S12.11 and Fig. 2A).

More complex models with 3 or more admixture events could be devised, but cannot be constrained fully by our data as the number of ancient genomes is still small and limited in space and time, with crucial periods and places missing. The study of archaic humans has revealed an ever-increasing complexity of admixture and unexpected links across time and space[21-25], and as more ancient DNA samples became available, and it is likely that the story of our more immediate prehistoric ancestors will be shown to be even more complex.

References

1 David Reich, Kumarasamy Thangaraj, Nick Patterson, Alkes L. Price, and Lalji Singh, 'Reconstructing Indian Population History', *Nature,* 461 (2009), 489-94.

2 N. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, and D. Reich, 'Ancient Admixture in Human History', *Genetics,* 192 (2012), 1065-93.

3 Kumarasamy Thangaraj, Gyaneshwer Chaubey, Toomas Kivisild, Alla G. Reddy, Vijay Kumar Singh, Avinash A. Rasalkar, and Lalji Singh, 'Reconstructing the Origin of Andaman Islanders', *Science,* 308 (2005), 996-96.

4 David Reich, Nick Patterson, Martin Kircher, Frederick Delfin, Madhusudan R Nandineni, Irina Pugach, Albert Min-Shan Ko, Ying-Chin Ko, Timothy A Jinam, Maude E Phipps, Naruya Saitou, Andreas Wollstein, Manfred Kayser, Svante Pääbo, and Mark Stoneking, 'Denisova Admixture and the First Modern Human Dispersals into Southeast Asia and Oceania', *American journal of human genetics,* 89 (2011), 516-28.

5 Maanasa Raghavan, Pontus Skoglund, Kelly E. Graf, Mait Metspalu, Anders Albrechtsen, Ida Moltke, Simon Rasmussen, Thomas W. Stafford Jr, Ludovic Orlando, Ene Metspalu, Monika Karmin, Kristiina Tambets, Siiri Rootsi, Reedik Magi, Paula F. Campos, Elena Balanovska, Oleg Balanovsky, Elza Khusnutdinova, Sergey Litvinov, Ludmila P. Osipova, Sardana A. Fedorova, Mikhail I. Voevoda, Michael DeGiorgio, Thomas Sicheritz-Ponten, Soren Brunak, Svetlana Demeshchenko, Toomas Kivisild, Richard Villems, Rasmus Nielsen, Mattias Jakobsson, and Eske Willerslev, 'Upper Palaeolithic Siberian Genome Reveals Dual Ancestry of Native Americans', *Nature,* advance online publication (2013).

6 D. Reich, N. Patterson, D. Campbell, A. Tandon, S. Mazieres, N. Ray, M. V. Parra, W. Rojas, C. Duque, N. Mesa, L. F. Garcia, O. Triana, S. Blair, A. Maestre, J. C. Dib, C. M. Bravi, G. Bailliet, D. Corach, T. Hunemeier, M. C. Bortolini, F. M. Salzano, M. L. Petzl-Erler, V. Acuna-Alonzo, C. Aguilar-Salinas, S. Canizales-Quinteros, T. Tusie-Luna, L. Riba, M. Rodriguez-Cruz, M. Lopez-Alarcon, R. Coral-Vazquez, T. Canto-Cetina, I. Silva-Zolezzi, J. C. Fernandez-Lopez, A. V. Contreras, G. Jimenez-Sanchez, M. J. Gomez-Vazquez, J. Molina, A. Carracedo, A. Salas, C. Gallo, G. Poletti, D. B. Witonsky, G. Alkorta-Aranburu, R. I. Sukernik, L. Osipova, S. A. Fedorova, R. Vasquez, M. Villena, C. Moreau, R. Barrantes, D. Pauls, L. Excoffier, G. Bedoya, F. Rothhammer, J. M. Dugoujon, G. Larrouy, W. Klitz, D. Labuda, J. Kidd, K. Kidd, A. Di Rienzo, N. B. Freimer, A. L. Price, and A. Ruiz-Linares, 'Reconstructing Native American Population History', *Nature,* 488 (2012), 370-4.

7 Ofer Bar-Yosef, *The Chronology of the Middle Paleolithic of the Levant*. ed. by K. Aoki T. Akazawa, and O. Bar-Yosef, *Neandertals and Modern Humans in Western Asia* (New York: Plenum Press, 1998).

8 J. I. Rose, V. I. Usik, A. E. Marks, Y. H. Hilbert, C. S. Galletti, A. Parton, J. M. Geiling, V. Cerny, M. W. Morley, and R. G. Roberts, 'The Nubian Complex of Dhofar, Oman: An African Middle Stone Age Industry in Southern Arabia', *PLoS One,* 6 (2011), e28239.

9 Ornella Semino, Giuseppe Passarino, † Peter J. Oefner, Alice A. Lin, Svetlana Arbuzova, Lars E. Beckman, Giovanna De Benedictis, Paolo Francalacci, Anastasia Kouvatsi, Svetlana Limborska, Mladen Marcikiæ, Anna Mika, Barbara Mika, Dragan Primorac, A. Silvana Santachiara-Benerecetti, L. Luca Cavalli-Sforza, and Peter A. Underhill, 'The Genetic Legacy of Paleolithic Homo Sapiens Sapiens in Extant Europeans: A Y Chromosome Perspective', *Science,* 290 (2000), 1155-59.

10 B. Trombetta, F. Cruciani, D. Sellitto, and R. Scozzari, 'A New Topology of the Human Y Chromosome Haplogroup E1b1 (E-P2) Revealed through the Use of Newly Characterized Binary Polymorphisms', *PLoS One,* 6 (2011), e16073.

11 Marie Lacan, Christine Keyser, François-Xavier Ricaut, Nicolas Brucato, Josep Tarrús, Angel Bosch, Jean Guilaine, Eric Crubézy, and Bertrand Ludes, 'Ancient DNA Suggests the Leading Role Played by Men in the Neolithic Dissemination', *Proceedings of the National Academy of Sciences* (2011).

12 Morten Rasmussen, Xiaosen Guo, Yong Wang, Kirk E. Lohmueller, Simon Rasmussen, Anders Albrechtsen, Line Skotte, Stinus Lindgreen, Mait Metspalu, Thibaut Jombart, Toomas Kivisild, Weiwei Zhai, Anders Eriksson, Andrea Manica, Ludovic Orlando, Francisco M. De La Vega, Silvana Tridico, Ene Metspalu, Kasper Nielsen, María C. Ávila-Arcos, J. Víctor Moreno-Mayar, Craig Muller, Joe Dortch, M. Thomas P. Gilbert, Ole Lund, Agata Wesolowska, Monika Karmin, Lucy A. Weinert, Bo Wang, Jun Li, Shuaishuai Tai, Fei Xiao, Tsunehiko Hanihara, George van Driem, Aashish R. Jha, François-Xavier Ricaut, Peter de Knijff, Andrea B. Migliano, Irene Gallego Romero, Karsten Kristiansen, David M. Lambert, Søren Brunak, Peter Forster, Bernd Brinkmann, Olaf Nehlich, Michael Bunce, Michael Richards, Ramneek Gupta, Carlos D. Bustamante, Anders Krogh, Robert A. Foley, Marta M. Lahr, Francois Balloux, Thomas Sicheritz-Pontén, Richard Villems, Rasmus Nielsen, Jun Wang, and Eske Willerslev, 'An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia', *Science,* 334 (2011), 94-98.

13 M. Lipson, P. R. Loh, A. Levin, D. Reich, N. Patterson, and B. Berger, 'Efficient Moment-Based Inference of Admixture Parameters and Sources of Gene Flow', *Mol Biol Evol,* 30 (2013), 1788-802.

14 Pontus Skoglund, Helena Malmström, Maanasa Raghavan, Jan Storå, Per Hall, Eske Willerslev, M. Thomas P. Gilbert, Anders Götherström, and Mattias Jakobsson, 'Origins and Genetic Legacy of Neolithic Farmers and Hunter-Gatherers in Europe', *Science,* 336 (2012), 466-69.

15 Federico Sánchez-Quinto, Hannes Schroeder, Oscar Ramirez, María C Ávila-Arcos, Marc Pybus, Iñigo Olalde, Amhed M V. Velazquez, María Encina Prada Marcos, Julio Manuel Vidal Encinas, Jaume Bertranpetit, Ludovic Orlando, M. Thomas P Gilbert, and Carles Lalueza-Fox, 'Genomic Affinities of Two 7,000-Year-Old Iberian Hunter-Gatherers', *Current biology : CB,* 22 (2012), 1494-99.

16 Siiri Rootsi, Lev A. Zhivotovsky, Marian Baldovic, Manfred Kayser, Ildus A. Kutuev, Rita Khusainova, Marina A. Bermisheva, Marina Gubina, Sardana A. Fedorova, Anne-Mai Ilumae, Elza K. Khusnutdinova, Mikhail I. Voevoda, Ludmila P. Osipova, Mark Stoneking, Alice A. Lin, Vladimir Ferak, Juri Parik, Toomas Kivisild, Peter A. Underhill, and Richard Villems, 'A Counter-Clockwise Northern Route of the Y-Chromosome Haplogroup N from Southeast Asia Towards Europe', *Eur J Hum Genet,* 15 (2006), 204-11.

17 H. Shi, X. Qi, H. Zhong, Y. Peng, X. Zhang, R. Z. Ma, and B. Su, 'Genetic Evidence of an East Asian Origin and Paleolithic Northward Migration of Y-Chromosome Haplogroup N', *PLoS One,* 8 (2013), e66102.

18 Po-Ru Loh, Mark Lipson, Nick Patterson, Priya Moorjani, Joseph K. Pickrell, David Reich, and Bonnie Berger, 'Inferring Admixture Histories of Human Populations Using Linkage Disequilibrium', *Genetics,* 193 (2013), 1233-54.

19 Bayazit Yunusbayev, Mait Metspalu, Mari Järve, Ildus Kutuev, Siiri Rootsi, Ene Metspalu, Doron M. Behar, Kärt Varendi, Hovhannes Sahakyan, Rita Khusainova, Levon Yepiskoposyan, Elza K. Khusnutdinova, Peter A. Underhill, Toomas Kivisild, and Richard Villems, 'The Caucasus as an Asymmetric Semipermeable Barrier to Ancient Human Migrations', *Molecular Biology and Evolution* (2011).

20 Guido Brandt, Wolfgang Haak, Christina J. Adler, Christina Roth, Anna Szécsényi-Nagy, Sarah Karimnia, Sabine Möller-Rieker, Harald Meller, Robert Ganslmeier, Susanne Friederich, Veit Dresely, Nicole Nicklisch, Joseph K. Pickrell, Frank Sirocko, David Reich, Alan Cooper, Kurt W. Alt, and Consortium The Genographic, 'Ancient DNA Reveals Key Stages in the Formation of Central European Mitochondrial Genetic Diversity', *Science,* 342 (2013), 257-61.

21 Richard E. Green, Johannes Krause, Adrian W. Briggs, Tomislav Maricic, Udo Stenzel, Martin Kircher, Nick Patterson, Heng Li, Weiwei Zhai, Markus Hsi-Yang Fritz, Nancy F. Hansen, Eric Y. Durand, Anna-Sapfo Malaspinas, Jeffrey D. Jensen, Tomas Marques-Bonet, Can Alkan, Kay Prüfer, Matthias Meyer, Hernán A. Burbano, Jeffrey M. Good, Rigo Schultz, Ayinuer Aximu-

Petri, Anne Butthof, Barbara Höber, Barbara Höffner, Madlen Siegemund, Antje Weihmann, Chad Nusbaum, Eric S. Lander, Carsten Russ, Nathaniel Novod, Jason Affourtit, Michael Egholm, Christine Verna, Pavao Rudan, Dejana Brajkovic, Željko Kucan, Ivan Gušic, Vladimir B. Doronichev, Liubov V. Golovanova, Carles Lalueza-Fox, Marco de la Rasilla, Javier Fortea, Antonio Rosas, Ralf W. Schmitz, Philip L. F. Johnson, Evan E. Eichler, Daniel Falush, Ewan Birney, James C. Mullikin, Montgomery Slatkin, Rasmus Nielsen, Janet Kelso, Michael Lachmann, David Reich, and Svante Pääbo, 'A Draft Sequence of the Neandertal Genome', *Science,* 328 (2010), 710-22.

22 David Reich, Richard E. Green, Martin Kircher, Johannes Krause, Nick Patterson, Eric Y. Durand, Bence Viola, Adrian W. Briggs, Udo Stenzel, Philip L. F. Johnson, Tomislav Maricic, Jeffrey M. Good, Tomas Marques-Bonet, Can Alkan, Qiaomei Fu, Swapan Mallick, Heng Li, Matthias Meyer, Evan E. Eichler, Mark Stoneking, Michael Richards, Sahra Talamo, Michael V. Shunkov, Anatoli P. Derevianko, Jean-Jacques Hublin, Janet Kelso, Montgomery Slatkin, and Svante Paabo, 'Genetic History of an Archaic Hominin Group from Denisova Cave in Siberia', *Nature,* 468 (2010), 1053-60.

23 Matthias Meyer, Martin Kircher, Marie-Theres Gansauge, Heng Li, Fernando Racimo, Swapan Mallick, Joshua G. Schraiber, Flora Jay, Kay Prüfer, Cesare de Filippo, Peter H. Sudmant, Can Alkan, Qiaomei Fu, Ron Do, Nadin Rohland, Arti Tandon, Michael Siebauer, Richard E. Green, Katarzyna Bryc, Adrian W. Briggs, Udo Stenzel, Jesse Dabney, Jay Shendure, Jacob Kitzman, Michael F. Hammer, Michael V. Shunkov, Anatoli P. Derevianko, Nick Patterson, Aida M. Andrés, Evan E. Eichler, Montgomery Slatkin, David Reich, Janet Kelso, and Svante Pääbo, 'A High-Coverage Genome Sequence from an Archaic Denisovan Individual', *Science,* 338 (2012), 222-26.

24 Matthias Meyer, Qiaomei Fu, Ayinuer Aximu-Petri, Isabelle Glocke, Birgit Nickel, Juan-Luis Arsuaga, Ignacio Martinez, Ana Gracia, Jose Maria Bermudez de Castro, Eudald Carbonell, and Svante Paabo, 'A Mitochondrial Genome Sequence of a Hominin from Sima De Los Huesos', *Nature,* advance online publication (2013).

25 Kay Prufer, Fernando Racimo, Nick Patterson, Flora Jay, Sriram Sankararaman, Susanna Sawyer, Anja Heinze, Gabriel Renaud, Peter H. Sudmant, Cesare de Filippo, Heng Li, Swapan Mallick, Michael Dannemann, Qiaomei Fu, Martin Kircher, Martin Kuhlwilm, Michael Lachmann, Matthias Meyer, Matthias Ongyerth, Michael Siebauer, Christoph Theunert, Arti Tandon, Priya Moorjani, Joseph Pickrell, James C. Mullikin, Samuel H. Vohr, Richard E. Green, Ines Hellmann, Philip L. F. Johnson, Helene Blanche, Howard Cann, Jacob O. Kitzman, Jay Shendure, Evan E. Eichler, Ed S. Lein, Trygve E. Bakken, Liubov V. Golovanova, Vladimir B. Doronichev, Michael V. Shunkov, Anatoli P. Derevianko, Bence Viola, Montgomery Slatkin, David Reich, Janet Kelso, and Svante Paabo, 'The Complete Genome Sequence of a Neanderthal from the Altai Mountains', *Nature,* advance online publication (2013).

Supplementary Information 13

Admixture estimates that do not require phylogenetic modeling

Iosif Lazaridis*, Nick Patterson and David Reich

* To whom correspondence should be addressed (lazaridis@genetics.med.harvard.edu)

In SI12 we identify a plausible model of the relationships of deeply diverged non-African populations that does not contradict the data to within the limits of our resolution, and then used this model to derive admixture proportions. One consequence of our modeling is to show that a range of puzzling observations can be reconciled with the evidence if one postulates at least one "ghost" population ("Basal Eurasians") contributing to present-day West Eurasian populations. In SI12 we also show that another such "ghost" population ("Ancient North Eurasians") could be reconciled with the recently published Paleolithic MA1 sample from Siberia[1].

In this section we estimate mixture proportions for European populations in way that does not require making assumptions about the deep phylogenetic relationships among non-African populations. One advantage of this is that it avoids errors that might arise due to forcing a set of populations into an explicit model. A second advantage is that it can be applied over a large number of world populations without precisely modeling events taking place outside West Eurasia.

We first estimate admixture proportions of European populations in terms of the two prehistoric Europeans (Loschbour and Stuttgart). Loschbour-related admixture appears to be general across Europe, on the basis of (i) the intermediate position of Europeans between Loschbour and the Near East (Fig. 1B), (ii) the fact that population pairs of the form (X=Loschbour, Y=Near East) often produce the lowest f_3*(European; X, Y)* statistics (Table 1, Extended Data Table 1), and (iii) the fact that Europeans have a positive f_4*(European, Stuttgart; Loschbour, Chimp)* statistic (Extended Data Fig. 4). Stuttgart-related admixture is a reasonable starting hypothesis because of (i) the geographical importance of the Linearbandkeramik as the first food producing culture in large parts of continental Europe, (ii) mtDNA evidence suggesting substantial persistence of early farmer lineages in modern Europeans[2], (iii) the fact that many Europeans have very negative f_3*(European; Stuttgart, MA1)* statistics (Table 1, Extended Data Table 1), (iv) the existence of Stuttgart/Sardinian-like individuals from a wide geographical range in Europe and from different times,[3, 4] and (v) the existence of the "European cline" in Fig. 1B which strongly suggests that many European populations were formed by admixture of a Stuttgart/Sardinian-like population and an unknown element presently mostly concentrated in northern Europe.

Our approach (Fig. S13.1) is to study statistics of the form f_4*(European, Stuttgart;* O_1*,* O_2*)* where O_1, O_2 are two non-West Eurasian populations from a set of 13 populations without any evidence of recent European admixture (SI 9). This assumption is necessary because this statistic can be interpreted[5] as the drift path overlap between *European*→*Stuttgart* and O_1→ O_2. If, say, O_1 has recent admixture from a French source, then the value of the statistic will be higher when *European=French* than when *European=Russian*, because of the additional common drift shared with the French, and not because the French and the Russians are differentially related to the non-recently mixed portion of O_1. A similar problem arises if a test European population has recent admixture from the outgroups. For example, recent Native American admixture ancestry will result in the statistic's value not only being affected by the relationship of the constituent elements to Native Americans, but also by the substantial common drift that ensued in the Americas down to the present.

In Extended Data Fig. 4, we plot the statistics f_4*(West Eurasian, Stuttgart; MA1, Chimp)* vs. f_4*(West Eurasian, Stuttgart; Loschbour, Chimp)*. Both Near Eastern and European populations are often positive for the first statistic (suggesting MA1-related gene flow in both Europe and the Near East), but only Europeans are positive for both, consistent with the hypothesis that Europeans have pre-

Neolithic hunter-gatherer related ancestry. Europeans form a cline of increasing common drift with both Loschbour and MA1, so we will derive them as a mixture of the following elements:

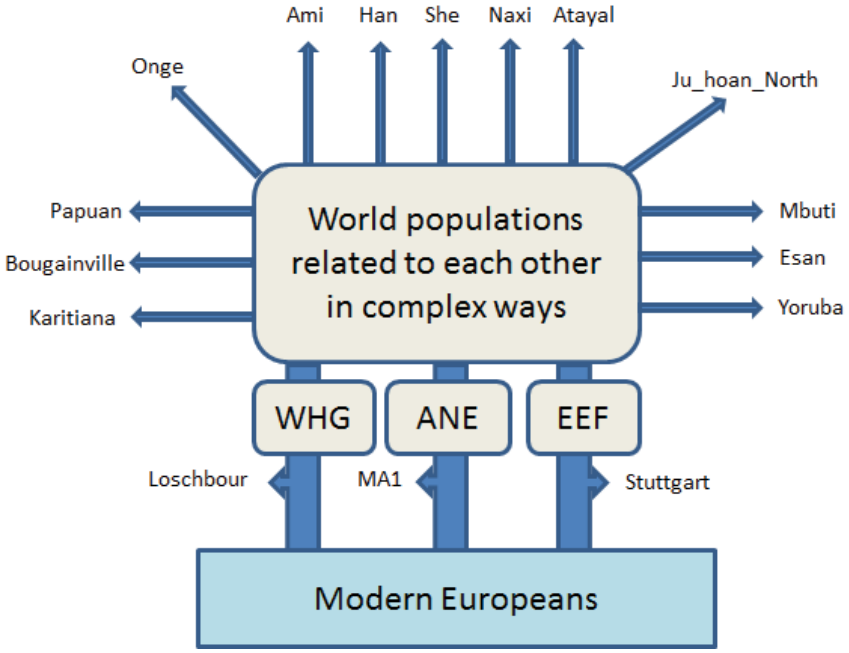
- 1-α fraction of ancestry of early European farmers (EEF): a sister group of Stuttgart
- α ancestry fraction of "Hunter", a population itself a mixture of:
  - β ancestry fraction of Loschbour-related west European hunter-gatherers (WHG)
  - 1-β ancestry fraction of MA1-related Ancient North Eurasians (ANE)

Thus, we can write:

$$\text{European} = (1-\alpha)\ \text{EEF} + \alpha(\beta\text{WHG}+(1-\beta)\text{ANE}) \quad \text{(S13.1)}$$

The above equation describes the fraction of ancestry inherited from a population of Stuttgart-like Early European Farmers and Loschbour/MA1-like pre-Neolithic hunter-gatherers, but this does not necessarily correspond to the actual populations historically involved. It is possible that this admixture took place in stages, so that, for example, the actual population responsible for the WHG/ANE ancestry in Europe already had some EEF ancestry. The historically involved populations will be revealed by ancient DNA studies of later periods of European history. However, our estimated α and β should correctly correspond to the ancestry proportions from the deep ancestors even in this case.

Figure S13.1: Admixture estimation that makes minimal assumptions about phylogeny. *We assume only that the three admixing populations (WHG, ANE, EEF) are sister groups of the ancient individuals (Loschbour, MA1, Stuttgart) and these are related in complex (but not modeled) ways with a set of outgroups. By exploiting correlations of* f_4*-statistics involving the ancient individuals and outgroups, we can estimate admixture proportions.*



We can write down an f_4-statistic involving Europeans and Stuttgart on the left-hand side as follows:

$$f_4(\text{European}, \text{Stuttgart}; O_1, O_2) = \quad \text{(S13.2)}$$
$$= \alpha\beta\ f_4(\text{Loschbour}, \text{Stuttgart}; O_1, O_2) + \alpha(1-\beta)\ f_4(\text{MA1}, \text{Stuttgart}; O_1, O_2)$$

The 1-α term has vanished because EEF and Stuttgart form a clade so that their allele frequency differences are uncorrelated to any of the outgroups. Using the 13 non-West Eurasians, we obtain 78 (O_1, O_2) pairs and thus 78 equations of the above form. We can then fit using least squares for the coefficients $A=\alpha\beta$ and $B=\alpha(1-\beta)$ and arrive at $\beta_{est} = 1/(1+B/A)$, $\alpha_{est} = A+B$ estimates of the parameters of interest, from which the estimated mixture proportions are (EEF=1- α_{est}, WHG= $\alpha_{est}\beta_{est}$, ANE= $\alpha_{est}(1-\beta_{est})$). We estimate standard errors using a jackknife[6] dropping one chromosome at a time[7].

The results are shown in Extended Data Table 2 together with other mixture estimates. We observe no systematic bias compared with the model-based estimates of SI12, as revealed by the number of standard errors by which the two estimates differ. None of the estimate differences exceed 1.9 standard errors. However, the mean and standard deviation of the estimate differences differ dramatically for each of the three ancestral proportions: 0.45±0.71 (EEF), -0.34±0.69 (WHG), and 0.06±0.65 (ANE) standard errors. Thus, the ANE estimates are rather precise, whereas the EEF and WHG estimates are more uncertain but are still consistent between the two methods.

We conclude that the method presented in this note and the fully model-based method presented in SI12 produce similar estimates for these populations, suggesting that the simple model devised in SI12 using Mbuti, Onge, Karitiana as the only non-west Eurasian populations and only two admixture events (basal admixture in Stuttgart and Ancient North Eurasian admixture in Karitiana) may capture some essential features of Eurasian prehistory.

Extended Data Table 2 includes, for completeness, aberrant estimates for six populations. We discuss the evidence for East Eurasian ancestry in Finns, Mordovians, and Russians in SI12; such ancestry is not accounted for in Equation 1, which assumes that all the ancestry of populations is EEF/WHG/ANE-related. The effect on the parameter fit is to produce negative EEF admixture; this is not surprising in view of Extended Data Fig. 6 which shows that Finns, Mordovians, and Russians differ from Stuttgart and most Europeans in sharing additional drift with Han, and the inclusion of Han and other East Asian populations in our set of world populations does not take this into account. The f_4-statistics used by our method are influenced both by the distant relationship of EEF/WHG/ANE to East Asians, and the more recent common drift shared by Finns, Mordovians, and Russians with some of them. Estonians, who exhibit the greatest discrepancy between the ancestry estimates that emerge from the full phylogenetic modeling in SI12 and the minimal phylogenetic modeling reported in this note, may harbor some of this ancestry as well.

The other three populations producing anomalous estimates in Extended Data Table 2 are Ashkenazi Jews, Sicilians, and Maltese. We observed in SI11 that these populations cannot be co-fit in the same admixture graph with most other Europeans, and this suggests that they do not fully trace their ancestry to the same EEF/WHG/ANE elements as most of Europe. Further evidence for this claim is presented in Extended Data Fig. 4 where all three populations have a negative value of f_4*(Test, Stuttgart; Loschbour, Chimp)*, and thus are inconsistent with a population of Stuttgart-related ancestry with additional Loschbour-related input, since such a population would have a zero or positive value of the statistic, as most Europeans do. All three populations strongly deviate towards the Near East in Extended Data Fig. 4 and Fig. 1B, and it is likely that they possess Near Eastern ancestry that is not mediated via Stuttgart.

In conclusion, the admixture estimates reported in this note show reasonable concordance with the fully model-based ones of SI12 for populations that have no evidence of additional ancestry beyond that which is represented by Stuttgart, Loschbour, and MA1. Additionally, populations that produce anomalous results in the present estimation coincide with those that fail to fit the model-based one, giving us more confidence in the results of both methods.

A caveat is that estimating mixture proportions on the basis of single ancient individuals is not easy, and in the case of MA1 we have to contend with the low coverage of the sample as well. Typically, ancestry estimation relies on the existence of large panels of individuals or allele frequency differences between populations to place the ancestry of additional single individuals[8, 9], the opposite

of what we are attempting here. Nonetheless, it is encouraging that these results establish many of the same patterns as the model-based ones. As more ancient genomes from the EEF/WHG/ANE groups become available, it may be possible to produce tighter estimates using methods such as this.

References

1 Maanasa Raghavan, Pontus Skoglund, Kelly E. Graf, Mait Metspalu, Anders Albrechtsen, Ida Moltke, Simon Rasmussen, Thomas W. Stafford Jr, Ludovic Orlando, Ene Metspalu, Monika Karmin, Kristiina Tambets, Siiri Rootsi, Reedik Magi, Paula F. Campos, Elena Balanovska, Oleg Balanovsky, Elza Khusnutdinova, Sergey Litvinov, Ludmila P. Osipova, Sardana A. Fedorova, Mikhail I. Voevoda, Michael DeGiorgio, Thomas Sicheritz-Ponten, Soren Brunak, Svetlana Demeshchenko, Toomas Kivisild, Richard Villems, Rasmus Nielsen, Mattias Jakobsson, and Eske Willerslev, 'Upper Palaeolithic Siberian Genome Reveals Dual Ancestry of Native Americans', *Nature,* advance online publication (2013).

2 Guido Brandt, Wolfgang Haak, Christina J. Adler, Christina Roth, Anna Szécsényi-Nagy, Sarah Karimnia, Sabine Möller-Rieker, Harald Meller, Robert Ganslmeier, Susanne Friederich, Veit Dresely, Nicole Nicklisch, Joseph K. Pickrell, Frank Sirocko, David Reich, Alan Cooper, Kurt W. Alt, and Consortium The Genographic, 'Ancient DNA Reveals Key Stages in the Formation of Central European Mitochondrial Genetic Diversity', *Science,* 342 (2013), 257-61.

3 Andreas Keller, Angela Graefen, Markus Ball, Mark Matzas, Valesca Boisguerin, Frank Maixner, Petra Leidinger, Christina Backes, Rabab Khairat, Michael Forster, Bjorn Stade, Andre Franke, Jens Mayer, Jessica Spangler, Stephen McLaughlin, Minita Shah, Clarence Lee, Timothy T. Harkins, Alexander Sartori, Andres Moreno-Estrada, Brenna Henn, Martin Sikora, Ornella Semino, Jacques Chiaroni, Siiri Rootsi, Natalie M. Myres, Vicente M. Cabrera, Peter A. Underhill, Carlos D. Bustamante, Eduard Egarter Vigl, Marco Samadelli, Giovanna Cipollini, Jan Haas, Hugo Katus, Brian D. O'Connor, Marc R. J. Carlson, Benjamin Meder, Nikolaus Blin, Eckart Meese, Carsten M. Pusch, and Albert Zink, 'New Insights into the Tyrolean Iceman's Origin and Phenotype as Inferred by Whole-Genome Sequencing', *Nat Commun,* 3 (2012), 698.

4 Pontus Skoglund, Helena Malmström, Maanasa Raghavan, Jan Storå, Per Hall, Eske Willerslev, M. Thomas P. Gilbert, Anders Götherström, and Mattias Jakobsson, 'Origins and Genetic Legacy of Neolithic Farmers and Hunter-Gatherers in Europe', *Science,* 336 (2012), 466-69.

5 N. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, and D. Reich, 'Ancient Admixture in Human History', *Genetics,* 192 (2012), 1065-93.

6 FrankM T. A. Busing, Erik Meijer, and RienVanDer Leeden, 'Delete-M Jackknife for Unequal M', *Statistics and Computing,* 9 (1999), 3-8.

7 Priya Moorjani, Nick Patterson, Joel N. Hirschhorn, Alon Keinan, Li Hao, Gil Atzmon, Edward Burns, Harry Ostrer, Alkes L. Price, and David Reich, 'The History of African Gene Flow into Southern Europeans, Levantines, and Jews', *PLoS Genet,* 7 (2011), e1001373.

8 David Alexander, and Kenneth Lange, 'Enhancements to the Admixture Algorithm for Individual Ancestry Estimation', *BMC Bioinformatics,* 12 (2011), 246.

9 Alkes L. Price, Johannah Butler, Nick Patterson, Cristian Capelli, Vincenzo L. Pascali, Francesca Scarnicci, Andres Ruiz-Linares, Leif Groop, Angelica A. Saetta, Penelope Korkolopoulou, Uri Seligsohn, Alicja Waliszewska, Christine Schirmer, Kristin Ardlie, Alexis Ramos, James Nemesh, Lori Arbeitman, David B. Goldstein, David Reich, and Joel N. Hirschhorn, 'Discerning the Ancestry of European Americans in Genetic Association Studies', *PLoS Genet,* 4 (2008), e236.

Supplementary Information 14

Segments identical due to shared descent between modern and archaic samples

Joshua G. Schraiber*, Montgomery Slatkin

*To whom correspondence should be addressed (jgschraiber@berkeley.edu)

We analyzed the sharing of tracts of identity by descent (IBD) between present-day and ancient samples by using the POPRES SNP genotyping dataset (Nelson *et al.,* 2008), along with sequence data generated for the analysis of the Denisova individual (Meyer *et al.*, 2012).

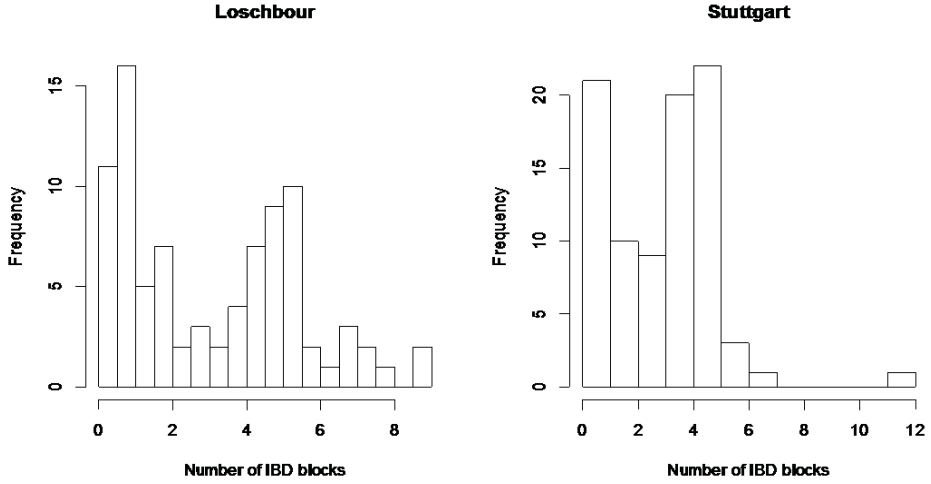
For every SNP in the POPRES dataset, we used the genotype calls for Loschbour and Stuttgart generated by the Genome Analysis Toolkit (GATK) (SI 2). We detected likely segments of IBD using RefinedIBD as implemented in BEAGLE 4 (Browning and Browning, 2013) with the settings "ibdtrim=20" and "ibdwindow=25". We kept all IBD tracts spanning at least 0.5 centimorgans (cM) and with a LOD score > 3. We note that in fact we are detecting segments that are identity by state (IBS), but previous studies have shown that they correlate strongly to IBD segments (e.g. Ralph and Coop 2013).

We quantified IBD sharing as the average number of IBD blocks shared between two populations, P_i and P_j,

$$S_{ij} = \frac{\sum_{k \in P_i} \sum_{l \in P_j} N_{kl}}{n_i n_j} \quad \text{(S8.1)}$$

where n_i is the number of individuals in population *i, k* and *l* index individuals, and N_{kl} is the number of IBD blocks shared between individuals *k* and *l*.

Figure S14.1. Histogram of IBD sharing between ancient and present-day samples. *In each panel, a histogram of the average number of IBD blocks shared between either Loschbour (panel A, mean = 3.18) or Stuttgart (Panel B, mean = 3.01) and present-day populations is shown.*



We detected substantial IBD sharing between present-day populations, replicating the results of Ralph and Coop (2013). In addition, our method inferred IBD sharing both between the ancient samples, and between the ancient and present-day samples (Figure S14.1).

We examined in detail the distribution of IBD sharing between present-day and ancient populations, and in Table S14.1 report the top 10 present-day populations that share IBD blocks with Loschbour and Stuttgart. According to Ralph and Coop (2013), most IBD sharing between present-day populations is due to ancestors living in the last 2-3 thousand years. On the surface, our results suggest that IBD sharing can potentially last for substantially longer.

Table S14.1. The 10 populations that share the most IBD with each of Loschbour and Stuttgart.

Loschbour		Stuttgart	
Present-day Population	**Mean number of shared IBD blocks**	**Present-day Population**	**Mean number of shared IBD blocks**
Denmark	9	Sardinian	12
European immigrants to North America	9	Slovakia	7
Finland	8	European immigrants to Zimbabwe	6
Ukraine	7.5	Macedonia	5.5
European immigrants to South Africa	7.5	Slovenia	5.5
French	7	Bulgaria	5
Sweden	6.6	Ukraine	5
Scotland	6.6	Latvia	5
Russia	6.2	Cyprus	5
Latvia	6	Swiss-Italian	4.8

Note: For each modern population listed, we report the average number of IBD blocks per individual

We hypothesize that our detection of segments of IBD beyond the threshold of the population separation time highlighted Ralph and Coop is likely due to these being segments of the genome that have very low recombination rates, allowing signals of IBD to persist over longer times (as a larger physical distance span is available for detection).

Alternatively, it is possible that some of the evidence for IBD is artifactual due to shared selective sweeps in a common ancestral population (IBS), which result in false-positive signals of IBD sharing as it is in fact difficult to detect real differences between any haplotypes in the region.

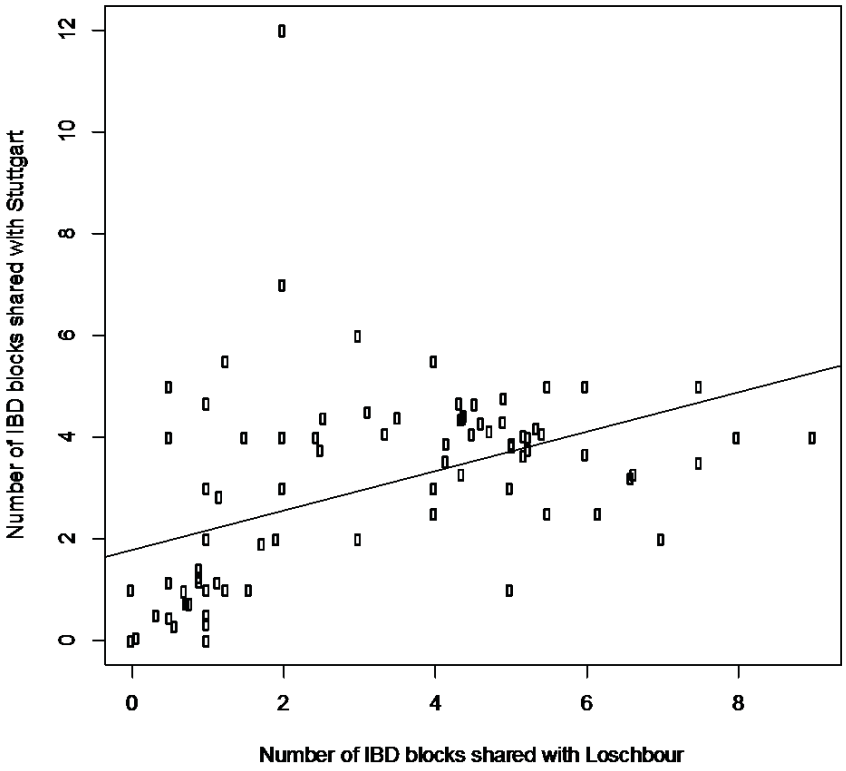
Whatever the explanation for the detected segments of shared IBD, we explored whether the ordering of populations based on the inferred IBD segments mirrored the genetic relationships we inferred from other aspects of the data. We observe areas of notable concordance.

- Evidence for deep relatedness of Loschbour and Stuttgart. The patterns of IBD sharing of Loschbour and Stuttgart to other world populations are positively correlated (Figure S14.2). This is consistent with these two populations being deeply related so that they have correlated levels of shared IBD to non-West Eurasian populations (e.g. Africans or eastern non-Africans). Loschbour shares slightly more IBD tracts with the present-day populations that happen to be in the POPRES dataset than does Stuttgart (3.18 vs. 3.01, respectively).

- Evidence that Loschbour is genetically closer to northern Europeans and that Stuttgart is genetically closer to southern Europeans. The top 10 populations in terms of IBD sharing with Loschbour tend to be in northern Europe or migrants from northern Europe. The top 10 populations in terms of IBD sharing with Stuttgart tend to be in southern Europe or migrants from southern Europe. These patterns are consistent with relatively higher proportions of WHG

ancestry in both Loschbour and northern European populations, and higher proportions of EEF ancestry in both Stuttgart and southern European populations.

Figure S14.2. IBD sharing between Loschbour and Stuttgart is correlated. *Each point corresponds to a modern population, plotted according to its average sharing with Loschbour (x-axis) and Stuttgart (y-axis). Spearman rank correlation = 0.59, slope of best fit line = 0.39.*



References

1. Nelson, M.R., et al., *The population reference sample, POPRES: a resource for population, disease, and pharmacological genetics research.* American Journal of Human Genetics, 2008. **83**(3): p. 347-58.
2. Browning, B.L. and Browning, S.R. *Improving the accuracy and efficiency of identity by descent detection in population data*. Genetics, 2013. **194**(2): p. 459-71.
3. Ralph, P. and Coop G. *The geography of recent genetic ancestry across Europe.* PLoS Biology, 2013. 11(5): p. e1001555.
4. Meyer, M. et al. *A high-coverage genome sequence from an archaic Denisovan individual*. Science, 2012. **338**(6104): p. 222-6.