ARTICLE

doi:10.1038/nature16152

Genome-wide patterns of selection in 230 ancient Eurasians

Iain Mathieson[1], Iosif Lazaridis[1,2], Nadin Rohland[1,2], Swapan Mallick[1,2,3], Nick Patterson[2], Songül Alpaslan Roodenberg[4], Eadaoin Harney[1,3], Kristin Stewardson[1,3], Daniel Fernandes[5], Mario Novak[5,6], Kendra Sirak[5,7], Cristina Gamba[5,8]†, Eppie R. Jones[8], Bastien Llamas[9], Stanislav Dryomov[10,11], Joseph Pickrell[1]†, Juan Luís Arsuaga[12,13], José María Bermúdez de Castro[14], Eudald Carbonell[15,16], Fokke Gerritsen[17], Aleksandr Khokhlov[18], Pavel Kuznetsov[18], Marina Lozano[15,16], Harald Meller[19], Oleg Mochalov[18], Vyacheslav Moiseyev[20], Manuel A. Rojo Guerra[21], Jacob Roodenberg[22], Josep Maria Vergès[15,16], Johannes Krause[23,24], Alan Cooper[9], Kurt W. Alt[19,25,26], Dorcas Brown[27], David Anthony[27], Carles Lalueza-Fox[28], Wolfgang Haak[9,23]*, Ron Pinhasi[5]* & David Reich[1,2,3]*

Ancient DNA makes it possible to observe natural selection directly by analysing samples from populations before, during and after adaptation events. Here we report a genome-wide scan for selection using ancient DNA, capitalizing on the largest ancient DNA data set yet assembled: 230 West Eurasians who lived between 6500 and 300 BC, including 163 with newly reported data. The new samples include, to our knowledge, the first genome-wide ancient DNA from Anatolian Neolithic farmers, whose genetic material we obtained by extracting from petrous bones, and who we show were members of the population that was the source of Europe's first farmers. We also report a transect of the steppe region in Samara between 5600 and 300 BC, which allows us to identify admixture into the steppe from at least two external sources. We detect selection at loci associated with diet, pigmentation and immunity, and two independent episodes of selection on height.

The arrival of farming in Europe around 8,500 years ago necessitated adaptation to new environments, pathogens, diets and social organizations. While indirect evidence of this adaptation can be detected in patterns of genetic variation in present-day people[1], these patterns are only echoes of past events, which are difficult to date and interpret, and are often confounded by neutral processes. Ancient DNA provides a direct way to study these patterns, and should be a transformative technology for studies of selection, just as it has transformed studies of human pre-history. Until now, however, the large sample sizes required to detect selection have meant that studies of ancient DNA have concentrated on characterizing effects at parts of the genome already believed to have been affected by selection[2–5].

Genome-wide ancient DNA from West Eurasia

We assembled genome-wide data from 230 ancient individuals from West Eurasia dated to between 6500 and 300 BC (Fig. 1a, Extended Data Table 1, Supplementary Data Table 1 and Supplementary Information section 1). To obtain this data set, we combined published data from 67 samples from relevant periods and cultures[4–6], with 163 samples for which we report new data, of which 83 have, to our knowledge, never previously been analysed (the remaining 80 samples include 67 whose targeted single nucleotide polymorphism (SNP) coverage we tripled from 390,000 ('390k capture') to 1,240,000 ('1240k capture')[7]; and 13 with shotgun data for which we generated new data using our targeted enrichment strategy[3,8]). The 163 samples for which we report new data are drawn from 270 distinct individuals who we screened for evidence of authentic DNA[7]. We used in-solution hybridization with synthesized oligonucleotide probes to enrich promising libraries for the targeted SNPs (Methods). The targeted sites include nearly all SNPs on the Affymetrix Human Origins and Illumina 610-Quad arrays, 49,711 SNPs on chromosome X, 32,681 SNPs on chromosome Y, and 47,384 SNPs with evidence of functional importance. We merged libraries from the same individual and filtered out samples with low coverage or evidence of contamination to obtain the final set of individuals. The 1240k capture gives access to genome-wide data from ancient samples with small fractions of human DNA and increases efficiency by targeting sites in the human genome that will actually be analysed. The effectiveness of the approach can be seen by comparing our results to the largest previously published ancient DNA study, which used a shotgun sequencing strategy[5]. Our median coverage on analysed SNPs

[1]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [2]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. [3]Howard Hughes Medical Institute, Harvard Medical School, Boston, Massachusetts 02115, USA. [4]Independent researcher, Santpoort-Noord, The Netherlands. [5]School of Archaeology and Earth Institute, Belfield, University College Dublin, Dublin 4, Ireland. [6]Institute for Anthropological Research, Zagreb 10000, Croatia. [7]Department of Anthropology, Emory University, Atlanta, Georgia 30322, USA. [8]Smurfit Institute of Genetics, Trinity College Dublin, Dublin 2, Ireland. [9]Australian Centre for Ancient DNA, School of Biological Sciences & Environment Institute, University of Adelaide, Adelaide, South Australia 5005, Australia. [10]Laboratory of Human Molecular Genetics, Institute of Molecular and Cellular Biology, Siberian Branch of the Russian Academy of Sciences, Novosibirsk 630090, Russia. [11]Department of Paleolithic Archaeology, Institute of Archaeology and Ethnography, Siberian Branch of the Russian Academy of Sciences, Novosibirsk 630090, Russia. [12]Centro Mixto UCM-ISCIII de Evolución y Comportamiento Humanos, 28040 Madrid, Spain. [13]Departamento de Paleontología, Facultad Ciencias Geológicas, Universidad Complutense de Madrid, 28040 Madrid, Spain. [14]Centro Nacional de Investigacíon sobre Evolución Humana (CENIEH), 09002 Burgos, Spain. [15]IPHES. Institut Català de Paleoecologia Humana i Evolució Social, Campus Sescelades-URV, 43007 Tarragona, Spain. [16]Area de Prehistoria, Universitat Rovira i Virgili (URV), 43002 Tarragona, Spain. [17]Netherlands Institute in Turkey, Istiklal Caddesi, Nur-i Ziya Sokak 5, Beyoğlu 34433, Istanbul, Turkey. [18]Volga State Academy of Social Sciences and Humanities, Samara 443099, Russia. [19]State Office for Heritage Management and Archaeology Saxony-Anhalt and State Museum of Prehistory, D-06114 Halle, Germany. [20]Peter the Great Museum of Anthropology and Ethnography (Kunstkamera) RAS, St Petersburg 199034, Russia. [21]Department of Prehistory and Archaeology, University of Valladolid, 47002 Valladolid, Spain. [22]The Netherlands Institute for the Near East, Leiden RA-2300, the Netherlands. [23]Max Planck Institute for the Science of Human History, D-07745 Jena, Germany. [24]Institute for Archaeological Sciences, University of Tübingen, D-72070 Tübingen, Germany. [25]Danube Private University, A-3500 Krems, Austria. [26]Institute for Prehistory and Archaeological Science, University of Basel, CH-4003 Basel, Switzerland. [27]Anthropology Department, Hartwick College, Oneonta, New York 13820, USA. [28]Institute of Evolutionary Biology (CSIC-Universitat Pompeu Fabra), 08003 Barcelona, Spain. †Present addresses: Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5–7, 1350 Copenhagen, Denmark (C.G.); New York Genome Center, New York, New York 10013, USA (J.P.).
*These authors contributed equally to this work.

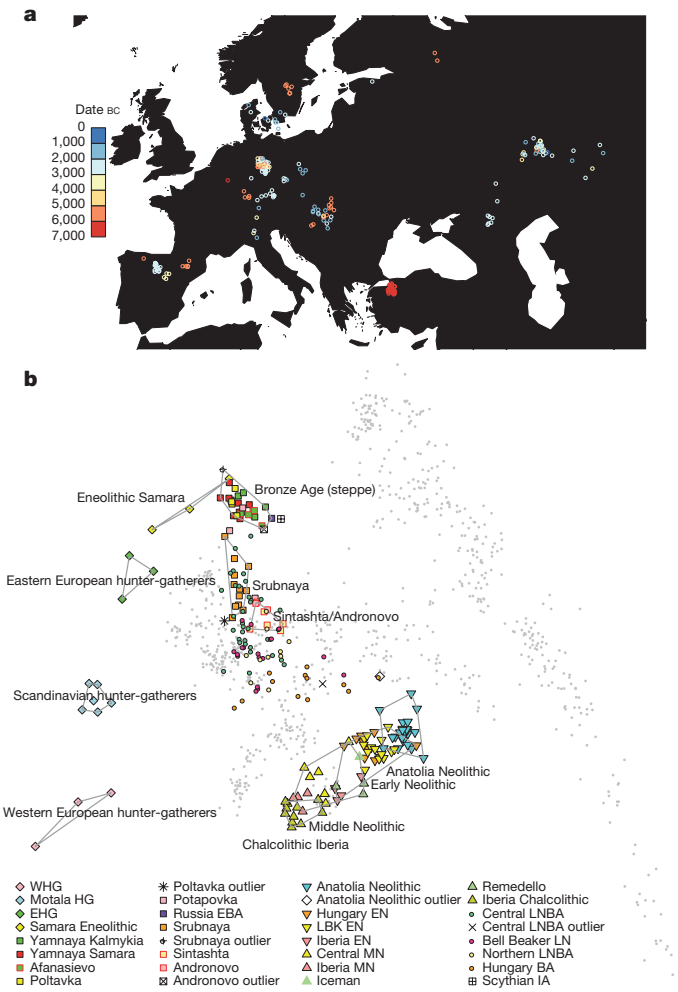


Figure 1 | Population relationships of samples. a, Locations colour-coded by date, with a random jitter added for visibility (8 Afanasievo and Andronovo samples lie further east and are not shown). **b**, Principal component analysis of 777 modern West Eurasian samples (grey), with 221 ancient samples projected onto the first two principal component axes and labelled by culture. E/M/LN, Early/Middle/Late Neolithic; LBK, *Linearbandkeramik*; E/WHG, Eastern/Western hunter-gatherer; EBA, Early Bronze Age; IA, Iron Age; LNBA, Late Neolithic and Bronze Age.

is approximately fourfold higher even while the mean number of reads generated per sample is 36-fold lower (Extended Data Fig. 1).

Insight into population transformations

To learn about the genetic affinities of the archaeological cultures for which genome-wide data are reported for the first time here, we studied either 1,055,209 autosomal SNPs when analysing 230 ancient individuals alone, or 592,169 SNPs when co-analysing them with 2,345 present-day individuals genotyped on the Human Origins array[4]. We removed 13 samples either as outliers in ancestry relative to others of the same archaeologically determined culture, or first-degree relatives (Supplementary Data Table 1).

Our sample of 26 Anatolian Neolithic individuals represents the first genome-wide ancient DNA data from the eastern Mediterranean. Our success at analysing such a large number of samples is due to the fact that in the case of 21 of the successful samples, we obtained DNA from the inner ear region of the petrous bone[9], which has been shown to increase the amount of DNA obtained by up to two orders of magnitude relative to teeth[3]. Principal component (PCA) and ADMIXTURE[10] analyses show that the Anatolian Neolithic samples do not resemble any present-day near-Eastern populations but are shifted towards Europe, clustering with early European farmers (EEF) from Germany, Hungary and Spain[7] (Fig. 1b and Extended Data Fig. 2). Further evidence that the Anatolian Neolithic and EEF were related comes from the high frequency (47%; $n = 15$) of Y-chromosome haplogroup G2a typical of ancient EEF samples[7] (Supplementary Data Table 1), and the low fixation index (F_{ST}; 0.005 – 0.016) between Neolithic Anatolians and EEF (Supplementary Data Table 2). These results support the hypothesis[7] of a common ancestral population of EEF before their dispersal along distinct inland/central European and coastal/Mediterranean routes. The EEF are slightly more shifted to Europe in the PCA than are the Anatolian Neolithic (Fig. 1b) and have significantly more admixture from Western hunter-gatherers (WHG), as shown by f_4-statistics ($|Z| > 6$ standard errors from 0) and negative f_3-statistics ($|Z| > 4$)[11] (Extended Data Table 2). We estimate that the EEF have 7–11% more WHG admixture than their Anatolian relatives (Extended Data Fig. 2, Supplementary Information section 2).

The Iberian Chalcolithic individuals from El Mirador cave are genetically similar to the Middle Neolithic Iberians who preceded them (Fig. 1b and Extended Data Fig. 2), and have more WHG ancestry than their Early Neolithic predecessors[7] ($|Z| > 10$) (Extended Data Table 2). However, they do not have a significantly different proportion of WHG ancestry (we estimate 23–28%) than the Middle Neolithic Iberians (Extended Data Fig. 2). Chalcolithic Iberians have no evidence of steppe ancestry (Fig. 1b and Extended Data Fig. 2), in contrast to central Europeans of the same period[5,7]. Thus, the steppe-related ancestry that is ubiquitous across present-day Europe[4,7] arrived in Iberia later than in central Europe (Supplementary Information section 2).

To understand population transformations in the Eurasian steppe, we analysed a time transect of 37 samples from the Samara region spanning ~5600–1500 BC and including the Eastern hunter-gatherer (EHG), Eneolithic, Yamnaya, Poltavka, Potapovka and Srubnaya cultures. Admixture between populations of Near Eastern ancestry and the EHG[7] began as early as the Eneolithic (5200–4000 BC), with some individuals resembling EHG and some resembling Yamnaya (Fig. 1b and Extended Data Fig. 2). The Yamnaya from Samara and Kalmykia, the Afanasievo people from the Altai (3300–3000 BC), and the Poltavka Middle Bronze Age (2900–2200 BC) population that followed the Yamnaya in Samara are all genetically homogeneous, forming a tight 'Bronze Age steppe' cluster in PCA (Fig. 1b), sharing predominantly R1b Y chromosomes[5,7] (Supplementary Data Table 1), and having 48–58% ancestry from an Armenian-like Near Eastern source (Extended Data Table 2) without additional Anatolian Neolithic or EEF ancestry[7] (Extended Data Fig. 2). After the Poltavka period, population change occurred in Samara: the Late Bronze Age Srubnaya have ~17% Anatolian Neolithic or EEF ancestry (Extended Data Fig. 2). Previous work documented that such ancestry appeared east of the Urals beginning at least by the time of the Sintashta culture, and suggested that it reflected an eastward migration from the Corded Ware peoples of central Europe[5]. However, the fact that the Srubnaya also had such ancestry indicates that the Anatolian Neolithic or EEF ancestry could have come into the steppe from a more eastern source. Further evidence that migrations originating as far west as central Europe may not have had an important impact on the Late Bronze Age steppe comes from the fact that the Srubnaya possess exclusively ($n = 6$) R1a Y chromosomes (Supplementary Data Table 1), and four of them (and one Poltavka male) belonged to haplogroup R1a-Z93, which is common in central/south Asians[12], very rare in present-day Europeans, and absent in all ancient central Europeans studied to date.

Twelve signals of selection

To study selection, we created a data set of 1,084,781 autosomal SNPs in 617 samples by merging 213 ancient samples with genome-wide sequencing data from four populations of European ancestry from the 1,000 Genomes Project[13]. Most present-day Europeans can be modelled as a mixture of three ancient populations related to Western hunter-gatherers (WHG), early European farmers (EEF) and steppe pastoralists (Yamnaya)[4,7], and so to scan for selection, we divided our

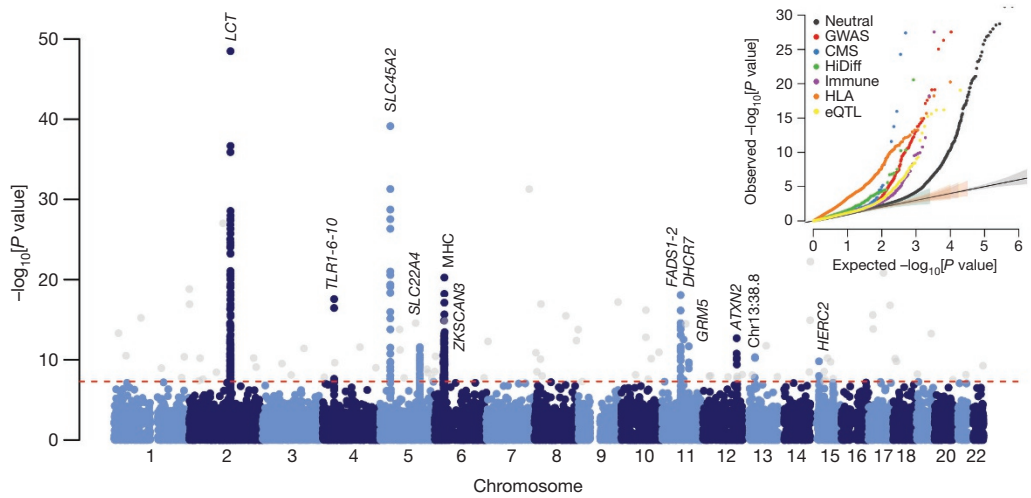


Figure 2 | Genome-wide scan for selection. GC-corrected $-\log_{10} P$ value for each marker (Methods). The red dashed line represents a genome-wide significance level of 0.5×10^{-8}. Genome-wide significant points filtered because there were fewer than two other genome-wide significant points within 1 Mb are shown in grey. Inset, quantile–quantile plots for corrected $-\log_{10} P$ values for different categories of potentially functional SNPs (Methods). Truncated at $-\log_{10}[P \text{ value}] = 30$. All curves are significantly different from neutral expectation. CMS, composite of multiple signals selection hits; HiDiff, highly differentiated between HapMap populations; Immune, immune-related; HLA, human leukocyte antigen type tag SNPs; eQTL, expression quantitative trait loci (see Methods).

samples into three groups based on which of these populations they clustered with most closely (Fig. 1b and Extended Data Table 1). We estimated mixture proportions for the present-day European ancestry populations and tested every SNP to evaluate whether its present-day frequencies were consistent with this model. We corrected for test statistic inflation by applying a genomic control correction analogous to that used to correct for population structure in genome-wide association studies[14]. Of approximately one million non-monomorphic autosomal SNPs, the ~50,000 in the set of potentially functional SNPs were significantly more inconsistent with the model than neutral SNPs (Fig. 2), suggesting pervasive selection on polymorphisms of functional importance. Using a conservative significance threshold of $P = 5.0 \times 10^{-8}$, and a genomic control correction of 1.38, we identified 12 loci that contained at least three SNPs achieving genome-wide significance within 1 Mb of the most associated SNP (Fig. 2, Extended Data Table 3, Extended Data Fig. 3 and Supplementary Data Table 3).

The strongest signal of selection is at the SNP (rs4988235) responsible for lactase persistence in Europe[15,16]. Our data (Fig. 3) strengthens previous reports that an appreciable frequency of lactase persistence in Europe only dates to the last 4,000 years[3,5,17]. The allele's earliest appearance in the dataset is in a central European Bell Beaker sample (individual I0112) dated to between 2450 and 2140 BC. Two other independent signals related to diet are located on chromosome 11 near *FADS1* and *DHCR7*. *FADS1* and *FADS2* are involved in fatty acid metabolism, and variation at this locus is associated with plasma lipid and fatty acid concentration[18]. The selected allele of the most significant SNP (rs174546) is associated with decreased triglyceride levels[18]. This locus has experienced independent selection in non-European populations[13,19,20] and is likely to be a critical component of adaptation to different diets. Variants at *DHCR7* and *NADSYN1* are associated with circulating vitamin D levels[21] and the most associated SNP in our analysis, rs7940244, is highly differentiated across closely related northern European populations[22,23], suggesting selection related to variation in dietary or environmental sources of vitamin D.

Two signals have a potential link to coeliac disease. One occurs at the ergothioneine transporter *SLC22A4* that is hypothesized to have experienced a selective sweep to protect against ergothioneine deficiency in agricultural diets[24]. Common variants at this locus are associated with increased risk for ulcerative colitis, coeliac disease, and irritable bowel disease and may have hitchhiked to high frequency as a result of this sweep[24–26]. However, the specific variant (rs1050152, L503F) that was thought to be the target did not reach high frequency until relatively recently (Extended Data Fig. 4). The signal at *ATXN2/SH2B3*—also associated with coeliac disease[25]—shows a similar pattern (Extended Data Fig. 4).

The second strongest signal in our analysis is at the derived allele of rs16891982 in *SLC45A2*, which contributes to light skin pigmentation and is almost fixed in present-day Europeans but occurred at much lower frequency in ancient populations. In contrast, the derived allele of *SLC24A5* that is the other major determinant of light skin pigmentation in modern Europe (and that is not significant in the genome-wide scan for selection) appears fixed in the Anatolian Neolithic, suggesting that its rapid increase in frequency to around 0.9 in Early Neolithic Europe was mostly due to migration (Extended Data Fig. 4). Another pigmentation signal is at *GRM5*, where SNPs are associated with pigmentation possibly through a regulatory effect on nearby *TYR*[27]. We also find evidence of selection for the derived allele of rs12913832 at *HERC2/OCA2*, which is at 100% frequency in the European hunter-gatherers we analysed, and is the primary determinant of light eye colour in present-day Europeans[28,29]. In contrast to the other loci, the range of frequencies in modern populations is within that of ancient populations (Fig. 3). The frequency increases with higher latitude, suggesting a complex pattern of environmental selection.

The *TLR1–TLR6–TLR10* gene cluster is a known target of selection in Europe, possibly related to resistance to leprosy, tuberculosis or other mycobacteria[30–32]. There is also a strong signal of selection at the major histocompatibility complex (MHC) on chromosome 6. The strongest signal is at rs2269424 near the genes *PPT2* and *EGFL8*, but there are at least six other apparently independent signals in the MHC (Extended Data Fig. 3); and the entire region is significantly more associated than the genome-wide average (residual inflation of 2.07 in the region on chromosome 6 between 29–34 Mb after genome-wide genomic control correction). This could be the result of multiple sweeps, balancing selection, or increased drift as a result of background selection reducing effective population size in this gene-rich region.

We find a surprising result in six Scandinavian hunter-gatherers (SHG) from Motala in Sweden. In three of six samples, we observe the haplotype carrying the derived allele of rs3827760 in the *EDAR*

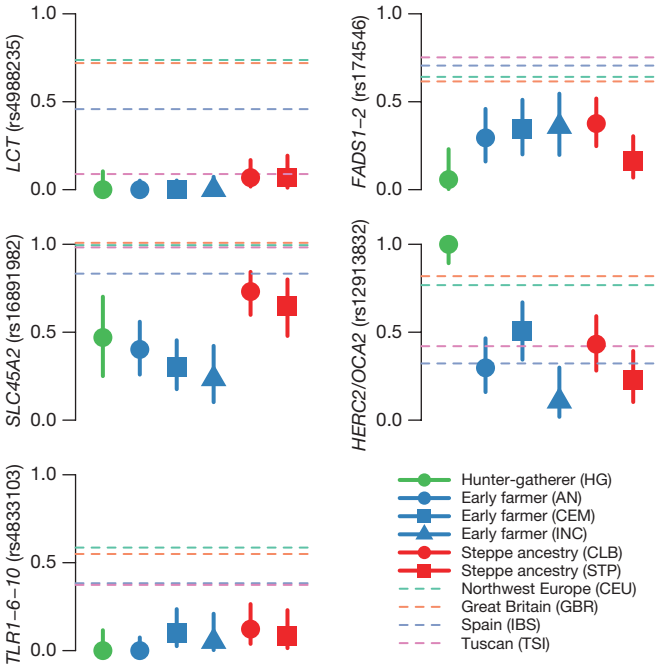


Figure 3 | Allele frequencies for five genome-wide significant signals of selection. Dots and solid lines show maximum likelihood frequency estimates and a 1.9-log-likelihood support interval for the derived allele frequency in each ancient population. Horizontal dashed lines show allele frequencies in the four modern 1000 Genomes populations. AN, Anatolian Neolithic; HG, hunter-gatherer; CEM, central European Early and Middle Neolithic; INC, Iberian Neolithic and Chalcolithic; CLB, central European Late Neolithic and Bronze Age; STP, steppe; CEU, Utah residents with northern and western European ancestry; IBS, Iberian population in Spain. The hunter-gatherer, early farmer and steppe ancestry classifications correspond approximately to the three populations used in the genome-wide scan with some differences (See Extended Data Table 1 for details).

gene (Extended Data Fig. 5), which affects tooth morphology and hair thickness[33,34], has been the target of a selective sweep in East Asia[35], and today is at high frequency in East Asians and Native Americans. The *EDAR* derived allele is largely absent in present-day Europe, except in Scandinavia, plausibly owing to Siberian movements into the region millennia after the date of the Motala samples. The SHG have no evidence of East Asian ancestry[4,7], suggesting that the *EDAR* derived allele may not have originated in the main ancestral population of East Asians as previously suggested[35]. A second surprise is that, unlike closely related WHGs, the Motala samples have predominantly derived pigmentation alleles at *SLC45A2* and *SLC24A5*.

Evidence of selection on height

We also tested for selection on complex traits. The best-documented example of this process in humans is height, for which the differences between northern and southern Europe have been driven by selection[36]. To test for this signal in our data, we used a statistic that tests whether trait-affecting alleles are both highly correlated and more differentiated, compared to randomly sampled alleles[37]. We predicted genetic heights for each population and applied the test to all populations together, as well as to pairs of populations (Fig. 4). Using 180 height-associated SNPs[38] (restricted to 169 for which we successfully obtained genotypes from at least two individuals from each population), we detect a significant signal of directional selection on height ($P = 0.002$). Applying this to pairs of populations allows us to detect two independent signals. First, the Iberian Neolithic and Chalcolithic samples show selection for reduced height relative to both the Anatolian Neolithic ($P = 0.042$) and the central European Early and Middle Neolithic ($P = 0.003$). Second, we detect a signal for increased height in the steppe populations ($P = 0.030$ relative to the central European Early and Middle Neolithic). These results suggest that the modern South–North gradient in height across Europe is due to both increased steppe ancestry in northern populations, and selection for decreased height in Early Neolithic migrants to southern Europe. We did not observe any other significant signals of polygenetic selection in five other complex traits we tested: body mass index[39] ($P = 0.20$), waist-to-hip ratio[40] ($P = 0.51$), type 2 diabetes[41] ($P = 0.37$), inflammatory bowel disease[26] ($P = 0.17$) and lipid levels[18] ($P = 0.50$).

Future studies of selection with ancient DNA

Our results, which take advantage of the massive increase in sample size enabled by optimized techniques for sampling from the inner-ear regions of the petrous bone, as well as in-solution enrichment methods for targeted SNPs, show how ancient DNA can be used to perform a genome-wide scan for selection. Our results also directly document selection on loci related to pigmentation, diet and immunity, painting a picture of populations adapting to settled agricultural life at high latitudes. For most of the signals, allele frequencies of modern Europeans are outside the range of any ancient populations, indicating that phenotypically, Europeans of 4,000 years ago were different in important respects from Europeans today, despite having overall similar ancestry. An important direction for future research is to increase the sample size for European selection scans (Extended Data Fig. 6), and to apply this approach to regions beyond Europe and to other species.

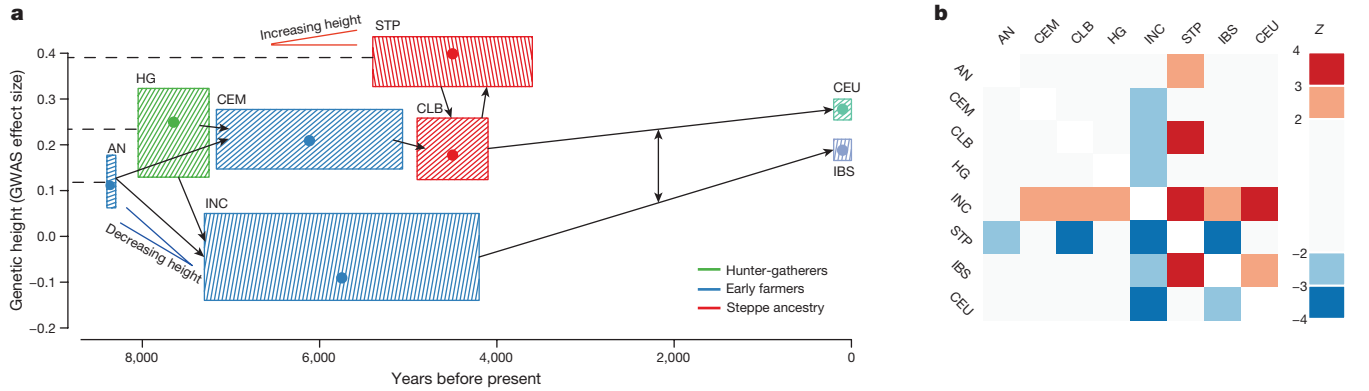


Figure 4 | Polygenic selection on height. a, Estimated genetic heights. Boxes show 0.05–0.95 posterior densities for population mean genetic height (Methods). Dots show the maximum likelihood point estimate. Arrows show major population relationships, dashed lines represent ancestral populations. The symbols < and > label potentially independent selection events resulting in an increase or decrease in height. **b**, *Z* scores for the pairwise polygenic selection test. Positive if the column population is taller than the row population.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 12 March; accepted 30 October 2015.
Published online 23 November 2015.

1. Grossman, S. R. *et al.* Identifying recent adaptations in large-scale genomic data. *Cell* **152,** 703–713 (2013).
2. Wilde, S. *et al.* Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. *Proc. Natl Acad. Sci. USA* **111,** 4832–4837 (2014).
3. Gamba, C. *et al.* Genome flux and stasis in a five millennium transect of European prehistory. *Nature Commun.* **5,** 5257 (2014).
4. Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513,** 409–413 (2014).
5. Allentoft, M. E. *et al.* Population genomics of Bronze Age Eurasia. *Nature* **522,** 167–172 (2015).
6. Keller, A. *et al.* New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nature Commun.* **3,** 698 (2012).
7. Haak, W. *et al.* Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522,** 207–211 (2015).
8. Olalde, I. *et al.* Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. *Nature* **507,** 225–228 (2014).
9. Pinhasi, R. *et al.* Optimal ancient DNA yields from the inner ear part of the human petrous bone. *PLoS ONE* **10,** e0129102 (2015).
10. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19,** 1655–1664 (2009).
11. Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192,** 1065–1093 (2012).
12. Underhill, P. A. *et al.* The phylogenetic and geographic structure of Y-chromosome haplogroup R1a. *Eur. J. Hum. Genet.* **23,** 124–131 (2015).
13. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
14. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55,** 997–1004 (1999).
15. Enattah, N. S. *et al.* Identification of a variant associated with adult-type hypolactasia. *Nature Genet.* **30,** 233–237 (2002).
16. Bersaglieri, T. *et al.* Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* **74,** 1111–1120 (2004).
17. Burger, J., Kirchner, M., Bramanti, B., Haak, W. & Thomas, M. G. Absence of the lactase-persistence-associated allele in early Neolithic Europeans. *Proc. Natl Acad. Sci. USA* **104,** 3736–3741 (2007).
18. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466,** 707–713 (2010).
19. Fumagalli, M. *et al.* Greenlandic Inuit show genetic signatures of diet and climate adaptation. *Science* **349,** 1343–1347 (2015).
20. Mathias, R. A. *et al.* Adaptive evolution of the FADS gene cluster within Africa. *PLoS ONE* **7,** e44926 (2012).
21. Wang, T. J. *et al.* Common genetic determinants of vitamin D insufficiency: a genome-wide association study. *Lancet* **376,** 180–188 (2010).
22. Price, A. L. *et al.* The impact of divergence time on the nature of population structure: an example from Iceland. *PLoS Genet.* **5,** e1000505 (2009).
23. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447,** 661–678(2007).
24. Huff, C. D. *et al.* Crohn's disease and genetic hitchhiking at IBD5. *Mol. Biol. Evol.* **29,** 101–111 (2012).
25. Hunt, K. A. *et al.* Newly identified genetic risk variants for celiac disease related to the immune response. *Nature Genet.* **40,** 395–402 (2008).
26. Jostins, L. *et al.* Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491,** 119–124 (2012).
27. Beleza, S. *et al.* Genetic architecture of skin and eye color in an African–European admixed population. *PLoS Genet.* **9,** e1003372 (2013).
28. Sturm, R. A. *et al.* A single SNP in an evolutionary conserved region within intron 86 of the *HERC2* gene determines human blue-brown eye color. *Am. J. Hum. Genet.* **82,** 424–431 (2008).
29. Eiberg, H. *et al.* 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).
30. Barreiro, L. B. *et al.* Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. *PLoS Genet.* **5,** e1000562 (2009).
31. Uciechowski, P. *et al.* Susceptibility to tuberculosis is associated with TLR1 polymorphisms resulting in a lack of TLR1 cell surface expression. *J. Leukoc. Biol.* **90,** 377–388 (2011).
32. Wong, S. H. *et al.* Leprosy and the adaptation of human toll-like receptor 1. *PLoS Pathog.* **6,** e1000979 (2010).
33. Fujimoto, A. *et al.* A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness. *Hum. Mol. Genet.* **17,** 835–843 (2008).
34. Kimura, R. *et al.* A common variation in EDAR is a genetic determinant of shovel-shaped incisors. *Am. J. Hum. Genet.* **85,** 528–535 (2009).
35. Kamberov, Y. G. *et al.* Modeling recent human evolution in mice by expression of a selected EDAR variant. *Cell* **152,** 691–702 (2013).
36. Turchin, M. C. *et al.* Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nature Genet.* **44,** 1015–1019 (2012).
37. Berg, J. J. & Coop, G. *et al.* A population genetic signal of polygenic adaptation. *PLoS Genet.* **10,** e1004412 (2014).
38. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467,** 832–838 (2010).
39. Speliotes, E. K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genet.* **42,** 937–948 (2010).
40. Heid, I. M. *et al.* Meta-analysis identifies 13 new loci associated with waist–hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nature Genet.* **42,** 949–960 (2010).
41. Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genet.* **44,** 981–990 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank P. de Bakker, J. Burger, C. Economou, E. Fornander, Q. Fu, F. Hallgren, K. Kirsanow, A. Mittnik, I. Olalde, A. Powell, P. Skoglund, S. Tabrizi and A. Tandon for discussions, suggestions about SNPs to include, or contribution to sample preparation or data curation. We thank S. Pääbo, M. Meyer, Q. Fu and B. Nickel for collaboration in developing the 1240k capture reagent. We thank J. M. V. Encinas and M. E. Prada for allowing us to resample La Braña 1. I.M. was supported by the Human Frontier Science Program LT001095/2014-L. C.G. was supported by the Irish Research Council for Humanities and Social Sciences (IRCHSS). F.G. was supported by a grant of the Netherlands Organization for Scientific Research, no. 380-62-005. A.K., P.K. and O.M. were supported by RFBR no. 15-06-01916 and RFH no. 15-11-63008 and O.M. by a state grant of the Ministry of Education and Science of the Russia Federation no. 33.1195.2014/k. J.K. was supported by ERC starting grant APGREID and DFG grant KR 4015/1-1. K.W.A. was supported by DFG grant AL 287 / 14-1. C.L.-F. was supported by a BFU2015-64699-P grant from the Spanish government. W.H. and B.L. were supported by Australian Research Council DP130102158. R.P. was supported by ERC starting grant ADNABIOARC (263441), and an Irish Research Council ERC support grant. D.R. was supported by US National Science Foundation HOMINID grant BCS-1032255, US National Institutes of Health grant GM100233, and the Howard Hughes Medical Institute.

Author Contributions W.H., R.P. and D.R. supervised the study. S.A.R., J.L.A., J.M.B., E.C., F.G., A.K., P.K., M.L., H.M., O.M., V.M., M.A.R., J.R., J.M.V., J.K., A.C., K.W.A., D.B., D.A., C.L., W.H., R.P. and D.R. assembled archaeological material. I.M., I.L., N.R., S.M., N.P., S.D., J.P, W.H. and D.R. analysed genetic data. N.R., E.H., K.St., D.F., M.N., K.Si., C.G., E.R.J., B.L., C.L. and W.H. performed wet laboratory ancient DNA work. I.M., I.L. and D.R. wrote the manuscript with input from all co-authors.

Author Information The aligned sequences are available through the European Nucleotide Archive under accession number PRJEB11450. The Human Origins genotype datasets including ancient individuals can be found at (http://genetics.med.harvard.edu/reich/Reich_Lab/Datasets.html). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to I.M. (iain_mathieson@hms.harvard.edu), W.H. (haak@shh.mpg.de), R.P. (ron.pinhasi@ucd.ie) or D.R. (reich@genetics.med.harvard.edu).

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Ancient DNA analysis. We screened 433 next-generation sequencing libraries from 270 distinct samples for authentic ancient DNA using previously reported protocols[7]. All libraries that we included in nuclear genome analysis were treated with uracil-DNA-glycosylase (UDG) to reduce characteristic errors of ancient DNA[42].

We performed in-solution enrichment for a targeted set of 1,237,207 SNPs using previously reported protocols[4,7,43]. The targeted SNP set merges 394,577 SNPs first reported in ref. 7 (390k capture), and 842,630 SNPs first reported in ref. 44 (840k capture). For 67 samples for which we newly report data in this study, there was pre-existing 390k capture data[7]. For these samples, we only performed 840k capture and merged the resulting sequences with previously generated 390k data. For the remaining samples, we pooled the 390k and 840k reagents together to produce a single enrichment reagent, which we called 1240k. We attempted to sequence each enriched library up to the point where we estimated that it was economically inefficient to sequence further. Specifically, we iteratively sequenced more and more from each sample and only stopped when we estimated that the expected increase in the number of targeted SNPs hit at least once would be less than about one for every 100 new read pairs generated. After sequencing, we filtered out samples with <30,000 targeted SNPs covered at least once, with evidence of contamination based on mitochondrial DNA polymorphism[43], a high rate of heterozygosity on chromosome X despite being male[45], or an atypical ratio of X to Y sequences.

Of the targeted SNPs, 47,384 are 'potentially functional' sites chosen as follows (with some overlap): 1,290 SNPs identified as targets of selection in Europeans by the Composite of Multiple Signals (CMS) test[1]; 21,723 SNPS identified as significant hits by genome-wide association studies, or with known phenotypic effect (GWAS); 1,289 SNPs with extremely differentiated frequencies between HapMap populations[46] (HiDiff); 9,116 'Immunochip' SNPs chosen for study of immunity-related phenotypes (Immune); 347 SNPs phenotypically relevant to South America (mostly altitude adaptation SNPs in *EGLN1* and *EPAS1*), 5,387 SNPs which tag HLA haplotypes and 13,672 expression quantitative trait loci[47] (eQTL).

Population history analysis. We used two data sets for population history analysis. 'HO' consists of 592,169 SNPs, taking the intersection of the SNP targets and the Human Origins SNP array[4]; we used this data set for co-analysis of present-day and ancient samples. 'HOIII' consists of 1,055,209 SNPs that additionally includes sites from the Illumina genotype array[48]; we used this data set for analyses only involving the ancient samples.

On the HO data set, we carried out principal components analysis in smartpca[49] using a set of 777 West Eurasian individuals[4], and projected the ancient individuals with the option 'lsqproject: YES'. We carried out admixture analysis on a set of 2,345 present-day individuals and the ancient samples after pruning for LD in PLINK 1.9 (https://www.cog-genomics.org/plink2)[50] with parameters '-indep-pairwise 200 25 0.4'. We varied the number of ancestral populations between $K=2$ and $K=20$, and used cross-validation (–cv.) to identify the value of $K=17$ to plot in Extended Data Fig. 2f.

We used ADMIXTOOLS[11] to compute *f*-statistics, determining standard errors with a block jackknife and default parameters. We used the option 'inbreed: YES' when computing f_3-statistics of the form f_3(ancient; Ref_1, Ref_2) as the ancient samples are represented by randomly sampled alleles rather than by diploid genotypes. For the same reason, we estimated F_{ST} genetic distances between populations on the HO data set with at least two individuals in smartpca also using the 'inbreed: YES' option.

We estimated ancestral proportions as in Supplementary Information section 9 of ref. 7, using a method that fits mixture proportions on a 'test' population as a mixture of *n* 'reference' populations by using f_4-statistics of the form f_4(test or ref, O_1; O_2, O_3) that exploit allele frequency correlations of the test or reference populations with triples of outgroup populations We used a set of 15 world outgroup populations[4,7]. In Extended Data Fig. 2, we added WHG and EHG as outgroups for those analyses in which they are not used as reference populations. We plot $\mathrm{resnorm} = \|\boldsymbol{t} - \boldsymbol{R}\hat{\boldsymbol{a}}^T\|_2^2$ the squared 2-norm of the residuals where $\hat{\boldsymbol{a}}$ is a vector of *n* estimated mixture proportions (summing to 1), $\boldsymbol{t}$ is a vector of $m\binom{m-1}{2}$ f_4-statistics of the form f_4(test, O_1; O_2, O_3) for *m* outgroups, and $\boldsymbol{R}$ is a $m\binom{m-1}{2} \times n$ matrix of the form f_4(ref, O_1; O_2, O_3) (Supplementary Information section 9 of ref. 7).

We determined sex by examining the ratio of aligned reads to the sex chromosomes[51]. We assigned Y-chromosome haplogroups to males using version 9.1.129 of the nomenclature of the International Society of Genetic Genealogy (http://www.isogg.org), restricting analysis using samtools[52] to sites with map quality and base quality of at least 30, and excluding two bases at the ends of each sequenced fragment.

Genome-wide scan for selection. For most ancient samples, we did not have sufficient coverage to make reliable diploid calls. We therefore used the counts of sequences covering each SNP to compute the likelihood of the allele frequency in each population. Suppose that at a particular site, for each population we have *M* samples with sequence level data, and *N* samples with full diploid genotype calls (Loschbour, Stuttgart and the 1,000 Genomes samples). For samples $i=1\ldots N$, with diploid genotype data, we observe *X* copies of the reference allele out of *2N* total chromosomes. For each of samples $i=(N+1)\ldots(N+M)$, with sequence level data, we observe R_i sequences with the reference allele out of T_i total sequences. Then, the likelihood of the population reference allele frequency, *p* given data $D=\{X, N, R_i, T_i\}$ is given by

$$L(p;D) = B(X,2N,p)\times \prod_{i=N+1}^{N+M}\left\{p^2 B(R_i,T_i,1-\varepsilon)+2p(1-p)B(R_i,T_i,0.5)+(1-p)^2 B(R_i,T_i,\varepsilon)\right\}$$

where $B(k,n,p)=\binom{n}{k}p^k(1-p)^{n-k}$ is the binomial probability distribution and ε is a small probability of error, which we set to 0.001. We write $\ell(p;D)$ for the log-likelihood. To estimate allele frequencies, for example in Fig. 3 or for the polygenic selection test, we maximized this likelihood numerically for each population.

To scan for selection across the genome, we used the following test. Consider a single SNP. Assume that we can model the allele frequencies $\boldsymbol{p}_{mod}$ in *A* modern populations as a linear combination of allele frequencies in *B* ancient populations $\boldsymbol{p}_{anc}$. That is, $\boldsymbol{p}_{mod} = \boldsymbol{C}\,\boldsymbol{p}_{anc}$, where $\boldsymbol{C}$ is an *A* by *B* matrix with rows summing to 1. We have data D_j from population *j* which is some combination of sequence counts and genotypes as described above. Then, writing $\overline{\boldsymbol{p}} = [\boldsymbol{p}_{anc}, \boldsymbol{p}_{mod}] = [\boldsymbol{p}_1 \cdots \boldsymbol{p}_{A+B}]$ the log-likelihood of the allele frequencies equals the sum of the log-likelihoods for each population.

$$\ell(\overline{\boldsymbol{p}}, \overline{D}) = \sum_{j=1}^{A+B} \ell(\boldsymbol{p}_j; D_j)$$

To detect deviations in allele frequency from expectation, we test the null hypothesis H_0: $\boldsymbol{p}_{mod} = \boldsymbol{C}\,\boldsymbol{p}_{anc}$ against the alternative H_1: $\boldsymbol{p}_{mod}$ unconstrained. We numerically maximize this likelihood in both the constrained and unconstrained model and use the fact that twice the difference in log-likelihood is approximately χ_A^2 distributed to compute a test statistic and *P* value.

We defined the ancient source populations by the 'Selection group 1' label in Extended Data Table 1 and Supplementary Table 1 and used the 1000 Genomes CEU, GBR, IBS and TSI as the present-day populations. We removed SNPs that were monomorphic in all four of these modern populations as well as in 1000 Genomes Yoruba (YRI). We do not use FIN as one of the modern populations, because they do not fit this three-population model well. We estimated the proportions of (HG, EF, SA) to be CEU = (0.196, 0.257, 0.547), GBR = (0.362, 0.229, 0.409), IBS = (0, 0.686, 0.314) and TSI = (0, 0.645, 0.355). In practice, we found that there was substantial inflation in the test statistic, most likely due to unmodelled ancestry or additional drift. To address this, we applied a genomic control correction[14], dividing all the test statistics by a constant, λ, chosen so that the median *P* value matched the median of the null χ_4^2 distribution. Excluding sites in the potentially functional set, we estimated $\lambda = 1.38$ and used this value as a correction throughout. One limitation of this test is that, although it identifies likely signals of selection, it cannot provide much information about the strength or date of selection. If the ancestral populations in the model are, in fact, close to the real ancestral populations, then any selection must have occurred after the first admixture event (in this case, after 6500 BC), but if the ancestral populations are mis-specified, even this might not be true.

To estimate power, we randomly sampled allele counts from the full data set, restricting to polymorphic sites with a mean frequency across all populations of <0.1. We then simulated what would happen if the allele had been under selection in all of the modern populations by simulating a Wright–Fisher trajectory with selection for 50, 100 or 200 generations, starting at the observed frequency. We took the final frequency from this simulation, sampled observations to replace the actual observations in that population, and counted the proportion of simulations that gave a genome-wide significant result after GC correction (Extended Data Fig. 6a). We resampled sequence counts for the observed distribution for each population to simulate the effect of increasing sample size, assuming that the coverage and distribution of the sequences remained the same (Extended Data Fig. 6b).

We investigated how the genomic control correction responded when we simulated small amounts of admixture from a highly diverged population (Yoruba;

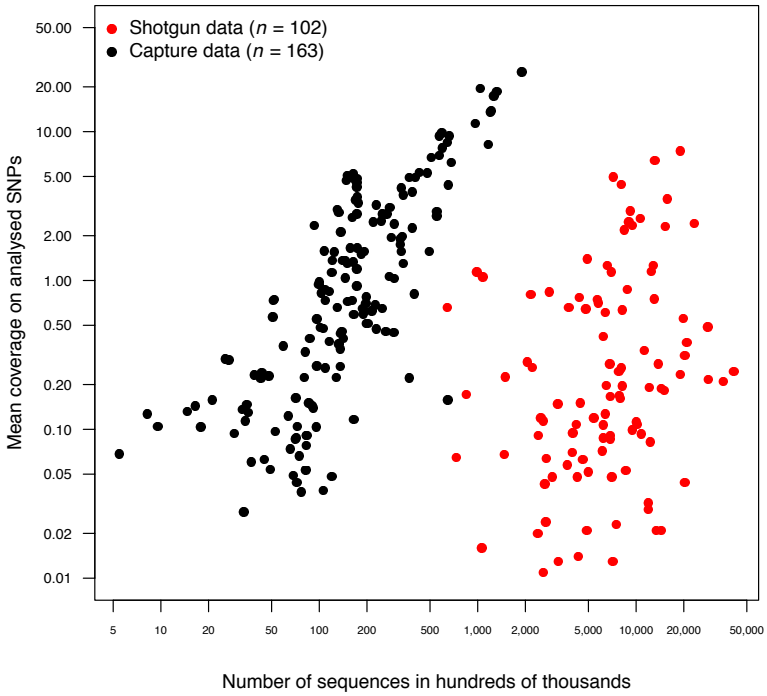
1000 Genomes YRI) into a randomly chosen modern population. The genomic inflation factor increases from around 1.38 to around 1.51 with 10% admixture, but there is little reduction in power (Extended Data Fig. 6c). Finally, we investigated how robust the test was to misspecification of the mixture matrix $\boldsymbol{C}$. We re-ran the power simulations using a matrix $\boldsymbol{C}' = x\boldsymbol{C} + (1-x)\boldsymbol{R}$ for $x \in [0, 1]$ where $\boldsymbol{R}$ was a random matrix chosen so that for each modern population the mixture proportions of the three ancient populations were jointly uniformly distributed on [0,1]. Increasing x increases the genomic inflation factor and reduces power, demonstrating the advantage of explicitly modelling the ancestries of the modern populations (Extended Data Fig. 6d).

Test for polygenic selection. We implemented the test for polygenic selection described by ref. 37. This evaluates whether trait-associated alleles, weighted by their effect size, are over-dispersed compared to randomly sampled alleles, in the directions associated with the effects measured by genome-wide association studies (GWAS). For each trait, we obtained a list of significant SNP associations and effect estimates from GWAS data, and then applied the test both to all populations combined and to selected pairs of populations. For height, we restricted the list of GWAS associations to 169 SNPs where we observed at least two chromosomes in all tested populations (Selection population 2). We estimated frequencies in each population by computing the maximum likelihood estimate (MLE), using the likelihood described above. For each test we sampled SNPs, frequency-matched in 20 bins, computed the test statistic Q_X and for ease of comparison converted these to Z scores, signed according the direction of the genetic effects. Theoretically Q_X has a χ^2 distribution but in practice, it is over-dispersed. Therefore, we report bootstrap P values computed by sampling 10,000 sets of frequency-matched SNPs.

To estimate population-level genetic height in Fig. 4a, we assumed a uniform prior on [0,1] for the frequency of all height-associated alleles, and then sampled from the posterior joint frequency distribution of the alleles, assuming they were independent, using a Metropolis–Hastings sampler with a $N(0,0.001)$ proposal density. We then multiplied the sampled allele frequencies by the effect sizes to get a distribution of genetic height.

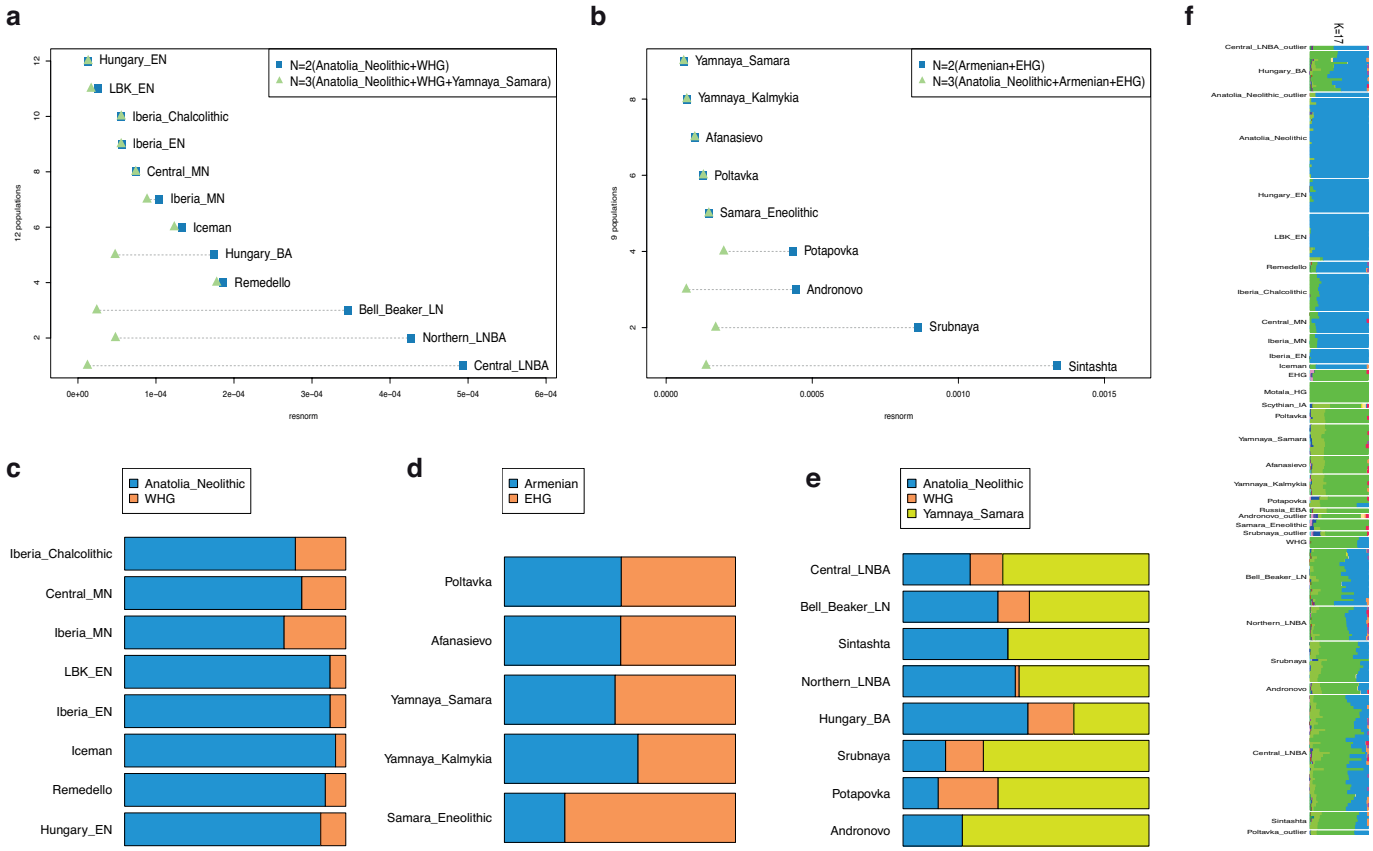
Code availability. Code implementing the selection analysis is available at https://github.com/mathii/europe_selection.

42. Briggs, A. W. *et al.* Removal of deaminated cytosines and detection of *in vivo* methylation in ancient DNA. *Nucleic Acids Res.* **38,** e87 (2010).
43. Fu, Q. *et al.* DNA analysis of an early modern human from Tianyuan Cave, China. *Proc. Natl Acad. Sci. USA* **110,** 2223–2227 (2013).
44. Fu, Q. *et al.* An early modern human from Romania with a recent Neanderthal ancestor. *Nature.* **524,** 216–219 (2015).
45. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* **15,** 356 (2014).
46. International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).
47. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501,** 506–511 (2013).
48. Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319,** 1100–1104 (2008).
49. Loh, P. R. *et al.* Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* **193,** 1233–1254 (2013).
50. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4** (2015).
51. Skoglund, P., Storå, J., Götherström, A. & Jakobsson, M. Accurate sex identification of ancient human remains using DNA shotgun sequencing. *J. Archaeol. Sci.* **40,** 4477–4482 (2013).
52. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
53. Norton, H. L. *et al.* Genetic evidence for the convergent evolution of light skin in Europeans and East Asians. *Mol. Biol. Evol.* **24,** 710–722 (2007).
54. Bokor, S. *et al.* Single nucleotide polymorphisms in the *FADS* gene cluster are associated with delta-5 and delta-6 desaturase activities estimated by serum fatty acid ratios. *J. Lipid Res.* **51,** 2325–2333 (2010).
55. Tanaka, T. *et al.* Genome-wide association study of plasma polyunsaturated fatty acids in the InCHIANTI Study. *PLoS Genet.* **5,** e1000338 (2009).
56. Ahn, J. *et al.* Genome-wide association study of circulating vitamin D levels. *Hum. Mol. Genet.* **19,** 2739–2745 (2010).
57. Gründemann, D. *et al.* Discovery of the ergothioneine transporter. *Proc. Natl Acad. Sci. USA* **102,** 5256–5261 (2005).
58. Chauhan, S. *et al.* ZKSCAN3 is a master transcriptional repressor of autophagy. *Mol. Cell* **50,** 16–28 (2013).
59. Soler Artigas, M. *et al.* Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. *Nature Genet.* **43,** 1082–1090 (2011).
60. Pruim, R. J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26,** 2336–2337 (2010).

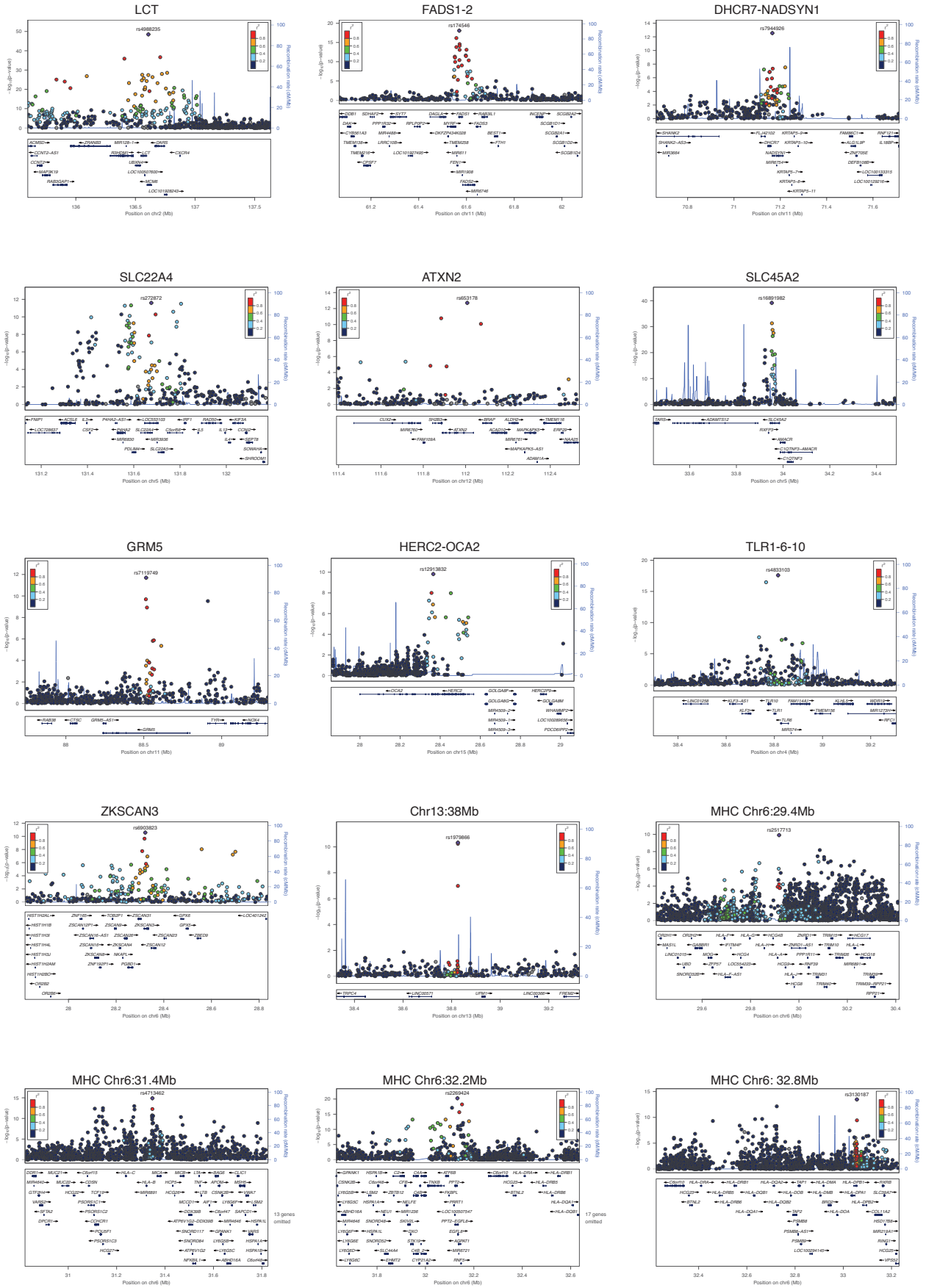


	Shotgun data (Literature)	Capture data (Newly reported)
Average number of raw read pairs (hundreds of thousands, similar to cost in dollars)	8730	241
Median coverage on analyzed SNPs	0.19	0.75

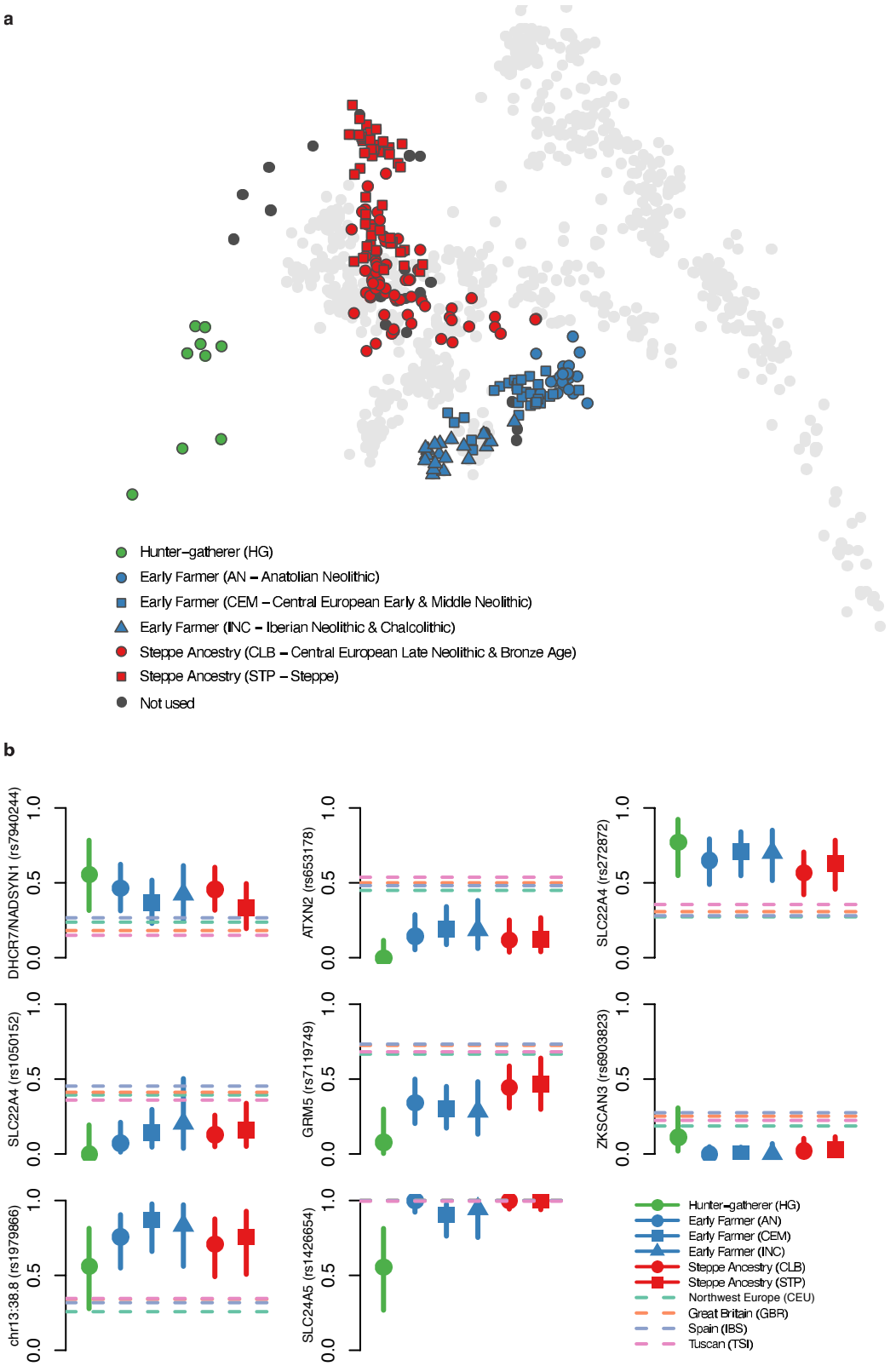
Extended Data Figure 1 | Efficiency and cost-effectiveness of 1240k capture. We plot the number of raw sequences against the mean coverage of analysed SNPs after removal of duplicates, comparing the 163 samples for which capture data are newly reported in this study, against the 102 samples analysed by shotgun sequencing in ref. 5. We caution that the true cost is more than that of sequencing alone.



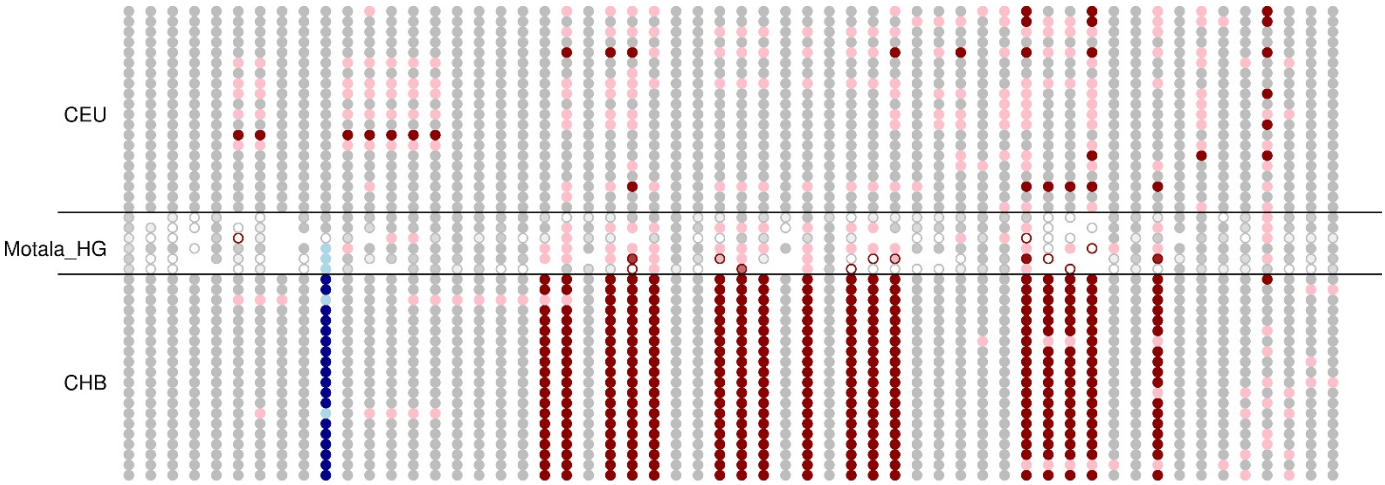
Extended Data Figure 2 | Early isolation and later admixture between farmers and steppe populations. a, Mainland European populations later than 3000 BC are better modelled with steppe ancestry as a third ancestral population, (closer correspondence between empirical and estimated f_4-statistics as estimated by resnorm; Methods). **b**, Later (post-Poltavka) steppe populations are better modelled with Anatolian Neolithic as a third ancestral population. **c**, Estimated mixture proportions of mainland European populations without steppe ancestry. **d**, Estimated mixture proportions of Eurasian steppe populations without Anatolian Neolithic ancestry. **e**, Estimated mixture proportions of later populations with both steppe and Anatolian Neolithic ancestry. **f**, Admixture plot at $k = 17$ showing population differences over time and space. EN, Early Neolithic; MN, Middle Neolithic; LN, Late Neolithic; BA, Bronze Age; LNBA, Late Neolithic and Bronze Age.



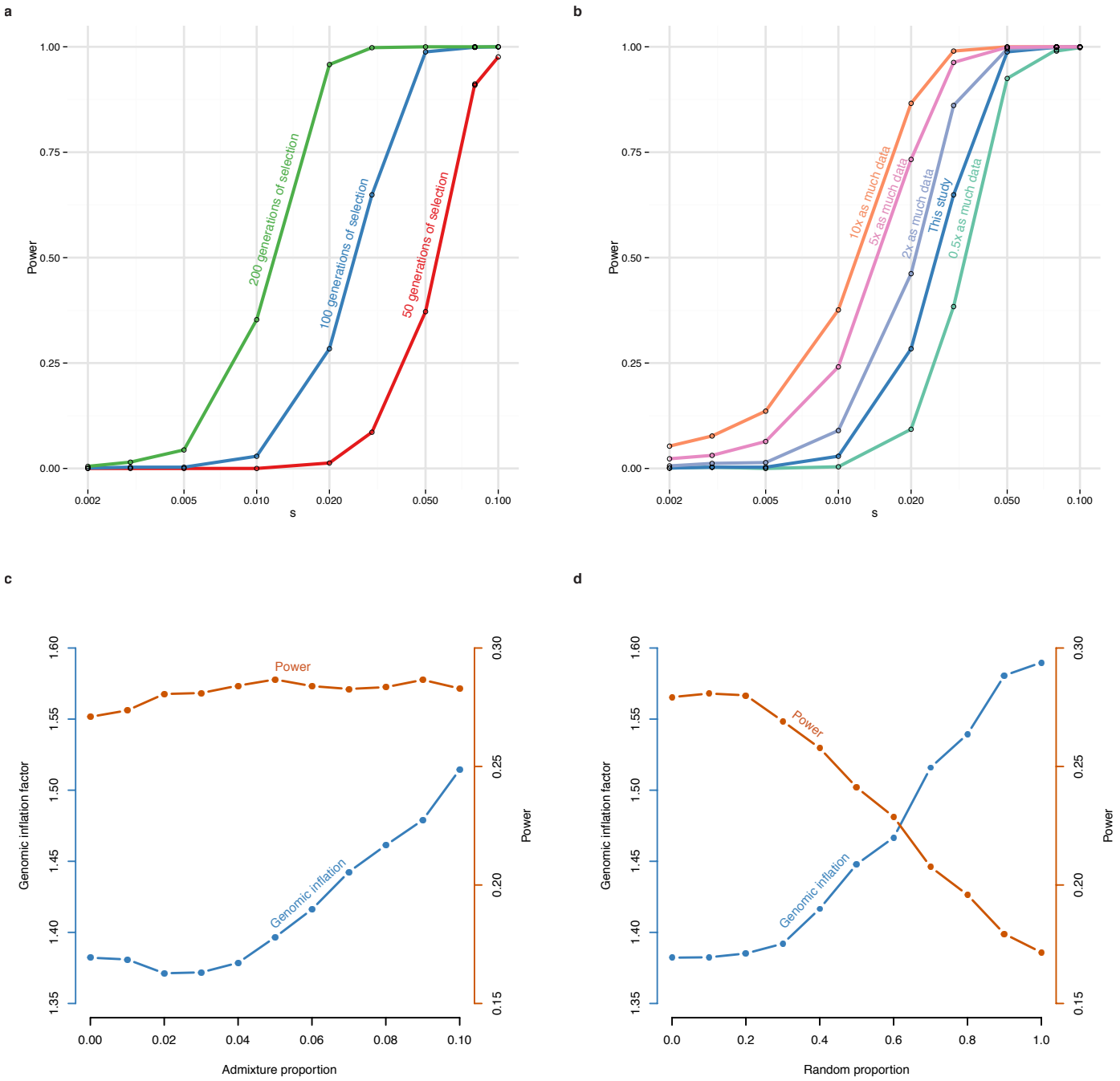
Extended Data Figure 3 | Regional association plots. Locuszoom[60] plots for genome-wide significant signals. Points show the $-\log_{10}$ *P* value for each SNP, coloured according to their linkage disequilibrium (LD; units of r^2) with the most associated SNP. The blue line shows the recombination rate, with scale on right hand axis in centimorgans per megabase (cM/Mb). Genes are shown in the lower panel of each subplot.



Extended Data Figure 4 | PCA of selection populations and derived allele frequencies for genome-wide significant signals. a, Ancient samples projected onto principal components of modern samples, as in Fig. 1, but labelled according to selection populations defined in Extended Data Table 1. **b**, Allele frequency plots as in Fig. 3. Six signals not included in Fig. 3—for *SLC22A4* we show both rs272872, which is our strongest signal, and rs1050152, which was previously hypothesized to be under selection, and we also show *SLC24A5*, which is not genome-wide significant but is discussed in the main text.



Extended Data Figure 5 | Motala haplotypes carrying the derived, selected *EDAR* allele. This figure compares the genotypes at all sites within 150 kb of rs3827760 (in blue) for the 6 Motala samples and 20 randomly chosen CHB (Chinese from Beijing) and CEU (Utah residents with northern and western European ancestry) samples. Each row is a sample and each column is a SNP. Grey means homozygous for the major (in CEU) allele. Pink denotes heterozygous and red indicates homozygous for the other allele. For the Motala samples, an open circle means that there is only a single sequence, otherwise the circle is coloured according to the number of sequences observed. Three of the Motala samples are heterozygous for rs3827760 and the derived allele lies on the same haplotype background as in present-day East Asians. The only other ancient samples with evidence of the derived *EDAR* allele in this data set are two Afanasievo samples dating to 3300–3000 BC, and one Scythian dating to 400–200 BC (not shown).



Extended Data Figure 6 | Estimated power of the selection scan. **a**, Estimated power for different selection coefficients (s) for a SNP that is selected in all populations for either 50, 100 or 200 generations. **b**, Effect of increasing sample size, showing estimated power for a SNP selected for 100 generations, with different amounts of data, relative to the main text. **c**, Effect of admixture from Yoruba (YRI) into one of the modern populations, showing the effect on the genomic inflation factor (blue, left axis) and the power to detect selection on a SNP selected for 100 generations with a selection coefficient of 0.02. **d**, Effect of mis-specification of the mixture proportions. Here 0 on the *x* axis corresponds to the proportions we used, and 1 corresponds to a random mixture matrix.

Extended Data Table 1 | 230 ancient individuals analysed in this study

By population	Population	Date range	N	Out	Rel	Eff N Chr	Selection population 1	Selection population 2
	WHG	8.2-8.0 kya	3	0	0	4.66	HG	HG
	Motala_HG	7.9-7.5 kya	6	0	0	5.19	HG	HG
	Anatolia_Neolithic	8.4-8.3 kya	24	1	1	22.49	EF	AN
	Hungary_EN	7.7-7.7 kya	10	0	0	8.81	EF	CEM
	LBK_EN	7.5-7.1 kya	15	0	0	11.15	EF	CEM
	Central_MN	5.9-5.8 kya	6	0	0	3.66	EF	CEM
	Iberia_EN	7.3-7.2 kya	4	0	1	3.54	EF	INC
	Iberia_MN	5.9-5.6 kya	4	0	0	3.47	EF	INC
	Iberia_Chalcolithic	4.8-4.2 kya	12	0	2	5.93	EF	INC
	Remedello	5.5-5.1 kya	3	0	0	0.93	EF	-
	Iceman	5.4-5.1 kya	1	0	0	1.90	EF	-
	Central_LNBA	4.9-4.6 kya	35	1	2	17.55	SA	CLB
	Yamnaya_Samara	5.4-4.9 kya	9	0	0	6.55	SA	STP
	Yamnaya_Kalmykia	5.3-4.7 kya	6	0	0	3.50	SA	STP
	Afanasievo	5.3-5.0 kya	5	0	0	3.01	SA	STP
	Poltavka	4.9-4.7 kya	4	1*	0	4.28	SA	STP
	Sintashta	4.3-4.1 kya	5	0	0	2.35	SA	STP
	Potapovka	4.2-4.1 kya	3	0	0	0.66	SA	STP
	Srubnaya	3.9-3.6 kya	12	1*	1	7.68	SA	STP
	Andronovo	3.8-3.6 kya	3	1*	0	3.87	SA	STP
	Russia_EBA	4.9-4.5 kya	1	0	0	0.21	SA	-
	Northern_LNBA	4.9-4.5 kya	10	0	0	3.81	SA	-
	Bell_Beaker_LN	4.5-4.5 kya	17	0	1	6.64	SA^	CLB
	Hungary_BA	4.2-4.1 kya	12	0	0	4.18	SA^	CLB
	EHG	7.7-7.6 kya	3	0	0	2.15	-	-
	Samara_Eneolithic	7.2-6.0 kya	3	0	0	1.07	-	-
	Scythian_IA	2.4-2.2 kya	1	0	0	1.26	-	-

By selection population	Selection population 1	Date range	N	Out	Rel	Eff N Chr	Description
	EF	8.4-4.2 kya	79	0	0	61.88	Early Farmer
	HG	8.2-7.5 kya	9	0	0	9.85	Hunter-gatherer
	SA	5.4-3.6 kya	93	3	0	52.14	Steppe Ancestry
	Selection population 2	**Date range**	**N**	**Out**	**Rel**	**Eff N Chr**	**Description**
	AN	8.4-8.3 kya	24	0	0	22.49	Anatolian Neolithic
	CEM	7.7-5.8 kya	31	0	0	23.62	Central European Early and Middle Neolithic
	INC	7.3-4.2 kya	20	0	0	12.95	Iberian Neolithic and Chalcolithic
	HG	8.2-7.5 kya	9	0	0	9.85	Hunter-gatherer
	CLB	4.9-4.1 kya	64	0	0	28.38	Central European Late Neolithic and Bronze Age
	STP	5.4-3.6 kya	47	3	0	30.58	Steppe

Population, samples grouped by a combination of date, source, archaeology and genetics; Date range, approximate date range of samples in this group; N, number of individuals sampled; Out, number of PCA outliers (marked with an asterisk if used in selection analysis); Rel, number of related individuals removed; Eff N Chr, average over sites of the effective number of chromosomes when we use genotype likelihoods, computed as 2 per called site for samples with genotype calls, or $2-0.5^{(c-1)}$ for samples with read depth c; Selection population 1, coarse population labels (marked with a caret if not used in genome-wide scan); Selection population 2, fine population labels. E/M/LN, Early/Middle/Late Neolithic; LBK, *Linearbandkeramik*; E/S/WHG, Eastern/Scandinavian/Western hunter-gatherer; EBA, Early Bronze Age; IA, Iron Age.

Extended Data Table 2 | Key *f*-statistics used to support claims about population history

A	B	C	D	f_4(A, B, C, D)	Z	Number of SNPs	Interpretation
Anatolia_Neolithic	LBK_EN	WHG	Chimp	-0.00114	-6.8	1003751	Early European Farmers had more WHG ancestry than Anatolian Neolithic
Anatolia_Neolithic	Hungary_EN	WHG	Chimp	-0.00212	-11.9	929553	
Anatolia_Neolithic	Iberia_EN	WHG	Chimp	-0.00244	-9.6	904437	
Iberia_EN	Iberia_Chalcolithic	WHG	Chimp	-0.00311	-10.5	802471	Iberian Chalcolithic had more WHG ancestry than Iberian Early Neolithic
Iberia_MN	Iberia_Chalcolithic	WHG	Chimp	0.00010	0.3	779905	Iberian Chalcolithic did not have more WHG ancestry than Iberian Middle Neolithic
EHG	Samara_Eneolithic	MA1	Chimp	0.00140	2.3	463388	First dilution of Ancient North Eurasian ancestry (prior to the Bronze Age Yamnaya culture)
EHG	Yamnaya_Samara	MA1	Chimp	0.00513	10.6	645211	
Samara_Eneolithic	Yamnaya_Samara	MA1	Chimp	0.00366	7.6	482492	
EHG	Yamnaya_Samara	Armenian	Chimp	-0.00191	-6.1	547370	Contribution of Near Eastern ancestry to the Bronze Age Yamnaya culture
EHG	Yamnaya_Kalmykia	Armenian	Chimp	-0.00180	-5.4	536989	
Samara_Eneolithic	Yamnaya_Samara	Armenian	Chimp	-0.00100	-3.3	405599	
EHG	Poltavka	Armenian	Chimp	-0.00175	-4.9	541983	
Yamnaya_Samara	Yamnaya_Kalmykia	MA1	Chimp	-0.00010	-0.3	675630	Stability of Ancient North Eurasian ancestry between Early Bronze Age Yamnaya from Kalmykia and Samara, and the Middle Bronze Age Poltavka
Yamnaya_Samara	Poltavka	MA1	Chimp	-0.00014	-0.4	673726	
Yamnaya_Kalmykia	Poltavka	MA1	Chimp	0.00012	0.3	659346	
Yamnaya_Samara	Srubnaya	MA1	Chimp	0.00151	5.1	691149	Second dilution of Ancient North Eurasian ancestry (prior to the Late Bronze Age Srubnaya culture)
Yamnaya_Kalmykia	Srubnaya	MA1	Chimp	0.00161	4.8	676735	
Poltavka	Srubnaya	MA1	Chimp	0.00164	4.5	674756	
Yamnaya_Samara	Srubnaya	LBK_EN	Chimp	-0.00225	-11.4	974659	Arrival of Early European Farmer-related ancestry prior to the Late Bronze Age Srubnaya culture. Statistics with Anatolia_Neolithic instead of LBK_EN are similar (Z<-8, not shown).
Yamnaya_Kalmykia	Srubnaya	LBK_EN	Chimp	-0.00264	-11.4	951827	
Poltavka	Srubnaya	LBK_EN	Chimp	-0.00210	-9.0	948968	
EHG	Yamnaya_Samara	Armenian	LBK_EN	-0.00080	-5.0	559478	Different source of dilution of Ancient North Eurasian ancestry prior to the Yamnaya (Near Eastern) vs. prior to the Srubnaya (Early European Farmer-related)
EHG	Yamnaya_Kalmykia	Armenian	LBK_EN	-0.00086	-5.2	548882	
EHG	Poltavka	Armenian	LBK_EN	-0.00069	-4.1	553996	
Yamnaya_Samara	Srubnaya	Armenian	LBK_EN	0.00138	13.1	585240	
Yamnaya_Kalmykia	Srubnaya	Armenian	LBK_EN	0.00142	11.3	574333	
Poltavka	Srubnaya	Armenian	LBK_EN	0.00134	10.7	577082	

Ref_1	Ref_2	*Test*	f_3(Test; Ref1, Ref2)	Z	Number of SNPs	Interpretation
WHG	Anatolia_Neolithic	Hungary_EN	-0.00412	-6.7	548445	Early European farmers were formed by admixture between Anatolia Neolithic and WHG (the non-significant signal in the Iberian_EN may be due to genetic drift specific to this population)
WHG	Anatolia_Neolithic	LBK_EN	-0.00257	-4.6	654357	
WHG	Anatolia_Neolithic	Iberia_EN	0.00179	1.4	389101	
EHG	Armenian	Poltavka	-0.00539	-3.9	213055	Early and Middle Bronze Age steppe pastoralists were formed by admixture between EHG and a population of Near Eastern ancestry
EHG	Armenian	Yamnaya_Kalmykia	-0.00537	-4.2	213996	
EHG	Armenian	Yamnaya_Samara	-0.00586	-6.2	276568	
LBK_EN	Yamnaya_Samara	Srubnaya	-0.00630	-11.2	584111	Srubnaya was formed by admixture between populations related to Yamnaya and Early European Farmers

Extended Data Table 3 | Twelve genome-wide significant signals of selection

Lead SNP	Chromosome	Position (hg19)	P Value	Range (Mb)	Genes	Potential function
rs4988235	2	136,608,646	3.19E-49	135.3-137.3	*MCM6*,*LCT*	Lactase persistence*[15]
rs16891982	5	33,951,693	7.05E-40	33.8-34.0	*SLC45A2*	Skin pigmentation*[53]
rs2269424	6	32,132,233	5.41E-21	29.9-33.1	MHC region	Immunity*
rs174546	11	61,569,830	8.18E-19	61.5-61.6	*FADS1*,*FADS2*	Fatty acid metabolism*[18,54,55]
rs4833103	4	38,815,502	2.58E-18	38.7-38.8	*TLR1*,*TLR6*,*TLR10*	Immunity[31,32]
rs653178	12	112,007,756	1.96E-13	111.9-112.6	*ATXN2*,*SH2B3*	Unknown
rs7944926	11	71,165,625	2.86E-13	71.2-71.2	*DHCR7*,*NADSYN1*	Vitamin D metabolism [21,56]
rs7119749	11	88,515,022	2.03E-12	88.5-88.9	*GRM5*	Skin pigmentation[27]
rs272872	5	131,675,864	2.56E-12	131.4-131.8	*SLC22A4*	Ergothioneine transport[57]
rs6903823	6	28,322,296	2.96E-11	28.3-28.7	*ZKSCAN3*,*ZSCAN31*	Autophagy[58], Lung function[59]
rs1979866	13	38,825,900	4.60E-11	38.1-38.8	-	Unknown
rs12913832	11	28,365,618	1.5E-10	29.9-30.1	*HERC2*,*OCA2*	Eye color*[28,29]

Chromosome/Position/Range, co-ordinates (hg19) of the SNP with the most significant signal, and the approximate range in which genome-wide significant SNPs are found. Genes, genes in which the top SNP is located, and selected nearby genes. Potential function, function of the gene, or specific trait under selection. Marked with an asterisk if the signal was still genome-wide significant in an analysis that used only the populations that correspond best to the three ancestral populations (WHG, Anatolian Neolithic and Bronze Age steppe), resulting in a less powerful test with the effective number of chromosomes analysed at the average SNP reduced from 125 to 50, a genomic control correction of 1.32, and five genome-wide significant loci that are a subset of the original twelve. Refs 53–59 are cited in this table.