Adaptation of the *FADS* gene family in Europe: Variation across time, geography and subsistence

Kaixiong Ye[1], Feng Gao[1], David Wang[1], Ofer Bar-Yosef[2], Alon Keinan[1]*

[1] Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY, USA

[2] Department of Anthropology, Harvard University, Cambridge, MA, USA

*corresponding author: alon.keinan@cornell.edu

Abstract: The *FADS* gene family encodes rate-limiting enzymes for the biosynthesis of omega-6 and omega-3 long chain polyunsaturated fatty acids (LCPUFAs), which is essential for individuals subsisting on LCPUFAs-poor diets (e.g. plant-based). Positive selection on *FADS* genes has been reported in multiple populations, but its presence and pattern in Europeans remain elusive. Here, with analyses of ancient and modern DNA, we demonstrated positive selection acted on variants centered on *FADS1* and *FADS2* both before and after the advent of farming in Europe, but adaptive alleles in these two periods are opposite. Selection signals in recent history also vary geographically, with the strongest in Southern Europe. We showed that adaptive alleles in recent farmers are associated with expression of *FADS* genes, enhanced LCPUFAs biosynthesis and reduced risk of inflammatory bowel diseases. Thus, the adaptation of *FADS* genes in Europe varies across time and geography, probably due to varying diet and subsistence.

Identifying genetic adaptations to local environment, including historical dietary practice, and elucidating their implications on human health and disease are of central interest in human evolutionary genomics[1]. The fatty acid desaturase (*FADS*) gene family consists of *FADS1*, *FADS2* and *FADS3*, which evolved by gene duplication[2]. *FADS1* and *FADS2* encode rate-limiting enzymes for the endogenous synthesis of omega-3 and omega-6 long-chain polyunsaturated fatty acids (LCPUFAs) from shorter-chain precursors from plants (Supplementary Fig. 1). LCPUFAs are indispensable for proper human brain development, cognitive function and immune response[3,4]. While omega-3 and omega-6 LCPUFAs can be consumed from animal-based diets, their endogenous synthesis is essential to compensate for their absence from plant-based diets. Adaptation (positive selection) acting on the *FADS* locus, a 100 kilobase (kb) region containing all three genes (Supplementary Fig. 2), has been identified in multiple populations[5-9]. Our recent study showed that a 22 bp insertion-deletion polymorphism (indel, rs66698963) within *FADS2*, which is associated with *FADS1* expression[10], has been adaptive in Africa, South Asia and parts of East Asia, possibly driven by local historical plant-based diets[8]. We further supported this hypothesis by the functional association of the adaptive insertion allele with more efficient endogenous synthesis[8]. In Greenlandic Inuit, who have traditionally subsisted on a LCPUFAs-rich marine diet, adaptation signals were also observed on the *FADS* locus, with adaptive alleles associated with less efficient endogenous synthesis[9].

In Europeans, positive selection on the *FADS* locus has only been reported recently in a study based on ancient DNA (aDNA)[11]. Evidence of positive selection from modern DNA (mDNA) is still lacking even though most of the above studies also performed similarly-powered tests in Europeans[5-8]. Moreover, although there are well-established differences in the Neolithization process and in dietary patterns across Europe[12-14], geographical differences of selection signals within Europe have not been investigated before. Furthermore, before the advent of farming, pre-Neolithic hunter-gatherers throughout Europe had been subsisting on animal-based diets with significant aquatic contribution[15-17], in contrast to the plant-heavy diets of recent European farmers[18-20]. We hypothesized that these drastic differences in subsistence strategy and dietary practice before and after the Neolithic revolution within Europe exert different selection pressures on the *FADS* locus. In this study, we combined analyses on ancient and modern DNA to investigate the geographical and temporal differences of selection signals on the *FADS* locus in Europe. We further interpreted the functional significance of adaptive alleles with analysis of expression quantitative trait loci (eQTLs) and genome-wide association studies (GWAS), as well as with anthropological findings within Europe.

Results

Evidence of recent positive selection in Europe from both ancient and modern DNA

To systematically evaluate the presence of positive selection on the *FADS* locus in Europe, we performed an array of selection tests using both ancient and modern samples. We first generated a uniform set of variants across the locus in a variety of aDNA data sets (Supplementary Table S1) via imputation (Methods). For all these variants, we conducted an aDNA-based test for recent positive selection (Methods)[11]. This test includes three groups of ancient European samples and four groups of modern samples. The three ancient groups represent the three major

ancestry sources of most present-day Europeans: Western and Scandinavian hunter-gatherers (WSHG), early European farmers (EF), and Steppe-Ancestry pastoralists (SA)[11,21-23]. The four groups of modern samples were drawn from the 1000 Genomes Project (1000GP), representing Tuscans (TSI), Iberians (IBS), British (GBR) and additional northern Europeans (CEU). Each modern population has been modelled as a linear mixture of the three ancestral sources with relative proportions estimated with genome-wide single nucleotide polymorphisms (SNPs)[11]. The frequencies of a neutral SNP in the four modern populations are expected to be the linear combinations of its frequencies in the three ancient sources (the null hypothesis H_0), while significant deviation from this expectation (the alternative hypothesis H_1) serves as a signal for the presence of positive selection during recent history of Europe (not more ancient than 8500 years ago)[11]. Our results confirmed the presence of significant selection signals on many SNPs in the *FADS* locus (Fig. 1A), including the previously identified peak SNP rs174546 (p = 1.04e-21)[11]. We observed the most significant signal in the locus for an imputed SNP, rs174594 (p = 1.29e-24), which was not included in the original study[11]. SNP rs174570, one of the top adaptive SNPs reported in Greenlandic Inuit[9], also carries significant signal (p = 7.64e-18) while indel rs66698963 has no evidence of positive selection (p = 3.62e-3, but see Supplementary Text). Overall, the entire peak of selection signals coincides with a linkage disequilibrium (LD) block (henceforth referred to as the *FADS1-FADS2* LD block) in Europeans, which extends over a long genomic region of 85 kb, covering the entirety of *FADS1* and most of the much longer *FADS2* (Supplementary Figs. 2 and 3). This suggests that the large number of SNPs showing genome-wide significant signals is likely the result of one causal variant targeted by strong selection and extensive hitchhiking of nearby SNPs.

We next performed several selection tests solely based on mDNA in European populations. Considering the five European populations from 1000GP, including samples of Finns (FIN) and the four samples described above, two haplotype-based selection tests, iHS[24] and nSL[25], revealed positive selection on the derived allele of the peak SNP from the aDNA-based test, rs174594, as well as many other SNPs in the *FADS1-FADS2* LD block (Fig. 1B, Supplementary Figs. 4 and 5). The normalized nSL values are significant in all five populations and the signal exhibits a gradient of being stronger in southern Europeans and weaker in northern Europeans, as per the following order (Fig. 1B): TSI (p = 0.00044), IBS (p = 0.0020), CEU (p = 0.0039), GBR (p = 0.0093), and FIN (p = 0.017). The iHS value is only significant in TSI (p = 0.026). Repeating the analysis in two whole-genome sequencing cohorts of British ancestry from the UK10K project revealed consistently significant positive selection on the derived allele (Supplementary Fig. 6). The other three variants of potential interest (rs174546, rs174570, and rs66698963, colored in Fig. 1A) exhibit no or borderline selection signals, with only rs174570 showing significant normalized nSL values in the two southernmost populations (TSI: p = 0.022; IBS: p = 0.050, Fig. 1B). Interestingly, it is the ancestral allele of rs174570 that was under positive selection, while its derived allele has been shown to be targeted by positive selection in Greenlandic Inuit[9].

We further applied two site frequency spectrum (SFS)-based selection tests that consider frequencies of all variants in a tested region (Methods). One of these two tests, Fay and Wu's H[26], consistently revealed a cluster of significant signals spanning a 14 kb genomic region surrounding the peak SNP rs174594 in all five 1000GP European populations ($p < 0.05$, Fig. 1C,

Supplementary Fig. 7). Similar results were observed with UK10K cohorts (Supplementary Fig. 8). Local peaks of H values also surround rs174546 and rs66698963, though they do not reach significance level (Fig. 1C, Supplementary Fig. 8). In all these tests, whether significant or not, the signals are gradually stronger towards Southern populations.

Taken together, standard tests on mDNA (Fig. 1B and 1C) support the results based on aDNA (Fig. 1A) of recent positive selection on the *FADS* locus and, specifically, on the *FADS1-FADS2* LD block. Moreover, all mDNA- and aDNA-based tests unanimously point to the same genomic region as the peak of selection signals (Fig. 1). The mDNA results additionally reveal a South-North gradient correlating with more pronounced selection signals in the South. As these tests across 1000GP populations are of comparable power (N=91-107 individuals), these results highlight an interesting possibility of stronger positive selection in Southern compared to Northern Europe.

Geographical differences of recent positive selection signals across Europe

To further rigorously evaluate geographical differences of recent positive selection signals on the *FADS* locus across Europe, we revisited the aDNA-based selection test[11]. We started by decomposing the original test for four representative SNPs (Fig. 2A) and then performed a revised version of the test separately in Northern and Southern Europeans for all variants in the *FADS* locus (Fig. 2B; Methods). Our first analysis included four SNPs, three of which (rs174594, rs174546, and rs174570) are top SNPs from this and previous studies[9,11] and are highlighted in all our analyses, while the fourth SNP (rs4246215) is the one showing the biggest difference in the upcoming South-North comparison analysis (Fig. 2B). The indel rs66698963 was not highlighted in this and all upcoming analyses because it has no significant selection signals in Europe (Fig. 1). The original aDNA-based test evaluates the frequencies of a tested allele in three ancient samples and four modern 1000GP samples under two hypotheses (H_0 and H_1). Under H_1, maximum likelihood estimates (MLEs) of frequencies in all samples are constrained only by observed allele counts and thus equivalent to the direct observed frequencies (Fig. 2A; blue bars). Among the four modern samples, the observed adaptive allele frequencies for all four SNPs exhibit a South-North gradient with the highest in Tuscans and the lowest in Finns, consistent with the gradient of selection signals observed before based on mDNA (Fig. 1). Among the three ancient samples, the observed allele frequencies, equivalent to the frequencies upon admixture (Fig. 2A, orange bars for ancient groups), are always the lowest and often zero in the WSHG sample.

Under H_0, the MLEs of frequencies are constrained by the observed allele counts and an additional assumption that an allele's frequencies in the four modern samples are each a linear combination of its frequencies in the three ancient samples. Considering the later assumption alone, we can predict the frequencies of adaptive alleles right after admixture for each modern population. Because the admixture contribution of WSHG, as estimated genome-wide, is higher towards the North, constituting of 0%, 0%, 19.6%, and 36.2% for TSI, IBS, CEU, and GBR, respectively[11], the predicted adaptive allele frequencies upon admixture for these four modern populations are usually lower in the North (Fig. 2A; orange bars in modern populations), suggesting higher starting frequencies in the South at the onset of selection. Further taking into account observed allele counts in modern populations, we obtained the MLEs of frequencies under H_0 (Fig. 2A; yellow bars in modern populations). As expected, the predicted allele

frequencies are higher in the South. But more importantly, the differences between H_0 and H_1 estimates in modern populations (Fig. 2A; indicated frequency differences between yellow and orange bars) are still higher in the South, suggesting more recent factors, in addition to higher starting frequencies, might contribute to the observation of stronger selection signals in the South.

We systematically evaluated geographical differences of selection signals for all SNPs in the *FADS* locus by applying the aDNA-based selection test separately for the two Southern and the two Northern populations (Methods). All SNPs across the *FADS1-FADS2* LD block that were significant in the combined analyses (Fig. 1A) were also significant in each of the two separate analyses, but many exhibited much stronger signals in the analysis with Southern populations (Fig. 2B; Supplementary Fig. 9). The maximum difference was found for SNP rs4246215, whose *p* value in Southern populations is 12 orders of magnitude stronger than that in Northern populations. SNP rs174594, rs174546 and rs174570 also have signals that are several orders (7, 11, and 10 respectively) of magnitude stronger in the South. A further decomposition of the selection test and comparison of maximum likelihoods under the null and alternative hypotheses between South and North revealed that a stronger deviation under the null hypothesis in the South is driving the signal (Supplementary Fig. S10). This is also manifested by the bigger differences between H_0 and H_1 estimates of adaptive allele frequencies in modern Southern Europeans for the four representative SNPs (Fig. 2A, indicated frequency differences between yellow and blue bars). It is noteworthy that the pattern of stronger signal in the South is observed only for some but not all SNPs, excluding the possibility of systemic bias and pointing at SNP-specific properties, likely for SNPs that are in LD with an underlying causal variant. Indeed, the most common haplotype (referred to as haplotype D; Methods) within the *FADS1-FADS2* LD block, also exhibits frequency patterns that are consistent with adaptive alleles of the four representative SNPs: higher frequencies in the South among modern European populations, while lowest frequency in WSHG among ancient groups (Fig. 2C). Hence, these results demonstrated stronger selection signals on the *FADS1-FADS2* LD block in Southern Europeans.

Opposite selection signals in pre-Neolithic European hunter-gatherers

Motivated by the very different diet of pre-Neolithic European hunter-gatherers, we set to test the action of natural selection on the *FADS* locus before the Neolithic revolution. The availability of aDNA for pre-Neolithic European hunter-gatherers over a long historic period offers this unprecedented opportunity. To this end, we started by examining the frequency trajectory of haplotype D, the candidate adaptive haplotype in recent European history after the Neolithic revolution. As noted above, its frequency increase drastically during recent European history (Fig. 2C, the contrast between orange and blue bars). In stark contrast, it shows a clear trajectory of decreasing frequency over time among pre-Neolithic hunter-gatherers[27] (Fig. 3A): starting from 32% in the ~30,000-year-old (yo) “Věstonice cluster”, through 21% in the ~15,000 yo “El Mirón cluster”, to 13% in the ~10,000 yo “Villabruna cluster”, and to being practically absent in the ~7,500 yo WSHG group. We hypothesized that there was positive selection on alleles opposite to the recently adaptive alleles that are associated with haplotype D.

To search for SNPs with evidence of positive selection during the pre-Neolithic period, we considered the allele frequency time series for all SNPs around the *FADS* locus. We applied to each SNP a rigorous, recently-published Bayesian method[28] to infer selection coefficients from

time series data while taking into account the European demographic history[29] (Methods). The test highlighted two SNPs (rs174570 and rs2851682) within the *FADS1-FADS2* LD block carrying suggestive evidence for the presence of positive selection during the pre-Neolithic period tested, approximately 30,000-7,500 years ago (Supplementary Fig. 11). The derived alleles of rs174570 and rs2851682 have similar frequency trajectories during this period, increasing from 35.7% to 77.8% (Fig. 3B). It is noteworthy that the derived allele of rs174570 has also been shown to be targeted by positive selection in modern Greenlandic Inuit[9]. Moreover, the ancestral alleles of rs174546 and rs174594 also experienced frequency increase from about 65% to almost fixation (Fig. 3B). However, presumably due to the high starting frequencies, results from the time series test are not significant for these two SNPs. Importantly, for each of these four SNPs, the allele experiencing frequency increase is opposite to the allele associated with haplotype D, with the latter allele experiencing extreme frequency increase after the Neolithic revolution.

We inferred selection coefficients concurrently with allele age for the derived alleles of rs174570 and rs2851682[28] (Methods). For rs1745470, the marginal maximum *a posteriori* (MAP) estimates of selection coefficients (s_1 and s_2 respectively) for heterozygote and homozygote are 0.28% (95% credible interval (CI): -0.025% – 1.3%) and 0.38% (95% CI: 0.038% – 0.92%) while the joint MAP of (s_1, s_2) is (0.24%, 0.34%). The age of the mutation giving rise to the derived allele is estimated to be 57,380 years (95% CI: 157,690 – 41,930 years) (Fig. 3C, Supplementary Fig. 12). For the derived allele of rs2851682, the marginal MAP for s_1 and s_2 are 0.31% (95% CI: 0.033% – 1.65%) and 0.40% (95% CI: 0.028% – 1.12%), while the joint MAP is (0.26%, 0.35%) and its allele age is 53,440 years (95% CI: 139,620 – 39,320 years) (Fig. 3D, Supplementary Fig. 13). As the observed allele frequency time series for the derived alleles of these two SNPs fall well within the 95% CI of the posterior distribution (Figs. 3C and 3D), these results support the presence of positive selection on these two alleles since their first appearance. For both SNPs, s_2 is larger than s_1, suggesting there was directional selection and the presence of derived allele was always beneficial. Additionally, we identified another haplotype in the *FADS1-FADS2* LD block, referred to as haplotype M2 (Methods, Supplementary Table 2), that appears in modern Europeans at frequency of 10% but is much more common in Eskimos (Fig. 4A). 99% of chromosomes carrying haplotype M2 in modern Europeans also carries the derived allele of rs174570, indicating a strong association between these two. Consistent with positive selection on the derived allele of rs174570, haplotype M2 exhibits increasing frequency over time in pre-Neolithic hunter-gatherers (Supplementary Table 2), suggesting that the causal allele is associated with this haplotype.

The temporal and global evolutionary trajectory of *FADS* haplotypes

Thus far we have revealed haplotype M2 and D in the *FADS1-FADS2* LD block as the candidate adaptive haplotype within Europe before and after the Neolithic revolution, respectively. To further reveal a more complete picture of the evolutionary trajectories of haplotypes in that long LD block, we performed more detailed analyses with global ancient and modern samples. Specifically, we conducted a haplotype network analysis (Fig. 4A, Supplementary Table 2) with 450 ancient haplotypes (422 from ancient European samples included in the two previous aDNA-based selection tests and additional 28 from representative ancient samples worldwide,

such as Neanderthal[30], Denisovan[31], Ust'-Ishim[32], Anzick[33], and Kennewick[34]) and 4,358 modern haplotypes (4,314 from 1000GP and 44 from modern Eskimos). Moreover, we examined the geographical frequency distribution of the resulting haplotypes in 29 previously-defined ancient Eurasian groups[11,27,35] with 600 haplotypes from 300 ancient samples (Fig. 4B, Supplementary Table 3; 422 haplotypes from ancient European samples included in the two previous aDNA-based tests and additional ones from the Middle East[35]) and also in 27 modern groups from 1000GP and modern Eskimos (Fig. 4C, Supplementary Table 4; 5,008 haplotypes from 1000GP and 44 from Eskimos).

The top five haplotypes in modern Europeans, designated as D, M1, M2, M3 and M4 from the most to the least common (63.4%, 15.3%, 10.2%, 4.7%, 4.3%, respectively), were all observed in aDNA and in modern Africans. They account for more than 95% of haplotypes in any extant non-African populations, but only for 42% in extant African populations and for 64% in the ancient samples (Fig. 4A, Supplementary Table 2). The difference between Africans and non-Africans is consistent with the general Out-of-Africa dispersal carrying with it only a subset of African haplotypes[36]. The additional difference between ancient and modern European samples is consistent with the action of positive selection as already illustrated for haplotype D and M2, which reduces haplotype diversity[37]. Among 450 aDNA haplotypes included in the haplotype network analysis, the most common haplotypes are M2 (22%), D (17%), and M1 (16%).

Haplotype D has a frequency of 32% in the oldest European hunter-gatherer group, the ~30,000 yo "Věstonice cluster", and a frequency of 42% in the ~14,000 yo Epipalaeolithic Natufian hunter-gatherers in the Levant (Fig. 4B, Supplementary Table 3), suggesting that it was of relatively high frequency of ~35% in the Out-of-Africa ancestors. This number is also similar to those in modern-day African populations (35% - 44%, Fig. 4C, Supplementary Table 4). As we have shown above among pre-Neolithic European hunter-gatherers, the D frequency decreased over time such that it was essentially absent by the advent of farming, possibly as a result of positive selection on haplotype M2. In addition to its absence in WSHG, D was not observed in the three ~7,500-year-old Eastern hunter-gatherers (EHG, Fig. 4B). D was re-introduced into Europe with the arrival of farmers and Steppe-Ancestry pastoralists. Since the admixture of the three ancient groups in Europe, the frequency of D has increased dramatically as a result of positive selection, possibly driven by the dietary changes associated with farming. At the same time, globally D also experienced dramatic frequency increase in South Asia and parts of East Asia (Fig. 4C). However, D was absent in modern-day Eskimos.

Haplotype M2 has frequencies of 29% in the "Věstonice cluster" and of 25% in Natufian hunter-gatherers, suggesting a medium frequency of ~27% at the time of Out-of-Africa dispersal (Supplementary Table 3). However, this number is much higher than its current frequencies in present-day Africans (0% - 3%, Supplementary Table 4), which might be a result of recent positive selection on other haplotypes[5,6,8]. During the pre-Neolithic period, M2 increased in frequency from 29% in the "Věstonice cluster" to 56% in WSHG and 50% in EHG (Supplementary Table 3). After Neolithic revolution, the frequency of M2 decreased dramatically to 10% among all present-day Europeans. There is also a South-North frequency gradient for M2: TSI (4%), IBS (7%), CEU (9%), GBR (10%), and FIN (22%). It is noteworthy that these two trends are opposite to those of haplotype D. Globally, in addition to its low frequency in Africa, M2 has low frequency in South Asia (1% - 5%) but high frequency in southern parts East Asia (44% - 53%, Supplementary Table 4). Its frequency in Eskimos is 27%.

Haplotype M1 has frequencies of 11% in the "Věstonice cluster" and of 8% in Natufian hunter-gatherers, suggesting a low frequency of ~10% at the time of Out-of-Africa dispersal (Supplementary Table 3). Similar to M2, this frequency is much higher than that in present-day Africans (0% - 6%, Supplementary Table 4). In contrast to D and M2, M1 had little frequency change during the pre-Neolithic period, maintaining at ~11% from the "Věstonice cluster" to WSHG (Supplementary Table 3). It also had little frequency change over time in Europe, with a frequency of ~15% in modern Europeans. Globally, M1 has overall low frequencies (<20%) except for Eskimos and American populations (Supplementary Table 4). With a frequency of 73%, it dominates the haplotypes observed in Eskimos, making it the candidate adaptive haplotype in this seafood-eating population[9].

The global frequency patterns of representative variants within the *FADS1-FADS2* LD block (rs174570, rs66698963, rs174594, rs174546, and rs2851682; Fig. 4D, Supplementary Figs. 15-19) mostly mirror those of key haplotypes, but with discrepancies that provide insights into casual variants and allele ages. One major discrepancy was found in Africa. The derived alleles of rs174570 and rs2851682 remains almost absent in Africa (Fig. 4D, Supplementary Figs. 15 and 19), consistent with their allele age estimates of ~55,000 years (Figs. 3C and 3D) and ruling out their possible involvement in the positive selection on *FADS* genes in Africa[5,6,8]. Considering the poor LD structure of the *FADS* locus in Africa (Supplementary Fig. 20), it is possible that selection in Africa may be on haplotypes and causal variants that are different from those in Europe.

Functional analyses of adaptive variants

Previous studies on adaptive evolution of the *FADS* locus suggested that alleles targeted by positive selection are also associated with expression levels of *FADS* genes[5,6,8]. To test this possibility in the context of this large-sale analysis, we considered data from the Genotype-Tissue Expression (GTEx) project[38]. Our results point to many SNPs on the *FADS1-FADS2* LD block being eQTLs of *FADS* genes. Out of a total of 44 tissues, these eQTLs at genome-wide significance level are associated with the expression of *FADS1*, *FADS2*, and *FADS3* in 12, 23, and 4 tissues, respectively, for a total of 27 tissues (Supplementary Figs. 21-23). Considering the peak SNP rs174594 alone, nominally significant associations with these three genes were found in 29, 28 and 4 tissues, respectively. More importantly, out of these tissues with association signals, the adaptive allele in recent European history is associated with higher expression of *FADS1*, lower expression of *FADS2* and higher expression of *FADS3* in 28, 27 and 4 tissues, respectively. The general trend that recently adaptive allele is associated with higher expression of *FADS1* but lower expression of *FADS2* was also observed for other representative SNPs (rs174546, rs174570, and rs2851682) in the *FADS1-FADS2* LD block.

Genome-wide association studies (GWAS) have revealed 178 association signals with 44 different traits in the 85 kb *FADS1-FADS2* LD block, as recorded in the GWAS catalog (Supplementary Tables 5-9)[39]. All effects reported in the following are based on GWAS conducted with individuals of European ancestry, while some are also replicated in other ethnic groups. Dissecting different associations, (1) the most prominent group of associated traits are polyunsaturated fatty acids (PUFAs, Supplementary Fig. 1), including LCPUFAs and their shorter chain precursors. Alleles on haplotype D are associated with higher levels of arachidonic

acid (20:4n-6, AA)[40-42], adrenic acid (22:4n-6, AdrA)[40,42-44], eicosapentaenoic acid (20:5n-3, EPA)[42,45] and docosapentaenoic acid (22:5n-3, DPA)[42,43,45], but with lower levels of dihomo-gamma-linolenic acid (20:3n-6, DGLA)[40-43], all of which suggest increased activity of delta-5 desaturase encoded by *FADS1*[42,46]. This is consistent with the association of recently adaptive alleles with higher *FADS1* expression. Surprisingly, alleles on haplotype D are associated with higher levels of gamma-linolenic acid (18:3n-6, GLA)[40,41,43] and stearidonic acid (18:4n-3, SDA)[42], but with lower levels of linoleic acid (18:2n-6, LA)[40,41,43,47] and alpha- linolenic acid (18:3n-3, ALA)[41,43,45], suggesting increased activity of delta-6 desaturase encoded by *FADS2*[41]. However, the above eQTL analysis suggested that adaptive alleles tend to be associated with lower *FADS2* expression. Some of these association signals have been replicated across Europeans[40,42-47], Africans[45], East Asians[41,45], and Hispanic/Latino[45]. (2) Besides PUFAs, recently adaptive alleles on haplotype D are associated with decreased cis/trans-18:2 fatty acids[48], which in turn is associated with lower risks for systemic inflammation and cardiac death[48]. Consistently, adaptive alleles are also associated with decreased resting heart rate[49,50], which reduces risks of cardiovascular disease and mortality. (3) With regards to other lipid levels, adaptive alleles have been associated with higher levels of high-density lipoprotein cholesterol (HDL)[51-56], low-density lipoprotein cholesterol (LDL)[51-53,57] and total cholesterol[51-53], but with lower levels of triglycerides[51,52,55,56]. (4) In terms of direct association with disease risk, adaptive alleles are associated with lower risk for inflammatory bowel diseases (IBD), both Crohn's disease[58-60] and ulcerative colitis[60].

Going beyond known associations from the GWAS catalog, we analyzed data from the two sequencing cohorts of the UK10K study. Focusing on the peak SNP rs174594, we confirmed the association of the recently adaptive allele with higher levels of TC, LDL, and HDL. We further revealed that its adaptive allele is associated with higher levels of additional lipids, Apo A1 and Apo B (Supplementary Fig. 24). Taken together, adaptive alleles in the *FADS1-FADS2* LD block, beyond their direct association with fatty acid levels, are associated with factors that are mostly protective against inflammatory and cardiovascular diseases, and indeed also show direct association with decreased risk of a type of inflammatory autoimmune diseases.

Discussion

Evidence for positive selection on *FADS* genes in Europe. For the first time, we revealed that patterns of positive selection on *FADS* genes within Europe vary geographically, between the North and the South, and temporally, before and after the Neolithic revolution. Positive selection on *FADS* genes within Europe was initially reported in a recent aDNA-based study[11]. Here, we repeated the aDNA-based analysis with much higher density of variants and confirmed the presence of positive selection. Moreover, we strengthened this discovery by providing independent evidence based on mDNA analyses. Both aDNA and mDNA results consistently pointed to the region surrounding SNP rs174594 as the peak of signals, suggesting the possibility of a causal variant in that region. Overall, selection signals revealed by both aDNA and mDNA analyses coincide with an 85 kb LD block covering *FADS1* and *FADS2*. Within this LD block, the most common haplotype in current Europeans, haplotype D, is the candidate adaptive haplotype. With regards to the timing of the selection event underlying these signals, because the aDNA-based analysis specifically models the frequency change from ancient to current samples,

the onset of selection must have occurred after the first admixture between early farmers and northwestern hunter-gatherers which was around 8,500 years ago[11]. One of the top adaptive SNPs reported in Greenlandic Inuit (rs174570)[9], also locates in the *FADS1-FADS2* LD block and carries adaptive signals in Europeans based on our aDNA-based analysis and haplotype-based test on mDNA. Interestingly, while its derived allele is adaptive Inuit[9], it is its ancestral allele that is adaptive in Europeans, suggesting the presence of opposite selection pressures, possibly because of very different diets in these two populations. The indel rs66698963, previously reported to be adaptive in Africans, South Asians, and parts of East Asians, does not carry significant adaptive signals in Europeans. However, there is a caveat that the imputation quality for this indel might not be good enough. This indel is also a copy number variation and has a very complex sequence context (Supplementary Text). 1000GP, the reference panel for imputation, consists of known genotype calling errors for this indel[8]. Both of our aDNA-based test and haplotype-based test revealed little signals for this indel, but the SFS-based test (Fay and Fu's H) unraveled a local peak around the indel, although not reaching genome-wide significance. Inaccurate imputation might explain this pattern because the first two tests are single-variant-based test while the third one draws information from all SNPs within a 5 kb window and thus is less affected by imputation inaccuracy of a single variant. Besides the *FADS1-FADS2* LD block, additional selection signals were detected with mDNA analyses (Figs. 1B and 1C) around the beginning of *FADS3*. Detailed analyses on this region are beyond the scope of this study and will be published separately.

For the first time, we demonstrated geographical differences of positive selection on *FADS* genes within Europe. The possibility of geographical differences was first suggested in our mDNA analyses (nSL, iHS, and Fay and Wu's H), with the strongest signals always observed in Southern Europeans, especially Tuscans. To formally evaluate the presence of geographical differences, we used four SNPs as examples and dissected different layers of forces, either demographic or selection, contributing to their final adaptive allele frequencies in current European populations. We revealed three layers of forces. First, among the three ancient samples, adaptive alleles always have the lowest frequencies or are even absent in western and Scandinavian hunter-gatherers (Fig. 2A). This is consistent with our observation that opposite selection forces operated in pre-Neolithic European hunter-gatherers and in more recent European farmers. Second, there are differential admixture proportions of ancient sources for Northern and Southern Europeans. The contribution of hunter-gatherers is higher towards the North, while the contribution of early farmers is higher towards the South. As a result, the predicted frequencies right after admixture are already higher in the South (Fig. 2A). Third, with a null model taking into account the first two layers and also observed allele counts in modern populations, we predicted current allele frequencies under neutrality. They are still lower than observed allele frequencies, calculated directly from observed allele counts, indicating the presence of positive selection as already detected in the aDNA-based test (Fig. 1A). More importantly, the bigger differences in the two Southern European populations compared to the two Northern populations suggest still stronger selection signals in the South (Fig. 2A), which might be a result of stronger selection pressure or earlier onset of selection in Southern Europe. These detailed analyses on the four SNPs were further confirmed by a global analysis on all SNPs in the region with aDNA-based tests separately applied on Northern and Southern Europeans. As the selection signal detected by the aDNA-based only describes the period starting from the ancient admixture to present and the exact timing of ancient admixture could be different for different populations, it is possible that ancient admixture finished earlier in

Southern Europe and there was a longer time for the action of selection, resulting in the stronger signals we detected. The other possibility is stronger selection pressure in the South, which is consistent with the dietary differences between Southern and Northern European farmers as discussed in the next section.

We also unraveled a novel discovery regarding the temporal differences of positive selection signals within Europe before and after the Neolithic revolution. Haplotype D in the *FADS1-FADS2* LD block, the candidate adaptive haplotype during recent European history, exhibits gradual frequency decrease over time among four groups of pre-Neolithic Hunter-gatherers, from approximately 30,000-7,500 years ago. With a recently-published Bayesian method[28] for inferring selection coefficients from allele frequency time series data, we identified two SNPs (rs174570 and rs2851682) with evidence of positive selection during this period. The ages of the derived alleles for these two SNPs are similar, about 55,000 years, after the Out-of-Africa dispersal. This is consistent with the near absence of these two alleles in modern Africans (Fig. 4D, Supplementary Figs. 15 and19). Although the trend of increasing frequency over time was also observed for other SNPs in the region (e.g. rs174546 and rs174594), the formal test did not reveal significant signals for them. Several factors could potentially contribute to reduced power of the test, including the higher starting frequencies for some SNPs, the small sample size for each group, and the use of samples of different ages in the same group. Future studies with much bigger sample size are needed to refine the selection signal for this pre-Neolithic period. Additionally, it will be of interest in the future to explore potential geographical differences among hunter-gatherers, especially considering the dietary differences between Northern and Southern pre-Neolithic hunter-gatherers, which are discussed in the next section.

Interpretation of positive selection signals in light of anthropological findings. The dispersal of the Neolithic package into Europe that began some 8,500 years ago caused a sharp dietary shift from an animal-based diet with significant aquatic contribution to a terrestrial plant-heavy diet including dairy products[15-20]. Before the Neolithic revolution, consumption of aquatic food had been prominent in diets of pre-Neolithic European hunter-gatherers[61]. The significant role of aquatic food, either marine or freshwater, has been established in sites along the Atlantic coast[17,62-64], around the Baltic sea[17], and along the Danube river[65]. The content of LCPUFAs are usually the highest in aquatic foods, lower in animal meat and milk, and almost negligible in most plants[66]. Consistent with the subsistence strategy and dietary pattern in pre-Neolithic hunter-gatherers, positive selection on *FADS* genes during this period was on alleles that are associated with less efficient endogenous synthesis of LCPUFAs, possibly compensating for the high dietary input. In addition to optimal absolute levels of LCPUFAs, maintaining a balanced ratio of omega-6 to omega-3 is also critical for human health[67]. It is also possible that positive selection on *FADS* genes in hunter-gatherers was in response to an unbalanced omega-6 to omega-3 ratio (e.g. too much omega-3 LCPUFAs). Similar selection signals on *FADS* genes have been observed in modern Greenlandic Inuit, who subsist on a seafood diet[9]. Specifically, the derived allele of SNP rs174570 carries positive selection signals in both pre-Neolithic European hunter-gatherers and modern Greenlandic Inuit. More generally, haplotype M2, the candidate adaptive haplotype during the pre-Neolithic period in Europe, is also common in the modern Eskimo samples examined in our study. It is noteworthy that aquatic food was less prevalent among pre-Neolithic hunter-gatherers around the Mediterranean basin, possibly due to the low productivity of the Mediterranean Sea[68-70]. It would be interesting to examine the geographical differences of selection signals among different European groups of pre-Neolithic

hunter-gatherers. However, aDNA from pre-Neolithic hunter-gatherers is still scarce and under-represented around the Mediterranean basin, prohibiting such an analysis at present.

The Neolithization of Europe[12,71,72] started in the Southeast region around 8,500 years ago when farming and herding spread into the Aegean and the Balkans. It continued in spite of a few temporary stops into central and northern Europe following the Danube River and its tributaries, and along the Mediterranean coast. It arrived at the Italian Peninsula about 8,000 years ago and shortly after reached the Iberia by 7,500 years ago. While farming rapidly spread across the loess plains of Central Europe and reach the Paris Basin by 7,000 years ago, it took another 1,000 or more years before it spread into Britain and Northern Europe around 6,000 years ago. From that time on, European farmers relied heavily on their domesticated animals and plants. Compared to pre-Neolithic hunter-gatherers, European farmers consumed much more plants but less aquatic foods[18-20,73]. Consistent with the lack of LCPUFAs in plant-based diets, positive selection on *FADS* genes during recent European history has been on alleles that are associated with enhanced endogenous synthesis of LCPUFAs from plant-derived precursors (LA and ALA). Positive selection for enhanced LCPUFAs was also observed before in Africans, South Asians and some East Asians, possibly driven by the local traditional plant-based diets[8].

Despite the overall trend of relying heavily on domesticated plants, there are geographical differences of subsistence strategies and dietary patterns among European farmers. In addition to the 2,000-year-late arrival of farming at Northern Europe, animal husbandry and the consumption of animal milk became gradually important as Neolithic farmers spread to the Northwest[18,72,74-76]. Moreover, similar to their pre-Neolithic predecessors, Northwestern European farmers close to the Atlantic Ocean or the Baltic Sea still consumed some marine food, more so than their Southern counterparts in the Mediterranean basin[77,78]. It is noteworthy that historic dairying practice in Northwestern Europe has driven the adaptive evolution of lactase persistence in Europe to reach the highest prevalence in this region[75]. In this study, we observed stronger positive selection signals on *FADS* genes during recent history in Southern than in Northern Europeans, even after considering the later arrival of farming and the lower starting allele frequencies in the North. The higher aquatic contribution and stronger reliance on animal meat and milk might be responsible for the weaker selection pressure in the North, although the possibilities of other environmental factors could not be ruled out.

Interpretation of eQTLs and GWAS results. Although liver is the primary site for the endogenous synthesis of LCPUFAs, the action of the pathway has been observed in a wide range of tissues[79,80], including heart[81], brain[81-83], both white and brown adipose tissues[84]. Moreover, while the synthesis rate and relevant enzyme levels in liver are regulated by dietary fatty acid inputs, they are not affected in other tissues[81], indicating that identifying eQTLs for *FADS* genes in the liver might need extra control for dietary inputs. Based on data from the GTEx project, eQTLs within the *FADS1-FADS2* LD block for the three *FADS* genes were identified in multiple tissues and in general recently adaptive alleles are associated with higher *FADS1* expression but lower *FADS2* expression. No genome-wide significant eQTLs for *FADS1* and *FADS2* were found in the liver, probably due to the complication of dietary inputs, which were not available to be controlled for during analysis. However, an apparent cluster of elevated association signals with *FADS1* was observed in the liver, although they do not reach genome-wide significance level (Supplementary Fig. 21). Furthermore, for the recently adaptive allele of peak SNP rs174594, the directions of association with *FADS1* and *FADS2* in the liver, although not significant, are consistent with the general trend – higher *FADS1* but lower *FADS2* expression.

The exact causal regulatory variants and the underlying mechanisms are still unknown, but variants disrupting the sterol response element (SRE) are among the most likely candidates[2].

GWAS revealed several potential beneficial effects of the recently adaptive alleles: enhanced efficiency of the overall LCPUFAs synthesis, lower risks of systemic inflammation, inflammatory bowel diseases, and cardiovascular diseases. The directions of association with PUFAs along the synthesis pathway (Supplementary Fig. 1) reflect the relative efficiency of rate-limiting enzymes, delta-5 and delta-6 desaturases: enhanced detal-5 desaturase activity is expected to reduce levels of its precursors, LA and ALA, but to increase levels of its products, GLA and SDA, while similarly enhanced delta-6 desaturase activity is expected to reduce DGLA and ETA, but to increase levels of AA, AdrA, EPA and DPA. While GWAS results are consistent with eQTLs analysis in revealing increased *FADS1* expression and enhanced delta-5 desaturase activity, they seem contradictory for *FADS2*: recently adaptive alleles are associated with lower *FADS2* expression but enhanced delta-6 desaturase activity. There are several possible explanations. First, the *FADS2* expression level might not directly correlate with the final delta-6 desaturase level because of post-transcriptional regulation. Second, the direction of *FADS2* eQTLs might be different in the liver from other tissues. Currently, there are marginal association signals for *FADS1* in the liver but no signals for *FADS2*. Additional analysis for *FADS2* is needed in the liver with proper control for dietary inputs. Third, there may be alternative splicing in addition to expression level change. Further experiments are needed to address this discrepancy and to unravel the underlying molecular mechanisms. Besides PUFAs, GWAS also revealed an overall trend that recently adaptive alleles are protective against inflammatory conditions, especially inflammatory bowel diseases. But there are exceptions: these alleles were also found to be associated with increased risk of rheumatoid arthritis[85] and colorectal cancer[86]. Because LCPUFAs-derived signaling molecules have both pro-inflammatory and anti-inflammatory effects (Supplementary Fig. 1), elucidating the effects of these adaptive alleles on specific diseases will require case-by-case analysis with special consideration of the relative contributions of omega-6 and omega-3 LCPUFAs. The effort in understanding the clinical significance of genetic variants in *FADS* genes might also reveal additional selection pressures beyond diet acting on these genes.

Conclusions

In summary, we demonstrated that in Europe an extended LD block covering *FAD1* and *FADS2* of the *FADS* gene family has been under strong recent positive selection both before and after the Neolithic revolution. During the recent history, positive selection also varies geographically, with selection signals and adaptive allele frequencies gradually increasing from Northern towards Southern Europe. The plant-heavy diet of European farmers, with its lack of LCPUFAs, is one possible environmental factor contributing to the recent positive selection. The higher consumption of aquatic resources and animal milk among Northwestern European farmers might contribute to the weaker selection signals observed in the North. Consistently, many alleles on the recently adaptive haplotype are eQTLs that increase *FADS1* expression, thereby increasing the efficacy of LCPUFAs synthesis. Additional evidence comes from a multitude of GWAS showing recently adaptive alleles associated with enhanced LCPUFAs biosynthesis. Before the advent of farming, the recently adaptive haplotype showed dramatic decrease in frequency across pre-Neolithic hunter-gatherers. While this could have been due to negative selection affecting alleles on the haplotype, time series analysis showed that it was driven by positive selection on

alleles opposite to those on the recently adaptive haplotype. Considering that pre-Neolithic hunter-gatherers subsisted on animal-based diets with significant aquatic contribution, limiting the rate of endogenous LCPUFAs synthesis by decreasing *FADS1* expression might be beneficial and contributed to that ancient adaptation. This discovery of subsistence-based temporal and geographical variations of selection in Europe supports and completes the global picture of the local adaptation of *FADS* genes: positive selection on alleles enhancing LCPUFAs biosynthesis in populations traditionally subsisting on plant-based diets[5,6,8], but positive selection on opposite alleles in populations subsisting on a LCPUFAs-rich marine diet[9]. This opposite pattern of positive selection in different dietary environment highlights the potential of matching diet to genome in the future nutritional practice. Finally, the vast number of traits associated with the adaptive region in the *FADS* genes, while raising the possibility of additional selection forces beyond diet, stresses the clinical and nutritional significance of understanding the evolutionary forces shaping the *FADS* gene family and other diet-related genes.

Methods

Data sets. The ancient DNA (aDNA) data set included in this study was compiled from two previous studies[27,35], which in turn were assembled from many other studies[11,21,22,30-34,87-96], in addition to new sequenced samples. These two data sets were downloaded from https://reich.hms.harvard.edu/datasets and were merged by removing overlapping samples. In total, there are 325 ancient samples included in this study (Supplementary Table 1). For the aDNA-based test for recent selection in Europe, a subset of 178 ancient samples were used and clustered into three groups as in the original study[11], representing the three major ancestral sources for most present-day European populations. These three groups are: West and Scandinavian hunter-gatherers (WSHG, N=9), early European farmers (EF, N=76), and individuals of Steppe-pastoralist Ancestry (SA, N=93). Three samples in the EF group in the original study were excluded from our analysis because they are genetic outliers to this group based on additional analysis[35]. For aDNA-based test for ancient selection in pre-Neolithic European hunter-gatherers, a subset of 42 ancient samples were used and four groups were defined. In addition to the WSHG (N=9), the other three groups were as originally defined in a previous study[27]: the "Věstonice cluster", composed of 14 pre-Last Glacial Maximum individuals from 34,000-26,000 years ago; the "El Mirón cluster", composed of 7 post-Last Glacial Maximum individuals from 19,000-14,000 years ago; the "Villabruna cluster", composed of 12 post-Last Glacial Maximum individuals from 14,000-7,000 years ago. There were three Western hunter-gatherers that were originally included in the "Villabruna cluster"[27], but we included them in WSHG in the current study because of their similar ages in addition to genetic affinity[11]. In haplotype network analysis, all aDNAs included in the two aDNA-based selection tests were also included in this analysis. In addition, we included some well-known ancient samples, such as the Neanderthal, Denisovan, and Ust'-Ishim. In total, there were 225 ancient samples (450 haplotypes). For geographical frequency distribution analysis, a total of 300 ancient samples were used and classified into 29 previously defined groups[11,27,35] based on their genetic affinity, sampling locations and estimated ages.

Data for the 1000 Genomes Project (1000GP, phase 3)[7] were downloaded from the official FTP site (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/). There are in total 2,504 individuals from 5 continental regions and 26 global populations. There are 7 populations of African ancestry (AFR, N=661): Yoruba in Ibadan, Nigeria (YRI, N=108), Luhya in Webuye, Kenya (LWK, N=99), Gambian in Western Divisions in the Gambia (GWD, N=113), Mende in Sierra Leone (MSL, N=85), Esan in Nigeria (ESN, N=99), Americans of African Ancestry in SW USA (ASW, N=61), African Caribbeans in Barbados (ACB, N=96); 5 populations of European ancestry (EUR, N=503): Utah Residents with Northern and Western European Ancestry (CEU, N=99), Toscani in Italia (TSI, N=107), Finnish in Finland (FIN, N=99), British in England and Scotland (GBR, N=91), Iberian Population in Spain (IBS, N=107); 5 populations of East Asian ancestry (EAS, N=504): Han Chinese in Beijing, China (CHB, N=103), Japanese in Tokyo, Japan (JPT, N=104), Southern Han Chinese (CHS, N=105), Chinese Dai in Xishuangbanna, China (CDX, N=93), Kinh in Ho Chi Minh City, Vietnam (KHV, N=99); 5 populations of South Asian ancestry (SAS, N=489): Gujarati Indian from Houston, Texas (GIH, N=103), Punjabi from Lahore, Pakistan (PJL, N=96), Bengali from Bangladesh (BEB, N=86), Sri Lankan Tamil from the UK (STU, N=102), Indian Telugu from the UK (ITU, N=102), and 4 populations of American ancestry (AMR=347): Mexican Ancestry from Los Angeles USA (MXL, N=64), Puerto Ricans from Puerto Rico (PUR, N=104), Colombians from Medellin, Colombia (CLM, N=94), Peruvians from Lima, Peru (PEL, N=85).

The data set for Human Genome Diversity Project (HGDP)[97] was downloaded from http://www.hagsc.org/hgdp/files.html. There were ~650K SNPs in 939 unrelated individuals from 51 populations. The data from the Population Reference Sample (POPRES)[98] were retrieved from dbGaP with permission. Only 3,192 Europeans were included in our analysis. The country of origin of each sample was defined with two approaches. Firstly, a "strict consensus" approach was used: an individual's country of origin was called if and only if all four of his/her grandparents shared the same country of origin. Secondly, a more inclusive approach was used to further include individuals that had no information about their grandparents. In this case, their countries of birth were used. Both approaches yielded similar results and only results from the inclusive approach are reported. The 22 Eskimo samples were extracted from the Human Origins dataset[22].

The two sequencing cohorts of UK10K were obtained from European Genome-phenome Archive with permission[99]. These two cohorts, called ALSPAC and TwinsUK, included low-depth whole-genome sequencing data and a range of quantitative traits for 3,781 British individuals of European ancestry (N=1,927 and 1,854 for ALSPAC and TwinsUK, respectively)[99].

Imputation for ancient and modern DNA. Genotype imputation was performed using Beagle 4.1[100] separately for the data sets of aDNA, HGDP and POPRES. The 1000GP phase 3 data were used as the reference panel[7]. Imputation was performed for a 5-Mb region surrounding the *FADS* locus (hg19:chr11: 59,100,000-64,100,000), although most of our analysis was restricted to a 200 kb region (hg19:chr11:61,500,000-61,700,000). For most of our analysis (e.g. estimated allele count or frequency for each group), genotype probabilities were taken into account without

setting a specific cutoff. For haplotype-based analysis (e.g. estimated haplotype frequency for each group), a cutoff of 0.8 was enforced and haplotypes were defined with missing data (if the genotype does not reach the cutoff) following the phasing information from imputation.

Genotype imputation for aDNA has been shown to be desirable and reliable[88]. We also evaluated the imputation quality for aDNA by comparing with the two modern data sets (Supplementary Fig. 25). Overall, the imputation accuracy for ungenotyped SNPs, measured with allelic R^2 and dosage R^2, is comparable between aDNA and HGDP, but is higher in aDNA when compared with POPRES. Note that the sample sizes are much larger for HGDP (N=939) and POPRES (N=3,192), compared to aDNA (N=325). The comparable or even higher imputation quality in aDNA was achieved because of the higher density of genotyped SNPs in the region.

Linkage disequilibrium and haplotype network analysis. Linkage disequilibrium (LD) analysis was performed with the Haploview software (version 4.2)[101]. Analysis was performed on a 200-kb region (chr11:61,500,000-61,700,000), covering all three *FADS* genes. Variants were included in the analysis if they fulfilled the following criteria: 1) biallelic; 2) minor allele frequency (MAF) in the sample not less than 5%; 3) with rsID; 4) *p* value for Hardy-Weinberg equilibrium test larger than 0.001. Analysis was performed separately for the combined UK10K cohort and each of the five European populations in 1000G.

Haplotype network analysis was performed with the R software package, pegas[102]. To reduce the number of SNPs and thus the number of haplotypes included in the analysis, we restricted this analysis to part of the 85 kb *FADS1-FADS2* LD block, starting 5 kb downstream of *FDAS1* to the end of the LD block (a 60-kb region). To further reduce the number of SNPs, in the analysis with all 1000GP European samples, we applied an iterative algorithm[103] to merge haplotypes that have no more than three nucleotide differences by removing the three corresponding SNPs. The algorithm stops when all remaining haplotypes are more than 3 nucleotides away. With this procedure, we were able to reduce the number of total haplotypes from 81 to 12, with the number of SNPs decreased from 88 to 34 (Supplementary Fig. 26). This set of 34 representative SNPs was used in all haplotype-based analysis in aDNA, 1000GP, HGDP and POPRES. Missing data (e.g. from a low imputation genotype probability) were included in the haplotype network analysis.

Of note, for the 12 haplotypes identified in 1000GP European samples, only five of them have frequency higher than 1% (Supplementary Table 2). These five haplotypes were designated as D, M1, M2, M3 and M4, from the most common to the least.

Ancient DNA-based test for recent selection in Europe. The ancient DNA-based selection test was performed as described before[11]. Briefly, most European populations could be modelled as a mixture of three ancient source populations at fixed proportions. The three ancient source populations are West or Scandinavian hunter-gatherers (WSHG), early European farmers (EF), and Steppe-Ancestry pastoralist (SA) (Supplementary Table 1). For modern European populations in 1000G, the ancestral proportions of these three populations estimated at genome-wide level are (0.196, 0.257, 0.547) for CEU, (0.362, 0.229, 0.409) for GBR, (0, 0.686, 0.314) for IBS, and (0, 0.645, 0.355) for TSI. FIN was not used because it does not fit this three-

population model[11]. Under neutrality, the frequencies of a SNP (e.g. reference allele) in present-day European populations are expected to be the linear combination of its frequencies in the three ancient source populations. This serves as the null hypothesis: $p_{mod} = Cp_{anc}$, where p_{mod} is the frequencies in A modern populations (A is always 3 in our test), p_{anc} is the frequencies in B ancient source populations while C is an AxB matrix with each row representing the estimated ancestral proportions for one modern population. The alternative hypothesis is that p_{mod} is unconstrained by p_{anc}. The frequency in each population is modelled with binomial distribution: $L(p; D) = B(X, 2N, p)$, where X is the number of designated allele observed while N is the sample size. In ancient populations, X is the expected number of designated allele observed, taking into account uncertainty in imputation. We write $\ell(p; D)$ for the log-likelihood. The log-likelihood for SNP frequencies in all three ancient populations and four modern populations are: $\ell(\vec{p};\ \vec{D}) = \sum_{i=1}^{A} \ell(p_i; D_i) + \sum_{j=1}^{B} \ell(p_j; D_j)$. Under the null hypothesis, there are A parameters in the model, corresponding to the frequencies in A ancient populations. Under the alternative hypothesis, there are A+B parameters, corresponding to the frequencies in A ancient populations and B modern populations. We numerically maximized the likelihood separately under each hypothesis and evaluate the statistic (twice the difference in log-likelihood) with the null χ_B^2distribution. Inflation was observed with this statistic in a previous genome-wide analysis and a $\lambda = 1.38$ was used for correction in the same cases of three ancient source populations and four present-day European populations [11]. Following this, we applied the same factor in correcting the p values in our analysis. For genotyped SNPs previously tested, similar scales of statistical significance were observed as in the previous study (Supplementary Fig. 27). We note that for the purpose of refining the selection signal with imputed variants, only relative significance levels across variants are informative.

In addition to combining signals from four present-day European populations, we further performed tests separately in the two South European populations (IBS and TSI) and in the two North European populations (CEU and GBR). In these two cases, B = 2 and the null distribution is χ_2^2. No genomic correction was performed for these two cases.

Ancient DNA-based test for ancient selection in pre-Neolithic European hunter-gatherers. A Bayesian method[28] was applied to infer natural selection from allele frequency time series data. The software was downloaded from https://github.com/Schraiber/selection. This method models the evolutionary trajectory of an allele under a specified demographic history and estimates selection coefficients (s_1 and s_2) for heterozygote and homozygote of the allele under study. This method has two modes, with or without the simultaneous estimation of allele age (with or without "-a" in the command line). Without the estimation of allele age, this method models the frequency trajectory only between the first and last time points provided and its estimates of selection coefficients describe the selection force during this period only. With the simultaneous estimation of allele age, this method models the frequency trajectory starting from the first appearance of the allele to the last time point provided. In this case, the selection coefficients describe the selection force starting from the mutation of the allele, which therefore should be the derived allele. For demographic history, we used the model with two historic epochs of bottleneck and recent exponential growth[29]. However, the recent epoch of exponential

growth does not have an impact on our analysis because for our analysis the most recent sample, WSHG, had an age estimate of around 7500 years ago, predating the onset of exponential growth (3520 years ago, assuming 25 years per generation). Four groups of pre-Neolithic European hunter-gatherers were included in our test: the Věstonice cluster (median sample age: 30,076 yo), the El Mirón cluster (14,959 yo), the Villabruna cluster (10,059 yo) and WSHG (7,769 yo).

The use of allele frequency time series data in this Bayesian method makes several assumptions, including 1) all samples are from a randomly mating population with continuity of genetic ancestry; and 2) samples are drawn at different time points[28]. Although there was population structure among pre-Neolithic hunter-gatherers, the four groups used in our study were clustered mainly based on their genetic affinity with additional filtering based on their archaeological contexts[27], therefore population structure in each group was minimized. There is also demonstrated shared genetic ancestry among these groups[27]. Each of the four groups includes samples of different ages and the median sample age was used to represent the sampling time of the group. This approach might introduce noise into the time series and thus reduce the power of the method, making the test conservative. Overall, our time series data do not deviate from these assumptions.

To identify SNPs with evidence of positive selection during the historic period covered by available ancient samples (from Věstonice to WSHG), we first ran the software for most SNPs in the *FADS* locus without the simultaneous estimation of allele age. SNPs with small frequency difference (< 5%) between the first (Věstonice) and last (WSHG) time points were not included in the analysis. For each tested SNP, the allele under analysis was the one showing increasing frequency at the last time point compared to the first. Allele frequency time series data, for each tested SNP, were provided to the software as the expected number of the allele (calculated based on genotype probability) and the sample size. Each software run generated 1,000 Markov chain Monte Carlo (MCMC) samples out of 1,000,000 MCMC simulations with a sampling frequency of every 1,000. The effective sample size of these 1,000 samples were evaluated with the R package, coda[104]. Only runs with effective sample size larger than 50 for four parameters (the sampling likelihood, path likelihood, α_1 estimate, and α_2 estimate) were used[28]. A maximum of 100 runs were attempted for each SNP until a run with sufficient effective sample size was achieved. Otherwise, the SNPs were discarded in our analysis. Visual examination of the observed frequency trajectory for multiple failed SNPs revealed that none of them showed increasing frequency over time and therefore they were unlikely to be under selection. For SNPs with successful software runs, the maximum *a posteriori* (MAP) estimates and the 90% credible intervals (CI) for s_1 and s_2 were calculated. Suggestive evidence for positive selection was called if the 90% CI does not overlap with 0. Second, for the two candidate SNPs (rs174570 and rs2851682) identified in the unbiased global analysis, we further ran the software with the simultaneous estimation of derived allele age. The inference results were plotted with R scripts accompanying the software and additional customized scripts (available upon request).

Modern DNA-based selection tests. We performed two types of selection tests for modern DNAs: site frequency spectrum (SFS)-based and haplotype-based tests. These tests were performed separately in each of the five European populations from 1000G and each of the two cohorts from UK10K. For SFS-based tests, we calculated genetic diversity (π), Tajima's D[105],

and Fay and Wu's H[26], using in-house Perl scripts (available upon request). We calculated these three statistics with a sliding-window approach (window size = 5 kb and moving step = 1 kb). Statistical significance for these statistics were assessed using the genome-wide empirical distribution. Haplotype-based tests, including iHS[24] and nSL[25], were calculated using software selscan (version 1.1.0a)[106]. Only common biallelic variants (Minor allele frequency > 5%) were included in the analysis. Genetic variants without ancestral information were further excluded. These two statistics were normalized in their respective frequency bins (1% interval) and the statistical significance of the normalized iHS and nSL were evaluated with the empirical genome-wide distribution. The haplotype bifurcation diagrams and EHH decay plots were drawn using an R package, rehh[107].

Geographical frequency distribution analysis. For plots of geographical frequency distribution, the geographical map was plotted with R software package, maps (https://CRAN.R-project.org/package=maps) while the pie charts were added with the mapplots package (https://cran.r-project.org/web/packages/mapplots/index.html). Haplotype frequencies were calculated based on haplotype network analysis with pegas[102], which groups haplotypes while taking into account missing data. SNP frequencies were either the observed frequency, if the SNP was genotyped, or the expected frequency based on genotype probability, if the SNP was imputed.

Targeted association analysis for peak SNP rs174594 in UK10K. We performed association analysis for rs174594 in two UK10K datasets – ALSPAC and TwinsUK[99]. For both datasets, we analyzed height, weight, BMI and lipid level related traits including total cholesterol (TC), low density lipoprotein (LDL), very low density lipoprotein (VLDL), high density lipoprotein (HDL), Apolipoprotein A-I (APOA1), Apolipoprotein B (APOB) and triglyceride (TRIG). We performed principal components analysis using smartpca from EIGENSTRAT software[108] with genome-wide autosomal SNPs and we added top 4 principal components as covariates for all association analysis. We also used age as a covariate for all association analysis. Sex was added as a covariate only for ALSPAC dataset since all individuals in TwinsUK dataset are female. For all lipid-related traits, we also added BMI as a covariate.

Data availability.

Ancient DNA: https://reich.hms.harvard.edu/datasets

1000 Genomes Project: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/

Human Genome Diversity Project (HGDP): http://www.hagsc.org/hgdp/files.html

Population Reference Sample (POPRES): dbGaP Study Accession: phs000145.v4.p2

UK10K: https://www.uk10k.org/data_access.html

Code availability. Most analyses were conducted with available software and packages as described in the corresponding subsections of Methods. Customized Perl and R scripts were used in performing SFS-based selection test, and for general plotting purposes. All these scripts are available upon request (Contact K.Y. at ky279@cornell.edu).

References

1 Fan, S., Hansen, M. E. B., Lo, Y. & Tishkoff, S. A. Going global by adapting local: A review of recent human adaptation. *Science* **354**, 54-59 (2016).

2 Nakamura, M. T. & Nara, T. Y. Structure, function, and dietary regulation of delta6, delta5, and delta9 desaturases. *Annu Rev Nutr* **24**, 345-376 (2004).

3 Raphael, W. & Sordillo, L. M. Dietary polyunsaturated fatty acids and inflammation: the role of phospholipid biosynthesis. *Int J Mol Sci* **14**, 21167-21188 (2013).

4 Bazinet, R. P. & Laye, S. Polyunsaturated fatty acids and their metabolites in brain function and disease. *Nat Rev Neurosci* **15**, 771-785 (2014).

5 Mathias, R. A. *et al.* Adaptive evolution of the FADS gene cluster within Africa. *PLoS One* **7**, e44926 (2012).

6 Ameur, A. *et al.* Genetic adaptation of fatty-acid metabolism: a human-specific haplotype increasing the biosynthesis of long-chain omega-3 and omega-6 fatty acids. *Am J Hum Genet* **90**, 809-820 (2012).

7 The 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).

8 Kothapalli, K. S. *et al.* Positive Selection on a Regulatory Insertion-Deletion Polymorphism in FADS2 Influences Apparent Endogenous Synthesis of Arachidonic Acid. *Mol Biol Evol* **33**, 1726-1739 (2016).

9 Fumagalli, M. *et al.* Greenlandic Inuit show genetic signatures of diet and climate adaptation. *Science* **349**, 1343-1347 (2015).

10 Reardon, H. T. *et al.* Insertion-deletions in a FADS2 intron 1 conserved regulatory locus control expression of fatty acid desaturases 1 and 2 and modulate response to simvastatin. *Prostaglandins Leukot Essent Fatty Acids* **87**, 25-33 (2012).

11 Mathieson, I. *et al.* Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**, 499-503 (2015).

12 Bar-Yosef, O. in *On Human Nature: Biology, Psychology, Ethics, Politics, and Religion* (eds M. Tibayrenc & F. J. Ayala) Ch. 19, 297-331 (Academic Press, 2017).

13 Coward, F., Shennan, S., Colledge, S., Conolly, J. & Collard, M. The spread of Neolithic plant economies from the Near East to northwest Europe: a phylogenetic analysis. *Journal of Archaeological Science* **35**, 42-56 (2008).

14 Bogaard, A. *et al.* Crop manuring and intensive land management by Europe's first farmers. *Proc Natl Acad Sci U S A* **110**, 12589-12594 (2013).

15 Richards, M. P. in *The Evolution of Hominin Diets: Integrating Approaches to the Study of Palaeolithic Subsistence* (eds J. J. Hublin & M. P. Richards) 251-257 (Springer Science; Business Media, 2009).

16 Richards, M. P., Schulting, R. J. & Hedges, R. E. Archaeology: sharp shift in diet at onset of Neolithic. *Nature* **425**, 366 (2003).

17 Richards, M. P., Price, T. D. & Koch, E. Mesolithic and Neolithic Subsistence in Denmark:New Stable Isotope Data. *Current Anthropology* **44**, 288-295 (2003).

18 Fraser, R. A., Bogaard, A., Schäfer, M., Arbogast, R. & Heaton, T. H. E. Integrating botanical, faunal and human stable carbon and nitrogen isotope values to reconstruct land use and palaeodiet at LBK Vaihingen an der Enz, Baden-Württemberg. *World Archaeology* **45**, 492-517 (2013).

19 Knipper, C. *et al.* What is on the menu in a Celtic town? Iron Age diet reconstructed at Basel-Gasfabrik, Switzerland. *Archaeological and Anthropological Sciences* (2016).

20 López-Costas, O., Müldner, G. & Martínez Cortizas, A. Diet and lifestyle in Bronze Age Northwest Spain: the collective burial of Cova do Santo. *Journal of Archaeological Science* **55**, 209-218 (2015).
21 Haak, W. *et al.* Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207-211 (2015).
22 Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409-413 (2014).
23 Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065-1093 (2012).
24 Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol* **4**, e72 (2006).
25 Ferrer-Admetlla, A., Liang, M., Korneliussen, T. & Nielsen, R. On detecting incomplete soft or hard selective sweeps using haplotype structure. *Mol Biol Evol* **31**, 1275-1291 (2014).
26 Fay, J. C. & Wu, C. I. Hitchhiking under positive Darwinian selection. *Genetics* **155**, 1405-1413 (2000).
27 Fu, Q. *et al.* The genetic history of Ice Age Europe. *Nature* **534**, 200-205 (2016).
28 Schraiber, J. G., Evans, S. N. & Slatkin, M. Bayesian Inference of Natural Selection from Allele Frequency Time Series. *Genetics* **203**, 493-511 (2016).
29 Gazave, E. *et al.* Neutral genomic regions refine models of recent rapid human population growth. *Proc Natl Acad Sci U S A* **111**, 757-762 (2014).
30 Prufer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43-49 (2014).
31 Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222-226 (2012).
32 Fu, Q. *et al.* Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**, 445-449 (2014).
33 Rasmussen, M. *et al.* The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature* **506**, 225-229 (2014).
34 Rasmussen, M. *et al.* The ancestry and affiliations of Kennewick Man. *Nature* **523**, 455-458 (2015).
35 Lazaridis, I. *et al.* Genomic insights into the origin of farming in the ancient Near East. *Nature* **536**, 419-424 (2016).
36 Tishkoff, S. A. & Verrelli, B. C. Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annu Rev Genomics Hum Genet* **4**, 293-340 (2003).
37 Fu, W. & Akey, J. M. Selection and adaptation in the human genome. *Annu Rev Genomics Hum Genet* **14**, 467-489 (2013).
38 GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648-660 (2015).
39 Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42**, D1001-1006 (2014).
40 Guan, W. *et al.* Genome-wide association study of plasma N6 polyunsaturated fatty acids within the cohorts for heart and aging research in genomic epidemiology consortium. *Circ Cardiovasc Genet* **7**, 321-331 (2014).
41 Dorajoo, R. *et al.* A genome-wide association study of n-3 and n-6 plasma fatty acids in a Singaporean Chinese population. *Genes Nutr* **10**, 53 (2015).
42 Shin, S. Y. *et al.* An atlas of genetic influences on human blood metabolites. *Nat Genet* **46**, 543-550 (2014).

43 Tintle, N. L. *et al.* A genome-wide association study of saturated, mono- and polyunsaturated red blood cell fatty acids in the Framingham Heart Offspring Study. *Prostaglandins Leukot Essent Fatty Acids* **94**, 65-72 (2015).
44 Xie, W. *et al.* Genetic variants associated with glycine metabolism and their role in insulin sensitivity and type 2 diabetes. *Diabetes* **62**, 2141-2150 (2013).
45 Lemaitre, R. N. *et al.* Genetic loci associated with plasma phospholipid n-3 fatty acids: a meta-analysis of genome-wide association studies from the CHARGE Consortium. *PLoS Genet* **7**, e1002193 (2011).
46 Gieger, C. *et al.* Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet* **4**, e1000282 (2008).
47 Kettunen, J. *et al.* Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nat Genet* **44**, 269-276 (2012).
48 Mozaffarian, D. *et al.* Genetic loci associated with circulating phospholipid trans fatty acids: a meta-analysis of genome-wide association studies from the CHARGE Consortium. *Am J Clin Nutr* **101**, 398-406 (2015).
49 Eijgelsheim, M. *et al.* Genome-wide association analysis identifies multiple loci related to resting heart rate. *Hum Mol Genet* **19**, 3885-3894 (2010).
50 den Hoed, M. *et al.* Identification of heart rate-associated loci and their effects on cardiac conduction and rhythm disorders. *Nat Genet* **45**, 621-631 (2013).
51 Global Lipids Genetics Consortium *et al.* Discovery and refinement of loci associated with lipid levels. *Nat Genet* **45**, 1274-1283 (2013).
52 Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707-713 (2010).
53 Aulchenko, Y. S. *et al.* Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nat Genet* **41**, 47-55 (2009).
54 Zabaneh, D. & Balding, D. J. A genome-wide association study of the metabolic syndrome in Indian Asian men. *PLoS One* **5**, e11961 (2010).
55 Kathiresan, S. *et al.* Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat Genet* **41**, 56-65 (2009).
56 Waterworth, D. M. *et al.* Genetic variants influencing circulating lipid levels and risk of coronary artery disease. *Arterioscler Thromb Vasc Biol* **30**, 2264-2276 (2010).
57 Sabatti, C. *et al.* Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat Genet* **41**, 35-46 (2009).
58 Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet* **42**, 1118-1125 (2010).
59 Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet* **47**, 979-986 (2015).
60 Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119-124 (2012).
61 Richards, M. P., Jacobi, R., Cook, J., Pettitt, P. B. & Stringer, C. B. Isotope evidence for the intensive use of marine foods by Late Upper Palaeolithic humans. *J Hum Evol* **49**, 390-394 (2005).
62 Richards, M. P. & Hedges, R. E. M. Stable Isotope Evidence for Similarities in the Types of Marine Foods Used by Late Mesolithic Humans at Sites Along the Atlantic Coast of Europe. *Journal of Archaeological Science* **26**, 717-722 (1999).
63 Lubell, D., Jackes, M., Schwarcz, H. & Knyf, M. The Mesolithic-Neolithic Transition in Portugal:Isotopic and Dental Evidence of Diet. *Journal of Archaeological Science* **21**, 201-216 (1994).

64 Richards, M. P. & Mellars, P. A. Stable isotopes and the seasonality of the Oronsay middens. *Antiquity* **72**, 178-184 (1998).
65 Bonsall, C. *et al.* Mesolithic and Early Neolithic in the Iron Gates: A Palaeodietary Perspective. *Journal of European Archaeology* **5**, 50-92 (1997).
66 Abedi, E. & Sahari, M. A. Long-chain polyunsaturated fatty acid sources and evaluation of their nutritional and functional properties. *Food Sci Nutr* **2**, 443-463 (2014).
67 Simopoulos, A. P. Evolutionary aspects of diet: the omega-6/omega-3 ratio and the brain. *Mol Neurobiol* **44**, 203-215 (2011).
68 Mannino, M. A., Thomas, K. D., Leng, M. J., Di Salvo, R. & Richards, M. P. Stuck to the shore? Investigating prehistoric hunter-gatherer subsistence, mobility and territoriality in a Mediterranean coastal landscape through isotope analyses on marine mollusc shell carbonates and human bone collagen. *Quaternary International* **244**, 88-104 (2011).
69 Mannino, M. A. *et al.* Origin and diet of the prehistoric hunter-gatherers on the mediterranean island of Favignana (Egadi Islands, Sicily). *PLoS One* **7**, e49802 (2012).
70 Lightfoot, E., Boneva, B., Miracle, P. T., Šlaus, M. & O'Connell, T. C. Exploring the Mesolithic and Neolithic transition in Croatia through isotopic investigations. *Antiquity* **85**, 73-86 (2015).
71 Bocquet-Appel, J.-P., Naji, S., Vander Linden, M. & Kozlowski, J. Understanding the rates of expansion of the farming system in Europe. *Journal of Archaeological Science* **39**, 531-546 (2012).
72 Rowley-Conwy, P. Westward Ho! The Spread of Agriculture from Central Europe to the Atlantic. *Current Anthropology* **52**, S431-S451 (2011).
73 Vigne, J.-D. in *The Neolithic Demographic Transition and its Consequences* (eds J.-P. Bocquet-Appel & O. Bar-Yosef) 179-205 (Springer Science+Business Media B.V., 2008).
74 Cramp, L. J. *et al.* Immediate replacement of fishing with dairying by the earliest farmers of the Northeast Atlantic archipelagos. *Proc Biol Sci* **281**, 20132372 (2014).
75 Curry, A. Archaeology: The milk revolution. *Nature* **500**, 20-22 (2013).
76 Salque, M. *et al.* Earliest evidence for cheese making in the sixth millennium BC in northern Europe. *Nature* **493**, 522-525 (2013).
77 Lidén, K., Eriksson, G., Nordqvist, B., Götherström, A. & Bendixen, E. “The wet and the wild followed by the dry and the tame” – or did they occur at the same time? Diet in Mesolithic – Neolithic southern Sweden. *Antiquity* **78**, 23-33 (2004).
78 Rottoli, M. & Castiglioni, E. Prehistory of plant growing and collecting in northern Italy, based on seed remains from the early Neolithic to the Chalcolithic (c. 5600–2100 cal b.c.). *Vegetation History and Archaeobotany* **18**, 91-103 (2008).
79 Cho, H. P., Nakamura, M. & Clarke, S. D. Cloning, expression, and fatty acid regulation of the human delta-5 desaturase. *J Biol Chem* **274**, 37335-37339 (1999).
80 Cho, H. P., Nakamura, M. T. & Clarke, S. D. Cloning, expression, and nutritional regulation of the mammalian Delta-6 desaturase. *J Biol Chem* **274**, 471-477 (1999).
81 Rapoport, S. I., Igarashi, M. & Gao, F. Quantitative contributions of diet and liver synthesis to docosahexaenoic acid homeostasis. *Prostaglandins Leukot Essent Fatty Acids* **82**, 273-276 (2010).
82 DeMar, J. C., Jr., Ma, K., Chang, L., Bell, J. M. & Rapoport, S. I. alpha-Linolenic acid does not contribute appreciably to docosahexaenoic acid within brain phospholipids of adult rats fed a diet enriched in docosahexaenoic acid. *J Neurochem* **94**, 1063-1076 (2005).
83 DeMar, J. C., Jr. *et al.* Brain elongation of linoleic acid is a negligible source of the arachidonate in brain phospholipids of adult rats. *Biochim Biophys Acta* **1761**, 1050-1059 (2006).
84 Qin, X. *et al.* Brown but not white adipose cells synthesize omega-3 docosahexaenoic acid in culture. *Prostaglandins Leukot Essent Fatty Acids* **104**, 19-24 (2016).

85 Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376-381 (2014).
86 Zhang, B. *et al.* Large-scale genetic study in East Asians identifies six new loci associated with colorectal cancer risk. *Nat Genet* **46**, 533-542 (2014).
87 Allentoft, M. E. *et al.* Population genomics of Bronze Age Eurasia. *Nature* **522**, 167-172 (2015).
88 Gamba, C. *et al.* Genome flux and stasis in a five millennium transect of European prehistory. *Nat Commun* **5**, 5257 (2014).
89 Olalde, I. *et al.* Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. *Nature* **507**, 225-228 (2014).
90 Raghavan, M. *et al.* Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* **505**, 87-91 (2014).
91 Jones, E. R. *et al.* Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nat Commun* **6**, 8912 (2015).
92 Fu, Q. *et al.* An early modern human from Romania with a recent Neanderthal ancestor. *Nature* **524**, 216-219 (2015).
93 Seguin-Orlando, A. *et al.* Paleogenomics. Genomic structure in Europeans dating back at least 36,200 years. *Science* **346**, 1113-1118 (2014).
94 Gunther, T. *et al.* Ancient genomes link early farmers from Atapuerca in Spain to modern-day Basques. *Proc Natl Acad Sci U S A* **112**, 11917-11922 (2015).
95 Olalde, I. *et al.* A Common Genetic Origin for Early Farmers from Mediterranean Cardial and Central European LBK Cultures. *Mol Biol Evol* **32**, 3132-3142 (2015).
96 Gallego Llorente, M. *et al.* Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent. *Science* **350**, 820-822 (2015).
97 Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100-1104 (2008).
98 Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456**, 98-101 (2008).
99 The UK10 Consortium *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82-90 (2015).
100 Browning, B. L. & Browning, S. R. Genotype Imputation with Millions of Reference Samples. *Am J Hum Genet* **98**, 116-126 (2016).
101 Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263-265 (2005).
102 Paradis, E. pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics* **26**, 419-420 (2010).
103 Dannemann, M., Andres, A. M. & Kelso, J. Introgression of Neandertal- and Denisovan-like Haplotypes Contributes to Adaptive Variation in Human Toll-like Receptors. *Am J Hum Genet* **98**, 22-33 (2016).
104 Plummer, M., Best, N., Cowles, K. & Vines, K. CODA: convergence diagnosis and output analysis for MCMC. *R News* **6**, 7-11 (2006).
105 Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585-595 (1989).
106 Szpiech, Z. A. & Hernandez, R. D. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol Biol Evol* **31**, 2824-2827 (2014).
107 Gautier, M. & Vitalis, R. rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics* **28**, 1176-1177 (2012).
108 Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-909 (2006).

109 Wang, J. *et al.* Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. *Nucleic Acids Res* **41**, D171-176 (2013).

Acknowledgements

We thank Montgomery Slatkin and Joshua Schraiber for their help in running their software, David Reich and Iain Mathieson for making their data publicly available, Leonardo Arbiza, Charles Liang, Daniel (Alex) Marburgh, Kumar Kothapalli, Tom Brenna, and all members of the Keinan lab for helpful discussion and comments on the manuscript. This work was supported by the National Institutes of Health (Grants R01HG006849 and R01GM108805 to AK) and the Edward Mallinckrodt, Jr. Foundation (AK).

Author contributions

K.Y. and A.K. conceived and designed the project. K.Y. performed the vast majority of data analysis with help from F.G. and D.W.. K.Y. and A.K. interpreted the results, with contribution from O.B.Y. in interpretation from an anthropological perspective. K.Y. and A.K. wrote the manuscript. All authors read, edited and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

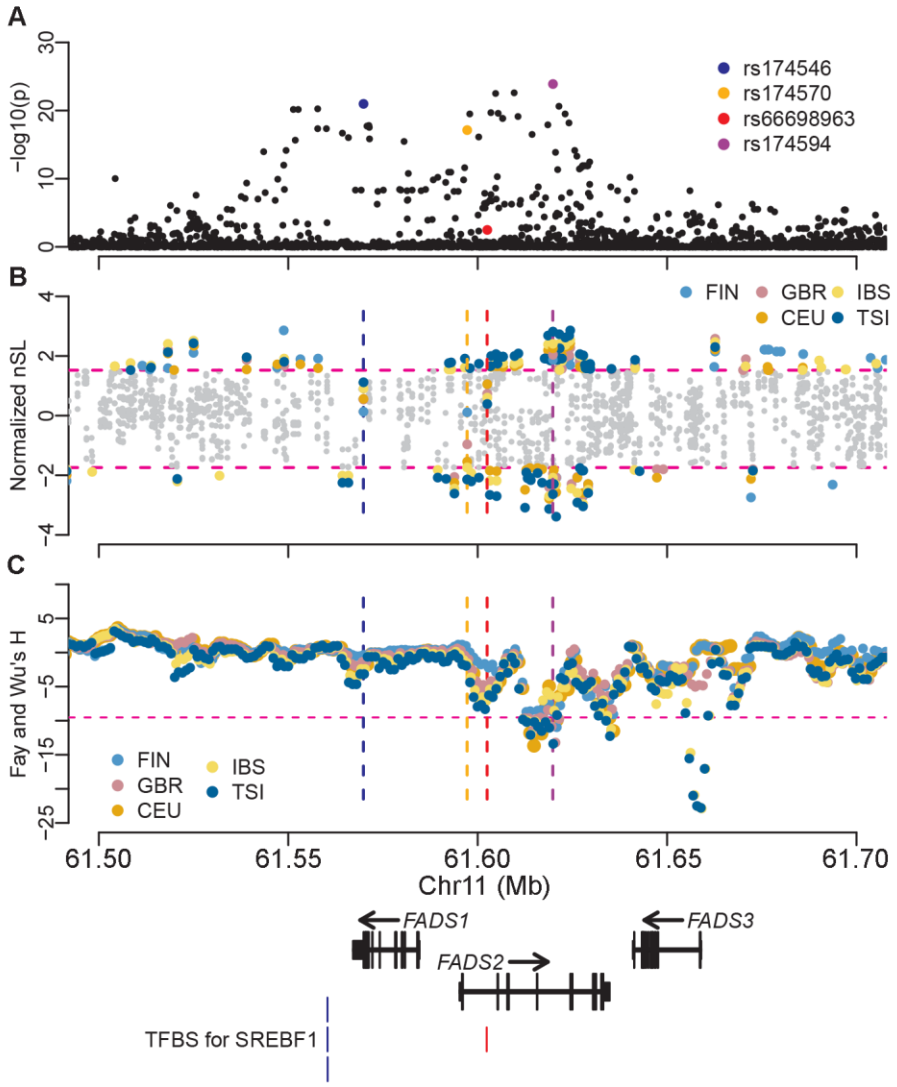


Fig. 1. Recent positive selection on the *FADS* locus in Europeans. (**A**) Ancient DNA-based selection test. The overall pattern is consistent with that previously described[11] (Supplementary Fig. S27). Four variants are highlighted: the most significant SNP (purple); the top SNP reported by Mathieson *et al.*[11] (blue); one of the top adaptive SNPs reported in Greenlandic Inuit[9] (orange); the indel reported to be targeted by positive selection in populations with historical plant-based diets[8] (red). (**B**) Haplotype-based selection test (nSL[25]) in modern Europeans. Tests were performed separately for each of the five 1000GP European groups. Only variants with significant values, beyond the 5% genome-wide significance cutoff (magenta dashed lines) are shown with population-specific colors, with exceptions for the four highlighted variants (positions indicated with vertical dashed lines, colored as in A). Consistent patterns were detected with iHS (Supplementary Fig. 4) and in UK10K cohorts (Supplementary Fig. 6). (**C**) Site frequency spectrum (SFS)-based selection test (Fay and Wu's H[26]) in modern Europeans. Tests were performed separately for each of the five 1000GP European groups. Positions for the four highlighted variants are indicated as in (B). The additional significant result observed around the transcription starting site of *FADS3* is beyond the scope of this paper, and detailed

results will be published separately (Ye, Keinan *et al.*). Similar patterns were observed in UK10K cohorts (Supplementary Fig. 8). At the bottom are the representative transcript models for the three *FADS* genes and the four transcription factor binding sites (TFBS) for SREBF1 from generated from ENCODE[109] (blue) and another previous study[10] (red). The 5% significance cutoffs in (B) and (C) are the most extreme ones among the five populations (but *p* values reported in the main text were based on population-specific empirical distributions). The five 1000 GP European populations are: CEU – Utah Residents (CEPH) with Northern and Western Ancestry; FIN – Finnish in Finland; GRB – British in England and Scotland; IBS – Iberian Population in Spain; TSI – Toscani in Italia.

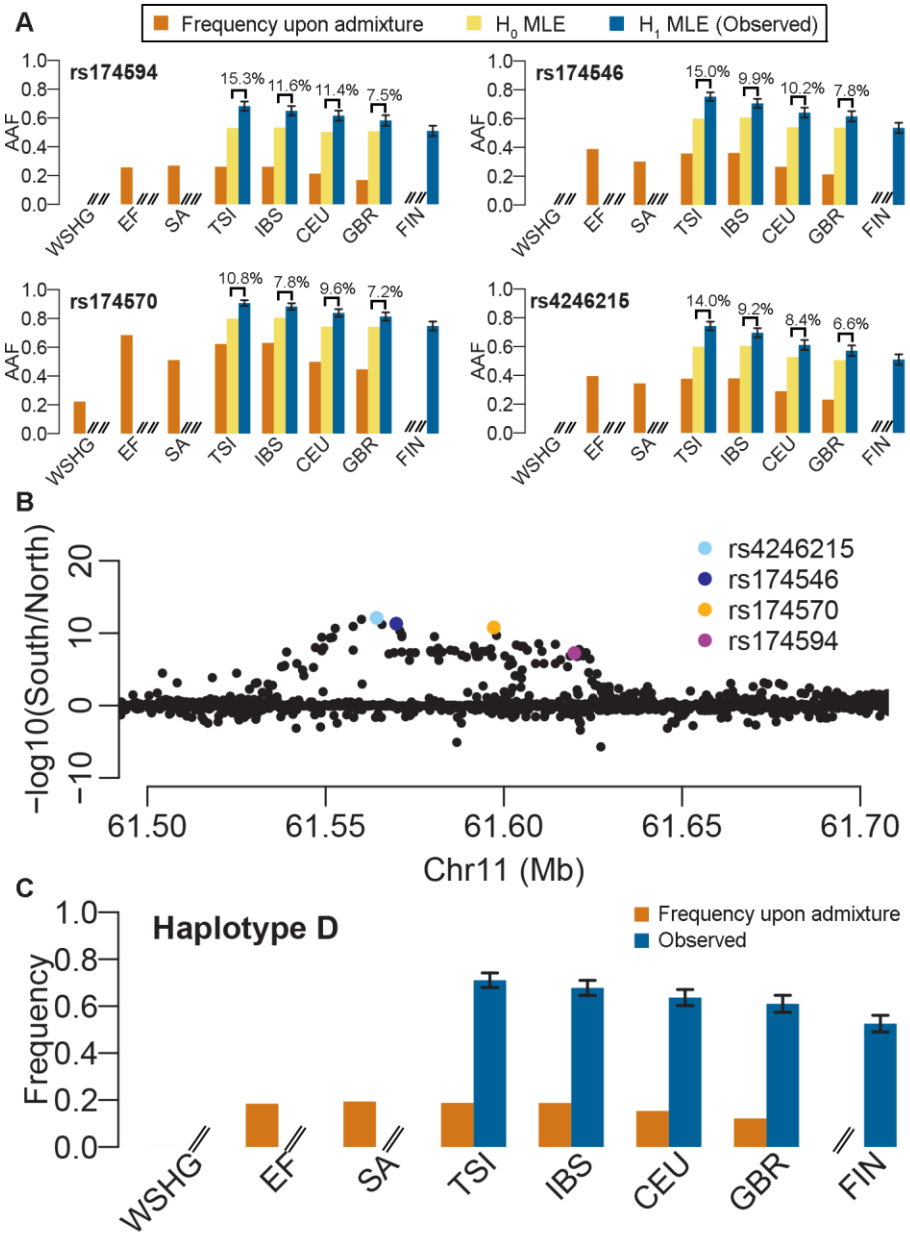
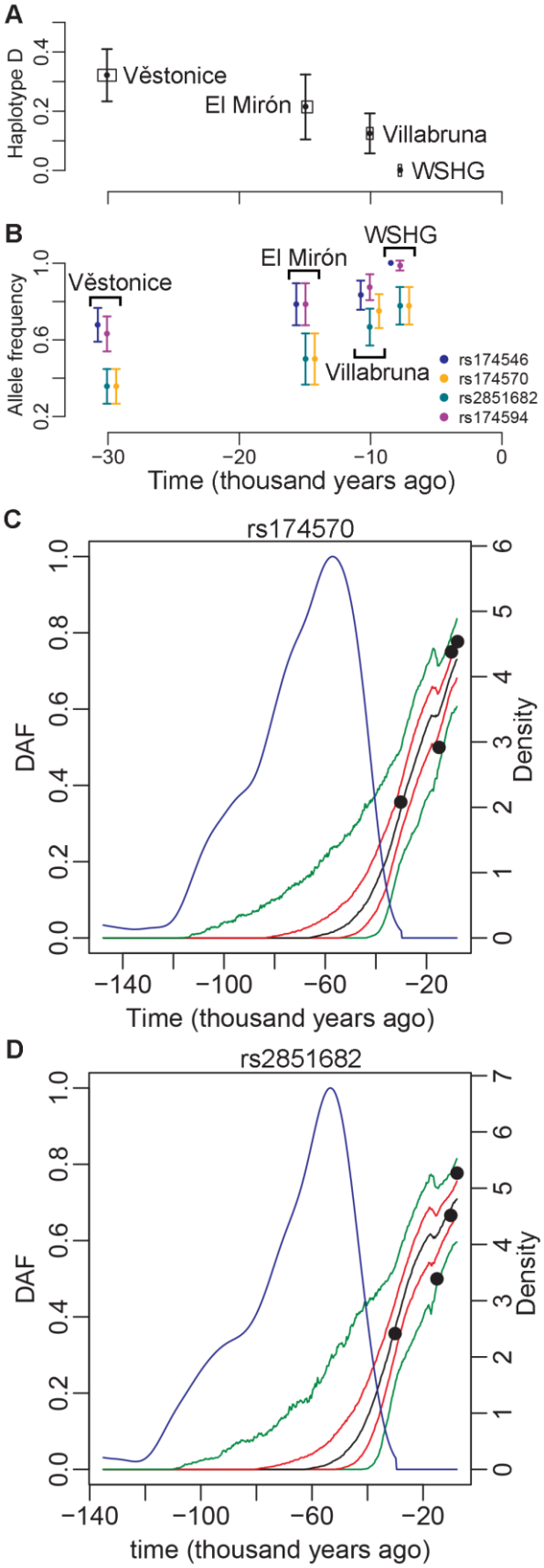


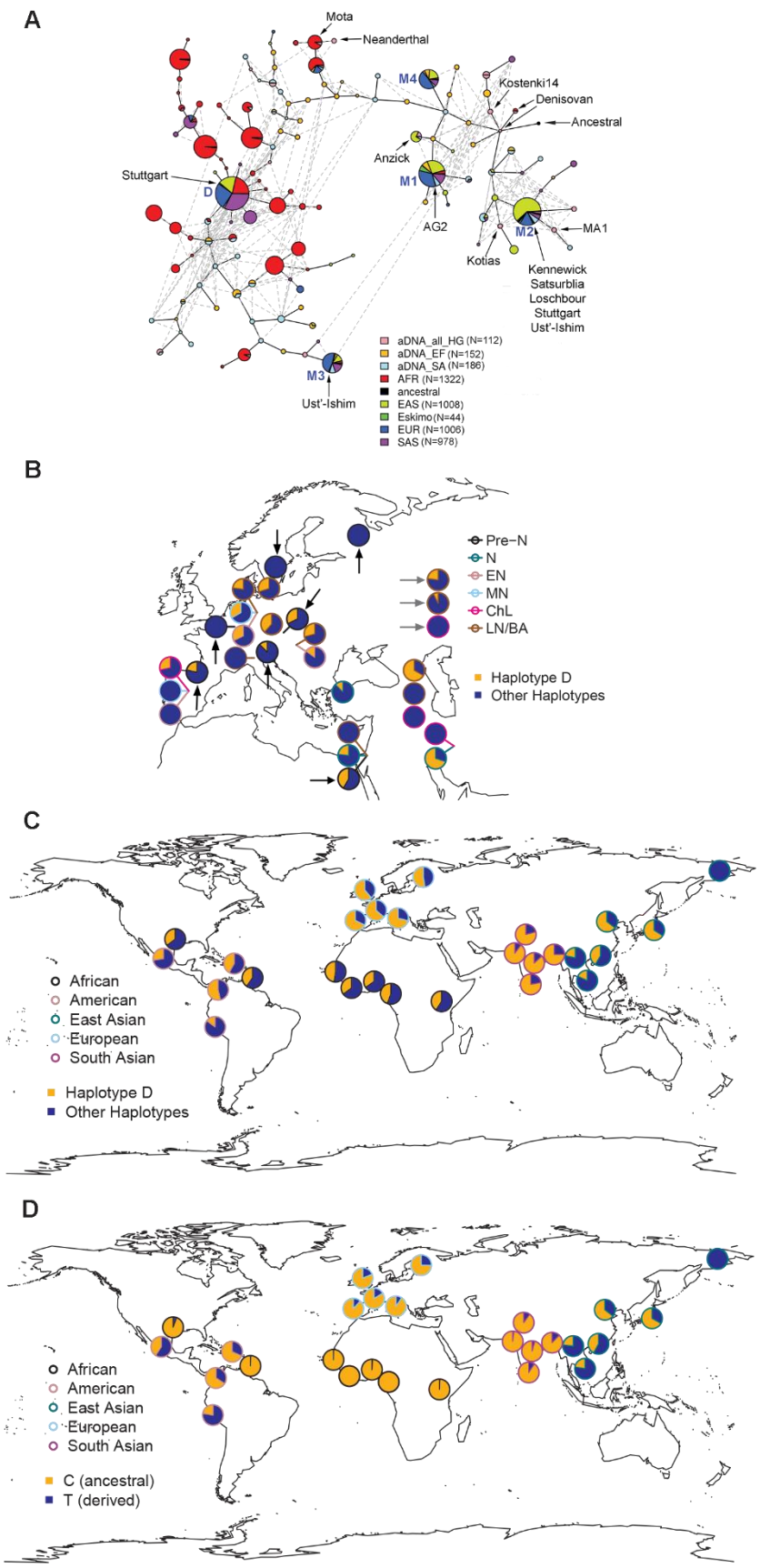
Fig. 2. Varying selection and frequency patterns between Southern and Northern Europe. (**A**) South-North frequency gradient for adaptive alleles of four representative SNPs under different scenarios. AAF refers to adaptive allele frequency. Orange bars represent frequencies upon admixture, which were directly observed in ancient groups and predicted for extant populations based on linear mixture of frequencies in ancient groups. Yellow bars represent frequencies estimated under the null hypothesis. Estimates for ancient groups were not shown because they are not relevant here. Blue bars represent frequencies estimated under the alternative hypothesis, whose only constraint is the observed data and therefore the MLEs are just the observed means. The estimates for ancient groups are the same as their frequencies upon admixture and are omitted on the plot. The absolute difference between H_0 and H_1 estimates are indicated above the corresponding bars. Please note that the frequences upon admixture in

WSHG are 0 for rs174594, rs174546 and rs4246215 and no bars were plotted. (**B**) Comparison of aDNA-based selection signals between Southern and Northern Europe. aDNA-based selection tests were performed separately for Southern (TSI and IBS) and Northern (CEU and GBR) Europeans. For each variant, the *p* values from these two tests were compared at a $-\log_{10}$ scale (y axis). SNPs of interest were colored as indicated. (**C**) South-North frequency gradient for the adaptive haplotype in extant populations. The adaptive haplotype is referred to as haplotype D. The two frequency types are just as in (A). The frequency upon admixture for WSHG is 0. In (A) and (C), FIN has only observed values. If values are not shown or not available, signs of "//" are indicated at corresponding positions. Error bars stand for standard errors.


A
Haplotype D
0.0
0.2
0.4
Věstonice
El Mirón
Villabruna
WSHG
B
Allele frequency
0.2
0.6
1.0
Věstonice
El Mirón
WSHG
Villabruna
rs174546
rs174570
rs2851682
rs174594
−30
−20
−10
0
Time (thousand years ago)
C
rs174570
DAF
Density
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
6
−140
−100
−60
−20
Time (thousand years ago)
D
rs2851682
DAF
Density
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
6
7
−140
−100
−60
−20
time (thousand years ago)

1095

Fig. 3. Temporal frequency pattern and selection signals in pre-Neolithic European hunter-gatherers. (**A**) The frequency of haplotype D over time in four groups of hunter-gatherers. Frequency for each group is plotted as a black point at the median age of samples. The horizontal box surrounding the point represents the medians of lower- and upper-bound estimates of sample ages. The error bar is standard error. Group names are indicated next to their frequencies. (**B**) Allele frequencies for four SNPs. It has similar format as in (A) except that small arbitrary values were added on their x coordinates in order to visualize all SNPs, which were colored as indicated in the legend. The alleles chosen are the ones increasing over time. They are derived alleles for rs174570 and rs2851682, and ancestral alleles for rs174546 and rs174594. (**C**) and (**D**) Posterior distribution on the derived allele frequency path for rs174570 and rs2851682, respectively. The sampled frequencies are indicated with black points, which are the same point estimates as in (B). The median, 25% and 75% quantiles, and 5% and 95% quantiles of the posterior distribution are indicated respectively with black, red and green lines. The posterior distribution on the age of derived allele is shown with a blue line, with values on the right y axis.


A
Mota
Neanderthal
M4
Kostenki14
Denisovan
Ancestral
Anzick
M1
Stuttgart
D
AG2
M2
MA1
Kotias
Kennewick
Satsurblia
Loschbour
Stuttgart
Ust'-Ishim
M3
Ust'-Ishim
aDNA_all_HG (N=112)
aDNA_EF (N=152)
aDNA_SA (N=186)
AFR (N=1322)
ancestral
EAS (N=1008)
Eskimo (N=44)
EUR (N=1006)
SAS (N=978)
B
Pre-N
N
EN
MN
ChL
LN/BA
Haplotype D
Other Haplotypes
C
African
American
East Asian
European
South Asian
Haplotype D
Other Haplotypes
D
African
American
East Asian
European
South Asian
C (ancestral)
T (derived)

1110

Fig. 4. Haplotype network and geographical frequency distribution. (**A**) Haplotype network for 1000G samples (2,157 individuals, excluding admixed American samples), 22 modern Eskimos and 225 aDNAs. Haplotype are defined on the *FADS1-FADS2* LD block. Each pie chart represents one unique haplotype and its size is proportional to $\log_2$(# of chromosomes carrying the haplotype) plus a minimum size so as to visualize the rare haplotypes. The sections in the pie provide the breakdown of the haplotype representation amongst populations with population-specific colors. Sample size for each group (# of haplotypes) is indicated in the legend. Because of the small sample size of some groups, it can be difficult to visualize their proportions in this plot. Detailed haplotype frequencies for each group could be found in Supplementary Table 2. The edges connecting haplotypes are of arbitrary length. Haplotypes for some well-known ancient samples are indicated with their names and arrows pointing at the corresponding pies. The top five haplotypes in modern Europeans, referred to as D, M1, M2, M3, and M4 from the most to least frequent common, are indicated with their names in blue font right next to the haplotypes. (**B**) Frequency of haplotype D in Eurasian ancient DNAs. Each pie represents one sampled group and is placed at the sampling location or nearby with a line pointing at the sampling location. The color of the pie chart border and its associated line indicates the archaeological period of the sample. If multiple samples of different periods were collected at the same geographical location, these samples are ordered vertically with the older samples at the bottom. Hunter-gatherer groups are indicated with black arrows and pastoralist groups with gray arrows, while others are farmers. Geographical locations for some hunter-gatherer groups (e.g. the Věstonice, El Mirón and Villabruna clusters) are only from representative samples. Detailed frequencies could be found in Supplementary Table 3. Pre-N: Pre-Neolithic; N: Neolithic; EN: Early Neolithic; MN: Mid-Neolithic; ChL: Chalcolithic; LN/BA: Late Neolithic/Bronze Age. (**C**) Frequency of haplotype D in present-day global populations. Detailed frequencies are given in Supplementary Table 4. (**D**) Frequency of SNP rs174570 in present-day global populations. All 26 populations from 1000GP and one Eskimo group are included. The color of the pie chart border represents the genetic ancestry. It is noteworthy that there are two samples in America that are actually of African ancestry. Similar global patterns were observed with HGDP samples (Supplementary Figs. 14 and 15).