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

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters.

Citation

Published Version
doi:10.1038/nature13673

Accessed
May 10, 2017 8:56:47 AM EDT

Citable Link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:14351136

Terms of Use
This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

(Article begins on next page)
Ancient human genomes suggest three ancestral populations for present-day Europeans

A full list of authors and affiliations appears at the end of the article.

Abstract

We sequenced the genomes of a ~7,000 year old farmer from Germany and eight ~8,000 year old hunter-gatherers from Luxembourg and Sweden. We analyzed these and other ancient genomes with 2,345 contemporary humans to show that most present Europeans derive from at least three highly differentiated populations: West European Hunter-Gatherers (WHG), who contributed ancestry to all Europeans but not to Near Easterners; Ancient North Eurasians (ANE) related to Upper Paleolithic Siberians, who contributed to both Europeans and Near Easterners; and Early European Farmers (EEF), who were mainly of Near Eastern origin but also harbored WHG-related ancestry. We model these populations’ deep relationships and show that EEF had ~44% ancestry from a “Basal Eurasian” population that split prior to the diversification of other non-African lineages.

Supplementary Information is linked to the online version of the paper at www.nature.com/nature. The human Origins dataset can be found at http://genetics.med.harvard.edu/reichlab/Reich_Lab/Datasets.html. The full version of the dataset (including additional samples) is available to researchers who send a signed letter to DR indicating that they will abide by specified usage conditions (SI).

Author contributions

BB, EEE, JBu, MS, SP, JKe, DR and JKr supervised the study. IL, NP, AM, GR, SM, KK, PHS, JGS, SC, ML, QF, HL, CdF, KP, WH, MMet, MMey and DR analyzed genetic data. FH, EF, DD, MF, J-MG, JW, AC and JKr obtained human remains. AM, CE, RBo, KB, SS, CP, NR and JKr processed ancient DNA. IL, NP, SN, NR, GA, HAB, GBa, EB, OB, RBa, GBe, HB-A, JBe, FBe, CMB, FBr, GBB, FC, MC, DECC, DCor, LD, GvD, SD, J-MD, SAD, IGR, MG, MH, BH, TH, UH, ARJ, SK-Y, RKh, EK, RKi, TK, WK, VK, AK, LL, SL, TL, RWBM, BM, EM, JMol, JMou, KN, DN, TN, LO, JP, FP, OLP, VR, FR, IR, RR, HS, ASaj, ASal, EBS, ATar, DT, ST, IU, OU, RVa, MVi, MVo, CW, LY, PZ, TZ, CC, MGT, AR-L, SAT, LS, KT, RVi, DCom, RS, MMet, SP and DR assembled the genotyping dataset. IL, NP, DR and JKr wrote the manuscript with help from all co-authors.

The aligned sequences are available through the European Nucleotide Archive under accession number PRJEB6272.

The authors declare competing financial interests: UH is an employee of Illumina, TL is an employee of AMGEN, and JM is an employee of 23andMe.
Loschbour rock shelter in Luxembourg, discovered in the context of hunter-gatherer artifacts (SI1; SI2); and seven ~8,000 year old samples (0.01–2.4-fold) from a hunter-gatherer burial in Motala, Sweden (the highest coverage individual was “Motala12”).

Sequence reads from all samples revealed >20% C→T and G→A deamination-derived mismatches at the ends of the molecules that are characteristic of ancient DNA (SI3). We estimate nuclear contamination rates to be 0.3% for Stuttgart and 0.4% for Loschbour (SI3), and mitochondrial (mtDNA) contamination rates to be 0.3% for Stuttgart, 0.4% for Loschbour, and 0.01–5% for the Motala individuals (SI3). Stuttgart has mtDNA haplogroup T2, typical of Neolithic Europeans (SI3), and Loschbour and all Motala individuals have the U5 or U2 haplogroups, typical of hunter-gatherers (SI4). Stuttgart is female, while Loschbour and five Motala individuals are male (SI5) and belong to Y-chromosome haplogroup I, suggesting that this was common in pre-agricultural Europeans (SI5).

We carried out large-scale sequencing of libraries prepared with uracil DNA glycosylase (UDG), which removes deaminated cytosines, thus reducing errors arising from ancient DNA damage (SI3). The ancient individuals had indistinguishable levels of Neanderthal ancestry when compared to each other (~2%) and to present-day Eurasians (SI6). The heterozygosity of Stuttgart (0.00074) is at the high end of present-day Europeans, while that of Loschbour (0.00048) is lower than in any present humans (SI2), reflecting a strong bottleneck in Loschbour’s ancestors as the genetic data show that he was not recently inbred (Extended Data Fig. 2). High copy numbers for the salivary amylase gene (AMY1) have been associated with a high starch diet (SI12); our data are consistent with this finding in that the ancient hunter gatherers La Braña (from Iberia) (SI2), Motala12, and Loschbour had 5, 6 and 13 copies respectively, whereas the Stuttgart farmer had 16 (SI7). Both Loschbour and Stuttgart had dark hair (>99% probability); and Loschbour, like La Braña and Motala12, likely had blue or intermediate-colored eyes (>75%) while Stuttgart likely had brown eyes (>99%) (SI8). Neither Loschbour nor La Braña carries the skin-lightening allele in SLC24A5 that is homozygous in Stuttgart and nearly fixed in Europeans today (SI2), but Motala12 carries at least one copy of the derived allele, showing that this allele was present in Europe prior to the advent of agriculture.

We compared the ancient genomes to 2,345 present-day humans from 203 populations genotyped at 594,924 autosomal single nucleotide polymorphisms (SNPs) with the Human Origins array (SI9) (Extended Data Table 1). We used ADMIXTURE (SI9 and Extended Data Fig. 3). Principal component analysis (PCA) (SI10) (Fig. 1B) indicates a discontinuity between the Near East and Europe, with each showing north-south clines bridged only by a few populations of mainly Mediterranean origin. We projected the newly sequenced and previously published ancient genomes onto the first two principal components (PCs) (Fig. 1B). Upper Paleolithic hunter-gatherers from Siberia like the MA1 (Mal’ta) individual project at the northern end of the PCA, suggesting an “Ancient North Eurasian” meta-population (ANE). European hunter-gatherers from Spain, Luxembourg, and Sweden fall beyond present-day Europeans in the direction of European differentiation from the Near East, and form a “West European Hunter-Gatherer” (WHG) cluster including Loschbour and La Braña, and a “Scandinavian Hunter-Gatherer” (SHG) cluster including
the Motala individuals and ~5,000 year old hunter-gatherers from the Pitted Ware Culture\(^4\). An “Early European Farmer” (EEF) cluster includes Stuttgart, the ~5,300 year old Tyrolean Iceman\(^1\) and a ~5,000 year old Swedish farmer\(^4\).

Patterns observed in PCA may be affected by sample composition (SI10) and their interpretation in terms of admixture events is not straightforward, so we rely on formal analysis of \(f\)-statistics\(^8\) to document mixture of at least three source populations in the ancestry of present Europeans. We began by computing all possible statistics of the form \(f_3(\text{Test}; \text{Ref}_1, \text{Ref}_2)\) (SI11), which if significantly negative show unambiguously\(^8\) that Test is admixed between populations anciently related to \(\text{Ref}_1\) and \(\text{Ref}_2\) (we choose \(\text{Ref}_1\) and \(\text{Ref}_2\) from 5 ancient and 192 present populations). The lowest \(f_3\)-statistics for Europeans are negative (93% are >4 standard errors below 0), with most showing strong support for at least one ancient individual being one of the references (SI11). Europeans almost always have their lowest \(f_3\) with either (EEF, ANE) or (WHG, Near East) (SI11, Table 1, Extended Data Table 1), which would not be expected if there were just two ancient sources of ancestry (in which case the best references for all Europeans would be similar). The lowest \(f_3\)-statistic for Near Easterners always takes Stuttgart as one of the reference populations, consistent with a Near Eastern origin for Stuttgart’s ancestors (Table 1). We also computed the statistic \(f_4(\text{Test, Stuttgart}; \text{MA1, Chimp})\), which measures whether MA1 shares more alleles with a Test population or with Stuttgart. This statistic is significantly positive (Extended Data Fig. 4, Extended Data Table 1) if Test is nearly any present-day West Eurasian population, showing that MA1-related ancestry has increased since the time of early farmers like Stuttgart (the analogous statistic using Native Americans instead of MA1 is correlated but smaller in magnitude (Extended Data Fig. 5), indicating that MA1 is a better surrogate than the Native Americans who were first used to document ANE ancestry in Europe\(^7,8\)). The analogous statistic \(f_4(\text{Test, Stuttgart}; \text{Loschbour, Chimp})\) is nearly always positive in Europeans and negative in Near Easterners, indicating that Europeans have more ancestry from populations related to Loschbour than do Near Easterners (Extended Data Fig. 4, Extended Data Table 1). Extended Data Table 2 documents the robustness of key \(f_4\)-statistics by recomputing them using transversion polymorphisms not affected by ancient DNA damage, and also using whole-genome sequencing data not affected by SNP ascertainment bias. Extended Data Fig. 6 shows the geographic gradients in the degree of allele sharing of present-day West Eurasians (as measured by \(f_4\)-statistics) with Stuttgart (EEF), Loschbour (WHG) and MA1 (ANE).

To determine the minimum number of source populations needed to explain the data for many European populations taken together, we studied the matrix of all possible statistics of the form \(f_4(\text{Test}_{\text{base}}, \text{Test}_i; \text{O}_{\text{base}}, \text{O}_j)\) (SI12). \(\text{Test}_{\text{base}}\) is a reference European population, \(\text{Test}_i\) is the set of all other European \(\text{Test}\) populations, \(\text{O}_{\text{base}}\) is a reference outgroup, and \(\text{O}_j\) is the set of other outgroups (ancient DNA samples, Onge, Karitiana, and Mbuti). The rank of the \((i, j)\) matrix reflects the minimum number of sources that contributed to the \(\text{Test}\) populations\(^16,17\). For a pool of individuals from 23 \(\text{Test}\) populations representing most present-day European groups, this analysis rejects descent from just two sources (\(P<10^{-12}\) by a Hotelling T-test\(^17\)). However, three source populations are consistent with the data after excluding the Spanish who have evidence for African admixture\(^18\sim20\) (\(P=0.019\), not
significant after multiple-hypothesis correction), consistent with the results from ADMIXTURE (SI9), PCA (Fig. 1B, SI10) and f-statistics (Extended Data Table 1, Extended Data Fig. 6, SI11, SI12). We caution that the finding of three sources could be consistent with a larger number of mixture events. Moreover, the source populations may themselves have been mixed. Indeed, the positive \( f_{4}(\text{Stuttgart, Test; Loschbour, Chimp}) \) statistics obtained when \texttt{Test} is Near Eastern (Extended Data Table 1) imply that the EEF had some WHG-related ancestry, which was greater than 0% and as high as 45% (SI13).

We used the ADMIXTUREGRAPH software\(^8,15\) to fit a model (a tree structure augmented by admixture events) to the data, exploring models relating the three ancient populations (Stuttgart, Loschbour, and MA1) to two eastern non-Africans (Onge and Karitiana) and sub-Saharan Africans (Mbuti). We found no models that fit the data with 0 or 1 admixture events, but did find a model that fit with 2 admixture events (SI14). The successful model (Fig. 2A) confirms the existence of MA1-related admixture in Native Americans\(^3\), but includes the novel inference that Stuttgart is partially (44 ± 10%) derived from a lineage that split prior to the separation of eastern non-Africans from the common ancestor of WHG and ANE. The existence of such “Basal Eurasian” admixture into Stuttgart provides a simple explanation for our finding that diverse eastern non-African populations share significantly more alleles with ancient European and Upper Paleolithic Siberian hunter-gatherers than with Stuttgart (that is, \( f_{4}(\text{Eastern non-African, Chimp; Hunter-gatherer, Stuttgart}) \) is significantly positive), but that hunter-gatherers appear to be equally related to most eastern groups (SI14). We verified the robustness of the model by reanalyzing the data using the unsupervised MixMapper\(^7\) (SI15) and TreeMix\(^21\) software (SI16), which both identified the same admixture events. The ANE/WHG split must have occurred >24,000 years ago (as it must predate the age of MA1\(^3\)), and the WHG/Eastern non-African split must have occurred >40,000 years ago (as it must predate the Tianyuan\(^22\) individual from China which clusters with Asians to the exclusion of Europeans). The Basal Eurasian split must be even older, and might be related to early settlement of the Levant\(^23\) or Arabia\(^24,25\) prior to the diversification of most Eurasians, or more recent gene flow from Africa\(^26\). However, the Basal Eurasian population shares much of the genetic drift common to non-African populations after their separation from Africans, and thus does not appear to represent gene flow between sub-Saharan Africans and the ancestors of non-Africans after the out-of-Africa bottleneck (SI14).

Fitting present-day Europeans into the model, we find that few populations can be fit as 2-way mixtures, but nearly all are compatible with 3-way mixtures of ANE/EEF/WHG (SI14). The mixture proportions from the fitted model (Fig. 2B; Extended Data Table 3) are encouragingly consistent with those obtained from a separate method that relates European populations to diverse outgroups using \( f_{4} \)-statistics, assuming only that MA1 is an unmixed descendent of ANE, Loschbour of WHG, and Stuttgart of EEF (SI17). We infer that EEF ancestry in Europe today ranges from ~30% in the Baltic region to ~90% in the Mediterranean, consistent with patterns of identity-by-descent (IBD) sharing\(^27,28\) (SI18) and shared haplotype analysis (chromosome painting)\(^29\) (SI19) in which Loschbour shares more segments with northern Europeans and Stuttgart with southern Europeans. Southern Europeans inherited their European hunter-gatherer ancestry mostly via EEF ancestors.
(Extended Data Fig. 6), while Northern Europeans acquired up to 50% of WHG ancestry above and beyond the WHG-related ancestry which they received through their EEF ancestors. Europeans have a larger proportion of WHG than ANE ancestry in general. By contrast, in the Near East there is no detectable WHG ancestry, but up to ~29% ANE in the Caucasus (SI14). A striking feature of these findings is that ANE ancestry is inferred to be present in nearly all Europeans today (with a maximum of ~20%), but was absent in both farmers and hunter-gatherers from central/western Europe during the Neolithic transition. At the same time, we infer that ANE ancestry was not completely absent from the larger European region at that time: we find that it was present in ~8,000 years old Scandinavian hunter-gatherers, since MA1 shares more alleles with Motala12 (SHG) than with Loschbour, and Motala12 fits as a mixture of 81% WHG and 19% ANE (SI14).

Two sets of European populations are poor fits for the model. Sicilians, Maltese, and Ashkenazi Jews have EEF estimates of >100% consistent with their having more Near Eastern ancestry than can be explained via EEF admixture (SI17). They also cannot be jointly fit with other Europeans (SI14), and they fall in the gap between European and Near Easterners (Fig. 1B). Finns, Mordovians and Russians (from the northwest of Russia) also do not fit (SI14; Extended Data Table 3) due to East Eurasian gene flow into the ancestors of these northeastern European populations. These populations (and Chuvash and Saami) are more related to East Asians than can be explained by ANE admixture (Extended Data Fig. 7), likely reflecting a separate stream of Siberian gene flow into northeastern Europe (SI14).

Several questions will be important to address in future ancient DNA work. Where and when did the Near Eastern farmers admix with European hunter-gatherers to produce the EEF? How did the ancestors of present-day Europeans first acquire their ANE ancestry? Discontinuity in central Europe during the late Neolithic (~4,500 years ago) associated with the appearance of mtDNA types absent in earlier farmers and hunter-gatherers raises the possibility that ANE ancestry may have also appeared at this time. Finally, it is important to study ancient genome sequences from the Near East to provide insights into the history of the Basal Eurasians.

Online Methods

Archeological context, sampling and DNA extraction

The Loschbour sample stems from a male skeleton excavated in 1935 at the Loschbour rock shelter in Heffingen, Luxembourg. The skeleton was AMS radiocarbon dated to 7,205 ± 50 years before present (OxA-7738; 6,220-5,990 cal BC). At the Palaeogenetics Laboratory in Mainz, material for DNA extraction was sampled from tooth 16 (an upper right M1 molar) after irradiation with UV-light, surface removal, and pulverization in a mixer mill. DNA extraction took place in the palaeogenetics facilities in the Institute for Archaeological Sciences at the University of Tübingen. Three extracts were made in total, one from 80 mg of powder using an established silica based protocol and two additional extracts from 90 mg of powder each with a protocol optimized for the recovery of short DNA molecules.

The Stuttgart sample was taken from a female skeleton excavated in 1982 at the site Viesenhäuser Hof, Stuttgart-Mühlhausen, Germany. It was attributed to the
Linearbandkeramik (5,500-4,800 BC) through associated pottery artifacts and the chronology was corroborated by radiocarbon dating of the stratigraphy\textsuperscript{34}. Both sampling and DNA extraction took place in the Institute for Archaeological Sciences at the University of Tübingen. Tooth 47 (a lower right M2 molar) was removed and material from the inner part was sampled with a sterile dentistry drill. An extract was made using 40 mg of bone powder\textsuperscript{33}.

The Motala individuals were recovered from the site of Kanaljorden in the town of Motala, Östergötland, Sweden, excavated between 2009 and 2013. The human remains at this site are represented by several adult skulls and one infant skeleton. All individuals are part of a ritual deposition at the bottom of a small lake. Direct radiocarbon dates on the remains range between 7,013 ± 76 and 6,701 ± 64 BP (6,361-5,516 cal BC), corresponding to the late Middle Mesolithic of Scandinavia. Samples were taken from the teeth of the nine best preserved skulls, as well as a femur and tibia. Bone powder was removed from the inner parts of the teeth or bones with a sterile dentistry drill. DNA from 100 mg of bone powder was extracted\textsuperscript{35} in the ancient DNA laboratory of the Archaeological Research Laboratory, Stockholm.

**Library preparation**

Illumina sequencing libraries were prepared using either double- or single-stranded library preparation protocols\textsuperscript{36,37} (SI1). For high-coverage shotgun sequencing libraries, a DNA repair step with Uracil-DNA-glycosylase (UDG) and endonuclease VIII (endo VIII) treatment was included in order to remove uracil residues\textsuperscript{38}. Size fractionation on a PAGE gel was also performed in order to remove longer DNA molecules that are more likely to be contaminants\textsuperscript{37}. Positive and blank controls were carried along during every step of library preparation.

**Shotgun sequencing and read processing**

All non-UDG-treated libraries were sequenced either on an Illumina Genome Analyzer IIX with 2×76 + 7 cycles for the Loschbour and Motala libraries, or on an Illumina MiSeq with 2×150 + 8 + 8 cycles for the Stuttgart library. We followed the manufacturer’s protocol for multiplex sequencing. Raw overlapping forward and reverse reads were merged and filtered for quality\textsuperscript{39} and mapped to the human reference genome (hg19/GRCh37/1000Genomes) using the Burrows-Wheeler Aligner (BWA)\textsuperscript{40} (SI2). For deeper sequencing, UDG-treated libraries of Loschbour were sequenced on 3 Illumina HiSeq 2000 lanes with 50-bp single-end reads, 8 Illumina HiSeq 2000 lanes of 100-bp paired-end reads and 8 Illumina HiSeq 2500 lanes of 101-bp paired-end reads. The UDG-treated library for Stuttgart was sequenced on 8 HiSeq 2000 lanes and 101-bp paired-end reads. The UDG-treated libraries for Motala 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). We also sequenced an additional 8 HiSeq 2000 lanes for Motala12, the Motala sample with the highest percentage of endogenous human DNA. For the Loschbour and Stuttgart high coverage individuals, diploid genotype calls were obtained using the Genome Analysis Toolkit (GATK)\textsuperscript{41}.  

*Nature. Author manuscript; available in PMC 2015 March 18.*
Enrichment of mitochondrial DNA and sequencing

To test for DNA preservation and mtDNA contamination non-UDG-treated libraries of Loschbour and all Motala samples were enriched for human mitochondrial DNA using a bead-based capture approach with present-day human DNA as bait\textsuperscript{42}. UDG-treatment was omitted in order to allow characterization of damage patterns typical for ancient DNA\textsuperscript{10}. The captured libraries were sequenced on an Illumina Genome Analyzer IIX platform with 2 × 76 + 7 cycles and the resulting reads were merged and quality filtered\textsuperscript{39}. The sequences were mapped to the Reconstructed Sapiens Reference Sequence, RSRS\textsuperscript{43}, using a custom iterative mapping assembler, MIA\textsuperscript{44} (SI4).

Contamination estimates

We assessed if the sequences had the characteristics of authentic ancient DNA using four approaches. First we searched for evidence of contamination by determining whether the sequences mapping to the mitochondrial genome were consistent with deriving from more than one individual\textsuperscript{44,45}. Second, for the high-coverage Loschbour and Stuttgart genomes, we used a maximum-likelihood-based estimate of autosomal contamination that uses variation at sites that are fixed in the 1000 Genomes data to estimate error, heterozygosity and contamination\textsuperscript{46} simultaneously. Third, we estimated contamination based on the rate of polymorphic sites on the X chromosome of the male Loschbour individual\textsuperscript{47} (SI3) Fourth, we analyzed non-UDG treated reads mapping to the RSRS to search for aDNA-typical damage patterns resulting in C\(\rightarrow\)T changes at the 5′-end of the molecule\textsuperscript{10} (SI3).

Phylogenetic analysis of the mitochondrial genomes

All nine complete mitochondrial genomes that fulfilled the criteria of authenticity were assigned to haplogroups using Haplofind\textsuperscript{48}. A Maximum Parsimony tree including present day humans and previously published ancient mtDNA sequences was generated with MEGA\textsuperscript{49}. The effect of branch shortening due to a lower number of substitutions in ancient lineages was studied by calculating the nucleotide edit distance to the root for all haplogroup R sequences (SI4).

Sex determination and Y-chromosome analysis

We assessed the sex of all sequenced individuals by using the ratio of (chrY) to (chrY +chrX) aligned reads\textsuperscript{50}. We downloaded a list of Y-chromosome SNPs curated by the International Society of Genetic Genealogy (ISOGG, \url{http://www.isogg.org}) v. 9.22 (accessed Feb. 18, 2014) and determined the state of the ancient individuals at positions where a single allele was observed and MAPQ ≥0. We excluded C/G or A/T SNPs due to uncertainty about the polarity of the mutation in the database. The ancient individuals were assigned haplogroups based on their derived state (SI5). We also used BEAST v1.7.5\textsuperscript{51} to assess the phylogenetic position of Loschbour using 623 males from around the world with 2,799 variant sites across 500kb of non-recombining Y-chromosome sequence\textsuperscript{52} (SI5).

Estimation of Neanderthal admixture

We estimate Neanderthal admixture in ancient individuals with the \(f_4\)-ratio or \(S\)-statistic\textsuperscript{8,53,54} 
\[
\alpha = \frac{f_4(Altai, Denisova; Test, Yoruba)}{f_4(Altai, Denisova; Vindija, Yoruba)}
\]
which uses whole genome data from Altai, a high coverage (52×) Neanderthal genome sequence\textsuperscript{55}, Denisova, a high coverage sequence\textsuperscript{37} from another archaic human population (31×), and Vindija, a low coverage (1.3×) Neanderthal genome from a mixture of three Neanderthal individuals from Vindija Cave in Croatia\textsuperscript{53}.

**Inference of demographic history and inbreeding**

We used the Pairwise Sequentially Markovian Coalescent (PSMC)\textsuperscript{56} to infer the size of the ancestral population of Stuttgart and Loschbour. This analysis requires high quality diploid genotype calls and cannot be performed in the low-coverage Motala samples. To determine whether the low effective population size inferred for Loschbour is due to recent inbreeding, we plotted the time-to-most-recent common ancestor (TMRCA) along each of chr1-22 to detect runs of low TMRCA.

**Analysis of segmental duplications and copy number variants**

We built read-depth based copy number maps for the Loschbour, Stuttgart and Motala12 genomes in addition to the Denisova and Altai Neanderthal genome and 25 deeply sequenced modern genomes\textsuperscript{55} (SI7). We built these maps by aligning reads, subdivided into their non-overlapping 36-bp constituents, against the reference genome using the mrsFAST aligner\textsuperscript{57}, and renormalizing read-depth for local GC content. We estimated copy numbers in windows of 500 unmasked base pairs slid at 100 bp intervals across the genome. We called copy number variants using a scale space filter algorithm. We genotyped variants of interest and compared the genotypes to those from individuals sequenced as part of the 1000 Genomes Project\textsuperscript{58}.

**Phenotypic inference**

We inferred likely phenotypes (SI8) by analyzing DNA polymorphism data in the VCF format\textsuperscript{59} using VCFtools (http://vcftools.sourceforge.net/). For the Loschbour and Stuttgart individuals, we included data from sites not flagged as LowQuality, with genotype quality (GQ) of ≥30, and SNP quality (QUAL) of ≥50. For Motala12, which is of lower coverage, we included sites having at least 2× coverage and that passed visual inspection of the local alignment using samtools tview (http://samtools.sourceforge.net)\textsuperscript{60}

**Human Origins dataset curation**

The Human Origins array consists of 14 panels of SNPs for which the ascertainment is well known\textsuperscript{8,61}. All population genetics analysis were carried out on a set of 594,924 autosomal SNPs, after restricting to sites that had >90% completeness across 7 different batches of sequencing, and that had >97.5% concordance with at least one of two subsets of samples for which whole genome sequencing data was also available. The total dataset consists of 2,722 individuals, which we filtered to 2,345 individuals (203 populations) after removing outlier individuals or relatives based on visual inspection of PCA plots\textsuperscript{14,62} or model-based clustering analysis\textsuperscript{13}. Whole genome amplified (WGA) individuals were not used in analysis, except for a Saami individual who we included because of the special interest of this population for Northeastern European population history (Extended Data Fig. 7).
ADMIXTURE analysis

We merged all Human Origins genotype data with whole genome sequencing data from Loschbour, Stuttgart, MA1, Motala12, Motala_merge, and LaBrana. We then thinned the resulting dataset to remove SNPs in linkage-disequilibrium with PLINK 1.07, using a window size of 200 SNPs advanced by 25 SNPs and an $r^2$ threshold of 0.4. We ran ADMIXTURE 1.23 for 100 replicates with different starting random seeds, default 5-fold cross-validation, and varying the number of ancestral populations $K$ between 2 and 20. We assessed clustering quality using CLUMPP. We used the ADMIXTURE results to identify a set of 59 “West Eurasian” (European/Near Eastern) populations based on values of a “West Eurasian” ancestral population at $K=3$ (SI9). We also identified 15 populations for use as “non-West Eurasian outgroups” based on their having at least 10 individuals and no evidence of European or Near Eastern admixture at $K=11$, the lowest $K$ for which Near Eastern/European-maximized ancestral populations appeared consistently across all 100 replicates.

Principal Components Analysis

We used smartpca (version: 10210) from EIGENSOFT 62,66 to carry out Principal Components Analysis (PCA) (SI10). We performed PCA on a subset on individuals and then projected others using the lsqproject: YES option that gives an unbiased inference of the position of samples even in the presence of missing data (especially important for ancient DNA).

$f_3$-statistics

We use the $f_3$-statistic\(^8\) $f_3(\text{Test}; \text{Ref}_1 \text{Ref}_2) = \frac{1}{N} \sum_{i=1}^{N} (t_i - r_{1,i})(t_i - r_{2,i})$, where $t_i$, $r_{1,i}$ and $r_{2,i}$ are the allele frequencies for the $i^{th}$ SNP in populations Test, Ref$_1$, Ref$_2$, respectively, to determine if there is evidence that the Test population is derived from admixture of populations related to Ref$_1$ and Ref$_2$ (SI11). A significantly negative statistic provides unambiguous evidence of mixture in the Test population\(^8\). We allow Ref$_1$ and Ref$_2$ to be any Human Origins population with 4 or more individuals, or Loschbour, Stuttgart, MA1, Motala12, LaBrana. We assess significance of the $f_3$-statistics using a block jackknife and a block size of 5cM. We report significance as the number of standard errors by which the statistic differs from zero (Z-score). We also perform an analysis in which we constrain the reference populations to be (i) EEF (Stuttgart) and WHG (Loschbour or LaBrana), (ii) EEF and a Near Eastern population, (iii) EEF and ANE (MA1), or (iv) any two present-day populations, and compute a $Z_{\text{diff}}$ score between the lowest $f_3$-statistic observed in the dataset, and the $f_3$-statistic observed for the specified pair.

$f_4$-statistics

We analyze $f_4$-statistics\(^8\) of the form $f_4(A; B; C; D) = \frac{1}{N} \sum_{i=1}^{N} (a_i - b_i)(c_i - d_i)$ to assess if populations A, B are consistent with forming a clade in an unrooted tree with respect to C, D. If they form a clade, the allele frequency differences between the two pairs should be uncorrelated and the statistic has an expected value of 0. We set the outgroup $D$ to be a sub-Saharan African population or Chimpanzee. We systematically tried all possible
combinations of the ancient samples or 15 “non-West Eurasian outgroups” identified by ADMIXTURE analysis as A, B, C to determine their genetic affinities (SI14). Setting A as a present-day test population and B as either Stuttgart or BedouinB, we documented relatedness to C=(Loschbour or MA1) or C=(MA1 and Karitiana) or C=(MA1 or Han) (Extended Data Figs. 4, 5, 7). Setting C as a test population and (A, B) a pair from (Loschbour, Stuttgart, MA1) we documented differential relatedness to ancient populations (Extended Data Fig. 6). We computed $D$-statistics using transversion polymorphisms in whole genome sequence data to confirm robustness to ascertainment and ancient DNA damage (Extended Data Table 2).

**Minimum number of source populations for Europeans**

We used qpWave to study the minimum number of source populations for a designated set of Europeans (SI12). We use $f_4$-statistics of the form $X(l, r) = f_4(l_0, l; r_0, r)$ where $l_0,r_0$ are arbitrarily chosen “base” populations, and $l, r$ are other populations from two sets $L$ and $R$ respectively. 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 $r+1 \leq n$. We set $L$ to include Stuttgart, Loschbour, MA1, Onge, Karitiana, Mbouti and $R$ to include 23 modern European populations who fit the model of SI14 and had admixture proportions within the interval [0,1] for the method with minimal modeling assumptions (SI17).

**Admixture proportions for Stuttgart in the absence of a Near Eastern ancient genome**

We used Loschbour and BedouinB as surrogates for “Unknown hunter-gatherer” and Near Eastern (NE) farmer populations that contributed to Stuttgart (SI13). Ancient Near Eastern ancestry in Stuttgart is estimated by the $f_4$-ratio $f_4(\text{Outgroup}, X; \text{Loschbour}, \text{Stuttgart})/f_4(\text{Outgroup}, X; \text{Loschbour}, \text{NE})$. A complication is that BedouinB is a mixture of NE and African ancestry. We therefore subtracted the effects of African ancestry using estimates of the BedouinB African admixture proportion from ADMIXTURE (SI9) or ALDER.

**Admixture graph modeling**

We used ADMIXTUREGRAPH (version 3110) to model population relationships between Loschbour, Stuttgart, Onge, and Karitiana using Mbuti as an African outgroup. We assessed model fit using a block jackknife of differences between estimated and fitted $f$-statistics for the set of included populations (we expressed the fit as a Z score). We determined that a model failed if $|Z|>3$ for at least one $f$-statistic. A basic tree model failed and we manually amended the model to test all possible models with a single admixture event, which also failed. Further manual amendment to include 2 admixture events resulted in 8 successful models, only one of which could be amended to also fit MA1 as an additional constraint. We successfully fit both the Iceman and LaBrana into this model as simple clades and Motala12 as a 2-way mixture. We also fit present-day West Eurasians as clades, 2-way mixtures, or 3-way mixtures in this basic model, achieving a successful fit for a larger number of European populations (n=26) as 3-way mixtures. We estimated the individual admixture proportions from the fitted model parameters. To test if fitted parameters for different populations are consistent with each other, we jointly fit all pairs of populations $A$ and $B$ by modifying.
ADMIXTUREGRAPH to add a large constant (10,000) to the variance term $f_3(A, A, B)$. By doing this, we can safely ignore recent gene flow within Europe that affects statistics that include both $A$ and $B$.

**Ancestry estimates from $f_4$-ratios**

We estimate EEF ancestry using the $f_4$-ratio\cite{8,15} $f_4(Mbuti, Onge; Loschbour, European) / f_4(Mbuti, Onge; Loschbour, Stuttgart)$, which produces consistent results with ADMIXTUREGRAPH (SI14). We use $f_4(Stuttgart, Loschbour; Onge MA1) / f_4(Mbuti, MA1; Onge, Loschbour)$ to estimate Basal Eurasian admixture into Stuttgart. We use $f_4(Stuttgart, Loschbour; Onge Karitiana) / f_4(Mbuti, MA1; Onge, Loschbour)$ to estimate ANE mixture in Karitiana (Fig. 2B). We use $f_4(Test, Stuttgart; Karitiana, Onge) / f_4(MA1, Stuttgart; Karitiana, Onge)$ to lower bound ANE mixture into North Caucasian populations.

**MixMapper analysis**

We carried out MixMapper 2.0\cite{7} analysis, a semi-supervised admixture graph fitting technique. First, we infer a scaffold tree of populations without strong evidence of mixture relative to each other (Mbuti, Onge, Loschbour and MA1). We do not include European populations in the scaffold as all had significantly negative $f_3$-statistics indicating admixture. We then ran MixMapper to infer the relatedness of the other ancient and present-day samples, fitting them onto the scaffold as 2- or 3-way mixtures. The uncertainty in all parameter estimates is measured by block bootstrap resampling of the SNP set (100 replicates with 50 blocks).

**TreeMix analysis**

We applied TreeMix\cite{21} to Loschbour, Stuttgart, Motala12, and MA1\cite{3}, LaBrana\cite{2} and the Iceman\cite{1}, along with the present-day samples of Karitiana, Onge and Mbuti. We restricted the analysis to 265,521 Human Origins array sites after excluding any SNPs where there were no-calls in any of the studied individuals. The tree was rooted with Mbuti and standard errors were estimated using blocks of 500 SNPs. We repeated the analysis on whole-genome sequence data, rooting with Chimp and replacing Onge with Dai since we did not have Onge whole genome sequence data\cite{55}. We varied the number of migration events ($m$) between 0 and 5.

**Inferring admixture proportions with minimal modeling assumptions**

We devised a method to infer ancestry proportions from three ancestral populations (EEF, WHG, and ANE) without strong phylogenetic assumptions (SI17). We rely on 15 “non-West Eurasian” outgroups and study $f_4(European, Stuttgart; O_1, O_2)$ which equals $\alpha \beta f_4(Loschbour, Stuttgart; O_1, O_2) + \alpha (1-\beta) f_4(MA1, Stuttgart; O_1, O_2)$ if European has $1-\alpha$ ancestry from EEF and $\beta$, $1-\beta$ ancestry from WHG and ANE respectively. This defines a system of $\binom{15}{2} = 105$ equations with unknowns $\alpha \beta$, $\alpha (1-\beta)$, which we solve with least squares implemented in the function lsfit in R to obtain estimates of $\alpha$ and $\beta$. We repeated this computation 22 times dropping one chromosome at a time\cite{20} to obtain block jackknife\cite{67} estimates of the ancestry proportions and standard errors, with block size equal to the...
number of SNPs per chromosome. We assessed consistency of the inferred admixture proportions with those derived from the ADMIXTUREGRAPH model based on the number of standard errors between the two (Extended Data Table 1).

**Haplotype-based analyses**

We used RefinedIBD from BEAGLE 4\textsuperscript{27} with the settings \texttt{ibdtrim}=20 and \texttt{ibdwindow}=25 to study IBD sharing between Loschbour and Stuttgart and populations from the POPRES dataset\textsuperscript{69}. We kept all IBD tracts spanning at least 0.5 centimorgans (cM) and with a LOD score >3 (SI18). We also used ChromoPainter\textsuperscript{29} to study haplotype sharing between Loschbour and Stuttgart and present-day West Eurasian populations (SI19). We identified 495,357 SNPs that were complete in all individuals and phased the data using Beagle 4\textsuperscript{27} with parameters \texttt{phase-its}=50 and \texttt{impute-its}=10. We did not keep sites with missing data to avoid imputing modern alleles into the ancient individuals. We used both unlinked (-k 1000) and linked modes (estimating -n and -M by sampling 10\% of individuals). We combined ChromoPainter output for chromosomes 1-22 using ChromoCombine\textsuperscript{29}. We carried out a PCA of the co-ancestry matrix using fineSTRUCTURE\textsuperscript{29}. 

**Extended Data**

![Extended Data Figure 1.](image)

Photographs of analyzed ancient samples.

(A) Loschbour skull; (B) Stuttgart skull, missing the lower right M2 we sampled; (C) excavation at Kanaljorden in Motala, Sweden; (D) Motala 1 in situ.
Extended Data Figure 2.
Pairwise Sequential Markovian Coalescent (PSMC) analysis.
(A) Inference of population size as a function of time, showing a very small recent
population size over the most recent period in the ancestry of Loschbour (at least the last 5–
10 thousand years). (B) Inferred time since the most recent common ancestor from the
PSMC for chromosomes 20, 21, 22 (top to bottom); Stuttgart is plotted on top and
Loschbour at bottom.
Extended Data Figure 3.
ADMIXTURE analysis (K=2 to K=20).
Ancient samples (Loschbour, Stuttgart, Motala_merge, Motala12, MA1, and LaBrana) are at left.
Extended Data Figure 4.
ANE ancestry is present in both Europe and the Near East but WHG ancestry is restricted to Europe, which cannot be due to a single admixture event.
(x-axis) We computed the statistic \( f_4^{(Test, Stuttgart; MA1, Chimp)} \), which measures where MA1 shares more alleles with a test population than with Stuttgart. It is positive for most European and Near Eastern populations, consistent with ANE (MA1-related) gene flow into both regions. (y-axis) We computed the statistic \( f_4^{(Test, Stuttgart; Loschbour, Chimp)} \), which measures whether Loschbour shares more alleles with a test sample than with Stuttgart. Only European populations show positive values of this statistic, providing evidence of WHG (Loschbour-related) admixture only in Europeans.
Extended Data Figure 5.
MA1 is the best surrogate for ANE for which we have data.
Europeans share more alleles with MA1 than with Karitiana, as we see from the fact that in a plot of $f_4(\text{Test, BedouinB}; \text{MA1, Chimp})$ and $f_4(\text{Test, BedouinB}; \text{Karitiana, Chimp})$, the European cline deviates in the direction of MA1, rather than Karitiana (the slope is >1 and European populations are above the line indicating equality of these two statistics).
Extended Data Figure 6.
The differential relatedness of West Eurasians to Stuttgart (EEF), Loschbour (WHG), and MA1 (ANE) cannot be explained by two-way mixture.
We plot on a West Eurasian map the statistic $f_4(\text{Test, Chimp}; A_1, A_2)$, where $A_1$ and $A_2$ are a pair of the three ancient samples representing the three ancestral populations of Europe. (A) In both Europe and the Near East/Caucasus, populations from the south have more relatedness to Stuttgart than those from the north where ANE influence is also important. (B) Northern European populations share more alleles with Loschbour than with Stuttgart, as they have additional WHG ancestry beyond what was already present in EEF. (C) We observe a striking contrast between Europe west of the Caucasus and the Near East in degree of relatedness to WHG. In Europe, there is a much higher degree of allele sharing with Loschbour than with MA1, which we ascribe to the 60–80% WHG/(WHG+ANE) ratio in most Europeans that we report in SI14. In contrast, the Near East has no appreciable WHG ancestry but some ANE ancestry, especially in the northern Caucasus. (Jewish populations are marked with a square in this figure to assist in interpretation as their ancestry is often anomalous for their geographic regions.)
Extended Data Figure 7.
Evidence for Siberian gene flow into far northeastern Europe.
Some northeastern European populations (Chuvash, Finnish, Russian, Mordovian, Saami) share more alleles with Han Chinese than with other Europeans who are arrayed in a cline from Stuttgart to Lithuanians/Estonians in a plot of $f_d(Test, BedouinB; Han, Mbuti)$ against $f_d(Test, BedouinB; MA1, Mbuti)$.

Extended Data Table 1
West Eurasians genotyped on the Human Origins array and key f-statistics.
Confirmation of key findings on transversions and on whole genome sequence data.

Extended Data Table 2

Note: \(Z_{diff}\) is the number of standard errors of the difference between the lowest \(I_{1}\)-statistic over all reference pairs and the lowest \(I_{1}\)-statistic for a subset of reference pairs.

Abbreviations used: Stu: Stuttgart; Los: Los Angeles.

<table>
<thead>
<tr>
<th>Reference</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>African</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>European</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jewish</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Middle Eastern</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native American</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>North African</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>South African</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sub-Saharan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table Notes:**
- **A:** Arabic
- **B:** African
- **C:** Asian
- **D:** European
- **E:** Jewish
- **F:** Middle Eastern
- **G:** Native American
- **H:** North African
- **I:** South African
- **J:** Sub-Saharan
- **K:** European
- **L:** Jewish
- **M:** Middle Eastern
- **N:** Native American
- **O:** North African
- **P:** South African
- **Q:** Sub-Saharan

**Table Key:**
- **A:** Arabic
- **B:** African
- **C:** Asian
- **D:** European
- **E:** Jewish
- **F:** Middle Eastern
- **G:** Native American
- **H:** North African
- **I:** South African
- **J:** Sub-Saharan

**Table Values:**
- **Z_{diff}:** Number of standard errors of the difference between the lowest \(I_{1}\)-statistic over all reference pairs and the lowest \(I_{1}\)-statistic for a subset of reference pairs.

<table>
<thead>
<tr>
<th>Reference</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>African</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>European</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jewish</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Middle Eastern</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native American</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>North African</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>South African</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sub-Saharan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table Notes:**
- **A:** Arabic
- **B:** African
- **C:** Asian
- **D:** European
- **E:** Jewish
- **F:** Middle Eastern
- **G:** Native American
- **H:** North African
- **I:** South African
- **J:** Sub-Saharan

**Table Key:**
- **A:** Arabic
- **B:** African
- **C:** Asian
- **D:** European
- **E:** Jewish
- **F:** Middle Eastern
- **G:** Native American
- **H:** North African
- **I:** South African
- **J:** Sub-Saharan

**Table Values:**
- **Z_{diff}:** Number of standard errors of the difference between the lowest \(I_{1}\)-statistic over all reference pairs and the lowest \(I_{1}\)-statistic for a subset of reference pairs.

<table>
<thead>
<tr>
<th>Reference</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>African</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>European</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jewish</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Middle Eastern</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native American</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>North African</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>South African</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sub-Saharan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table Notes:**
- **A:** Arabic
- **B:** African
- **C:** Asian
- **D:** European
- **E:** Jewish
- **F:** Middle Eastern
- **G:** Native American
- **H:** North African
- **I:** South African
- **J:** Sub-Saharan

**Table Key:**
- **A:** Arabic
- **B:** African
- **C:** Asian
- **D:** European
- **E:** Jewish
- **F:** Middle Eastern
- **G:** Native American
- **H:** North African
- **I:** South African
- **J:** Sub-Saharan

**Table Values:**
- **Z_{diff}:** Number of standard errors of the difference between the lowest \(I_{1}\)-statistic over all reference pairs and the lowest \(I_{1}\)-statistic for a subset of reference pairs.
Either by single population analysis or co-fitting population pairs and averaging over fits, minimal assumptions are from SI17. The estimates from the full modeling are from SI14.

Admixtures proportions for European populations. The estimates from the model with minimal assumptions are from SI17. The estimates from the full modeling are from SI14 either by single population analysis or co-fitting population pairs and averaging over fits (these averages are the results plotted in Fig. 2B). Populations that do not fit the models are not reported.
### Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

### Authors


Nature. Author manuscript; available in PMC 2015 March 18.

Affiliations

1Department of Genetics, Harvard Medical School, Boston, MA, 02115, USA 2Broad Institute of Harvard and MIT, Cambridge, MA, 02142, USA 3Institute for Archaeological Sciences, University of Tübingen, Tübingen, 72074, Germany 4Max Planck Institute for Evolutionary Anthropology, Leipzig, 04103, Germany 5Johannes Gutenberg University Mainz, Institute of Anthropology, Mainz, D-55128, Germany 6Department of Genome Sciences, University of Washington, Seattle, WA, 98195, USA 7Department of Integrative Biology, University of California, Berkeley, CA, 94720-3140, USA 8Department of Mathematics and Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA 9Archaeological Research Laboratory, Stockholm University, 114 18, Sweden 10Key Laboratory of Vertebrate Evolution and Human Origins of Chinese Academy of Sciences, IVPP, CAS, Beijing, 100049, China 11Australian Centre for Ancient DNA, School of Earth and Environmental Sciences, University of Adelaide, Adelaide, South Australia, SA 5005, Australia 12The Cultural Heritage Foundation, Västerås, 722 12, Sweden 13National Museum of Natural History, L-2160, Luxembourg 14National Center of Archaeological Research, National Museum of History and Art, L-2345, Luxembourg 15Department of Paleoanthropology, Senckenberg Center for Human Evolution and Paleoenvironment, University of Tübingen, Tübingen, D-72070, Germany 16State Office for Cultural Heritage Management Baden-Württemberg, Osteology, Konstanz, D-78467, Germany 17Center for Global Health and Child Development, Kisumu, 40100, Kenya 18Institutes of Evolution, Immunology and Infection Research, School of Biological Sciences, University of Edinburgh, Edinburgh, EH9 3JT, UK 19Biochemistry Department, Faculty of Medicine, Sultan Qaboos University, Alkhod, Muscat, 123, Oman 20Laboratorio de Genética Molecular Poblacional, Instituto Multidisciplinario de Biología Celular (IMBICE), CCT-CONICET & CICPBA, La Plata, B1906APO, Argentina 21Research Centre for Medical Genetics, Moscow, 115478, Russia 22Vavilov Institute for General Genetics, Moscow, 119991, Russia 23Escuela de Biología, Universidad de Costa Rica, San José, 2060, Costa Rica 24Institute of
Biology, Research group GENMOL, Universidad de Antioquia, Medellín, Colombia
25Rambam Health Care Campus, Haifa, 31096, Israel 28Department of Medical Genetics and Szentagothai Research Center, University of Pécs, Pécs, H-7624 Hungary
27Al Akhawayn University in Ifrane (AUI), School of Science and Engineering, Ifrane, 53000, Morocco 28Forensic Genetics Laboratory, Institute of Legal Medicine, Università Cattolica del Sacro Cuore, Rome, 00168, Italy
29Department of Zoology, University of Oxford, Oxford, OX1 3PS, UK 30Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN, UK
31Laboratorio di Genetica Molecolare, IRCCS Associazione Oasi Maria SS, Troina, 94018, Italy 32Belgorod State University, Belgorod, 308015, Russia 33Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, M5G 1L5, Canada 34Servicio de Huellas Digitales Genéticas, School of Pharmacy and Biochemistry, Universidad de Buenos Aires, 1113 CABA, Argentina 35Institute of Cytology and Genetics, Siberian Branch of Russian Academy of Sciences, Novosibirsk, 630090, Russia 36Institute of Linguistics, University of Bern, Bern, CH-3012, Switzerland 37Laboratory of Human Molecular Genetics, Institute of Molecular and Cellular Biology, Russian Academy of Science, Siberian Branch, Novosibirsk, 630090, Russia 38Anthropologie Moléculaire et Imagerie de Synthèse, CNRS UMR 5288, Université Paul Sabatier Toulouse III, Toulouse, 31000, France 39Yakut Research Center of Complex Medical Problems and North-Eastern Federal University, Yakutsk, 677010, Russia 40Department of Human Genetics, University of Chicago, Chicago, IL, 60637, USA 41ARL Division of Biotechnology, University of Arizona, Tucson, AZ, 85721, USA 42Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY, 11794, USA 43Department of Clinical Science, University of Bergen, Bergen, 5021, Norway 44NextBio, part of Illumina, Santa Clara, CA, USA 95050 45Dept. of Medical Genetics, National Human Genome Center, Medical University Sofia, Sofia, 1431, Bulgaria 46Institute of Biochemistry and Genetics, Ufa Research Centre, Russian Academy of Sciences, Ufa, 450054, Russia 47Department of Genetics and Fundamental Medicine, Bashkir State University, Ufa, 450074, Russia 48College of Medicine, University of Illinois at Chicago, Chicago, IL, 60607, USA 49Division of Biological Anthropology, University of Cambridge, Cambridge, United Kingdom CB2 1QH 50Department of Human and Medical Genetics, Vilnius University, Vilnius, LT-08661, Lithuania 51Estonian Biocentre, Evolutionary Biology group, Tartu, 51010, Estonia 52Translational Medicine and Neurogenetics, Institut de Génétique et de Biologie Moléculaire et Cellulaire, Illkirch, 67404, France 53Gladstone Institutes, San Francisco, CA, 94158, USA 54Department of Evolutionary Biology, University of Tartu, Tartu, 51010, Estonia 55Centro de Investigaciones Biomédicas de Guatemala, Ciudad de Guatemala, Guatemala 56Research Department, 23andMe, Inc. Mountain View, CA, 94043, USA 57Cultural Anthropology Program, University of Oulu, Oulu, 90014, Finland 58Department of Biochemistry, Muhimbili University of Health and Allied Sciences, Dar es Salaam, Tanzania 59Research Institute of Health, North-Eastern Federal University, Yakutsk, 677000, Russia 60Dipartimento di Fisica e Chimica, Università di Palermo, Palermo, 90128, Italy 61Instituto de Alta Investigación,
Acknowledgments

We are grateful to Cynthia Beall, Neil Bradman, Amha Gebremedhin, Damian Labuda, Mari Nelis and Anna Di Rienzo for sharing DNA samples; to Detlef Weigel, Christa Lanz, Verena Schnümmen, Peter Bauer and Olaf Riess for support and access to DNA sequencing facilities; to Philip Johnson for advice on contamination estimation; to Garrett Hellenthal for help with the ChromoPainter software; and to Pontus Skoglund for sharing graphics software. We thank Kenneth Nordveldt for alerting us to newly discovered Y-chromosome SNPs. We downloaded the POPRES data from dbGaP at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000145.v4.p2 through dbGaP accession number phs000145.v1.p2. We thank all the volunteers who donated DNA. We thank the staff of the Unità Operativa Complessa di Medicina Trasfusionale, Azienda
Ospedaliera Umberto I, Siracusa, Italy for assistance in sample collection; and The National Laboratory for the Genetics of Israeli Populations for facilitating access to DNA. We thank colleagues at the Applied Genomics at the Children’s Hospital of Philadelphia, especially Hakon Hakonarson, Cecilia Kim, Kelly Thomas, and Cuiping Hou, for genotyping samples on the Human Origins array. JK is grateful for support from DFG grant # KR 4015/1-1, the Carl-Zeiss Foundation and the Baden Württemberg Foundation. SP, GR, QF, CF, KP, SC and JKe acknowledge support from the Presidential Innovation Fund of the Max Planck Society. GR was supported by an NSERC fellowship. JGS acknowledges use of the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by NSF grant number OCI-1053575. EB and OB were supported by RFBR grants 13-06-00670, 13-04-01711, 13-04-90420 and by the Molecular and Cell Biology Program of the Presidium, Russian Academy of Sciences. BM was supported by grants OTKA 73430 and 103983. ASaj was supported by a Finnish Professorposto (Paulo Foundation) Grant. The Lithuanian sampling was supported by the LITGEN project (VP1-3.1-SMM-07-K-01-013), funded by the European Social Fund under the Global Grant Measure. AS was supported by Spanish grants SAF2011-26983 and EM 2012/045. OU was supported by Ukrainian SFFS grant F53.4/071. SAT was supported by an Indian CSIR Network Project (GENESIS: BSC0121). LS was supported by an Indian CSIR Bhatnagar Fellowship.

References


Figure 1. Map of West Eurasian populations and Principal Component Analysis
(a) Geographical locations of analyzed samples, with color coding matching the PCA. We show all sampling locations for each population, which results in multiple points for some
(e.g., Spain). (b) PCA on all present-day West Eurasians, with ancient and selected eastern non-African samples projected. European hunter-gatherers fall beyond present-day Europeans in the direction of European differentiation from the Near East. Stuttgart clusters with other Neolithic Europeans and present-day Sardinians. MA1 falls outside the variation of present-day West Eurasians in the direction of southern-northern differentiation along dimension 2.
Figure 2. Modeling of West Eurasian population history

(a) A three-way mixture model that is a fit to the data for many populations. Present-day samples are colored in blue, ancient in red, and reconstructed ancestral populations in green. Solid lines represent descent without mixture, and dashed lines represent admixture. We print mixture proportions and one standard error for the two mixtures relating the highly divergent ancestral populations. (We do not print the estimate for the “European” population as it varies depending on the population). (b) We plot the proportions of ancestry from each of three inferred ancestral populations (EEF, ANE and WHG).
**Table 1**

Lowest $f_3$-statistics for each West Eurasian population

<table>
<thead>
<tr>
<th>Ref$_1$</th>
<th>Ref$_2$</th>
<th>Target for which these two references give the lowest $f_3(X; \text{Ref}_1, \text{Ref}_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>WHG</td>
<td>EEF</td>
<td>Sardinian***</td>
</tr>
<tr>
<td>WHG</td>
<td>Near East</td>
<td>Basque, Belarusian, Czech, English, Estonian, Finnish, French_South, Icelandic, Lithuanian, Mordovian, Norwegian, Orcadian, Scottish, Spanish, Spanish_North, Ukrainian</td>
</tr>
<tr>
<td>WHG</td>
<td>Siberian</td>
<td>Russian</td>
</tr>
<tr>
<td>EEF</td>
<td>ANE</td>
<td>Abkhasian***, Albanian, Ashkenazi_Jew****, Bergamo, Bulgarian, Chechen****, Croatian, Cypriot****, Druze**, French, Greek, Hungarian, Lezgin, Maltese, Sicilian, Turkish_Jew, Tuscan</td>
</tr>
<tr>
<td>EEF</td>
<td>Native American</td>
<td>Adygei, Balkar, Iranian, Kumyk, North_Ossetian, Turkish</td>
</tr>
<tr>
<td>EEF</td>
<td>South Asian</td>
<td>Armenian, Georgian****, Georgian_Jew*, Iranian_Jew***, Iraqi_Jew***</td>
</tr>
</tbody>
</table>

Note: WHG = Loschbour or LaBraña; EEF=Stuttgart; ANE=MA1; Native American=Piapoco; African=Esan, Gambian, or Kgalagadi; South Asian=GujaratiC or Vishwabrahmin. Statistics are negative with Z<−4 unless otherwise noted: † (positive) or *, **, ***, ****, to indicate Z less than 0, −1, −2, and −3 respectively. The complete list of statistics can be found in Extended Data Table 1.

*Nature.* Author manuscript; available in PMC 2015 March 18.