THE SELECTION LANDSCAPE AND GENETIC LEGACY OF ANCIENT EURASIANS

Evan K. Irving-Pease[1]§, Alba Refoyo-Martínez[1]§, Andrés Ingason[2,1]§, Alice Pearson[3,4]§, Anders Fischer[5,6]§, William Barrie[7]§, Karl-Göran Sjögren[8], Alma S. Halgren[9], Ruairidh Macleod[7,10], Fabrice Demeter[1,11], Rasmus A. Henriksen[1], Tharsika Vimala[1], Hugh McColl[1], Andrew Vaughn[12], Aaron J. Stern[12], Leo Speidel[13,14], Gabriele Scorrano[1], Abigail Ramsøe[1], Andrew J. Schork[2,15], Anders Rosengren[2,1], Lei Zhao[1], Kristian Kristiansen[16,1], Peter H. Sudmant[9,12]@, Daniel J. Lawson[17]@, Richard Durbin[3,18]@, Thorfinn Korneliussen[1]@, Thomas Werge[19,20]@, Morten E. Allentoft[1,21]@, Martin Sikora[1]@, Rasmus Nielsen[22,1]*@, Fernando Racimo[1]*@, Eske Willerslev[1,7]*@

Affiliations
[1]Lundbeck Foundation GeoGenetics Centre, Globe Institute, University of Copenhagen, Copenhagen, Denmark, [2]Institute of Biological Psychiatry, Mental Health Services, Copenhagen University Hospital, Roskilde, Denmark, [3]Department of Genetics, University of Cambridge, UK, [4]Department of Zoology, University of Cambridge, UK, [5]Department of Historical Studies, University of Gothenburg, 405 30 Gothenburg, Sweden, [6]Sealand Archaeology, Gl. Roesnaesvej 27, 4400 Kalundborg, Denmark, [7]GeoGenetics Group, Department of Zoology, University of Cambridge, UK, [8]Department of Historical Studies, Gothenburg University, Sweden, [9]Department of Integrative Biology, University of California Berkeley, [10]Research department of Genetics, Evolution and Environment, University College London, [11]National Museum of Natural History, Paris, France, [12]Center for Computational Biology, University of California, Berkeley, USA, [13]Ancient Genomics Laboratory, The Francis Crick Institute, London, UK, [14]Genetics Institute, University College London, London, UK, [15]Neurogenomics Division, The Translational Genomics Research Institute (TGEN), Phoenix, AZ, USA, [16]Department of Historical Studies, University of Gothenburg, SE-41255, Gothenburg, Sweden, [17]Institute of Statistical Sciences, School of Mathematics, University of Bristol, Bristol, UK, [18]Wellcome Sanger Institute, Cambridge, UK, [19]Department of Clinical Medicine and Lundbeck Center for Geogenetics, GLOBE Institute, University of Copenhagen, [20]Institute of Biological Psychiatry, Mental Health Center Sct Hans, Copenhagen University Hospital, Denmark, [21]Trace and Environmental DNA (TrEnD) Laboratory, School of Molecular and Life Science, Curtin University, Australia, [22]Departments of Integrative Biology and Statistics, UC Berkeley, Berkeley 94720, USA

* Corresponding authors; email: rasmus_nielsen@berkeley.edu, fracimo@sund.ku.dk, ew482@cam.ac.uk
§ Joint first authors
@ Joint last authors

Summary

The Eurasian Holocene (beginning c. 12 thousand years ago) encompassed some of the most significant changes in human evolution, with far-reaching consequences for the dietary, physical and mental health of present-day populations. Using an imputed dataset of >1600 complete ancient genome sequences, and new computational methods for locating selection in time and space, we reconstructed the selection landscape of the transition from hunting and gathering, to farming and pastoralism across West Eurasia. We identify major selection signals related to metabolism, possibly associated with the dietary shift occurring in this period. We show that the selection on loci such as the FADS cluster, associated with fatty acid metabolism, and the lactase persistence locus, began earlier than previously thought. A substantial amount of selection is also found in the HLA region and other loci associated with immunity, possibly due to the increased exposure to pathogens during the Neolithic, which may explain the current high prevalence of auto-immune disease, such as psoriasis, due to genetic trade-offs. By using ancient populations to infer local ancestry tracks in hundreds of thousands of samples from the UK Biobank, we find strong genetic differentiation among ancient Europeans in loci associated with anthropometric traits and susceptibility to several diseases that contribute to present-day disease burden. These were previously thought to be caused by local selection, but in fact can be attributed to differential genetic contributions from various source populations that are ancestral to present-day Europeans. Thus, alleles associated with increased height seem to have

increased in frequency following the Yamnaya migration into northwestern Europe around 5,000 years ago. Alleles associated with increased risk of some mood-related phenotypes are overrepresented in the farmer ancestry component entering Europe from Anatolia around 11,000 years ago, while western hunter-gatherers show a strikingly high contribution of alleles conferring risk of traits related to diabetes. Our results paint a picture of the combined contributions of migration and selection in shaping the phenotypic landscape of present-day Europeans that suggests a combination of ancient selection and migration, rather than recent local selection, is the primary driver of present-day phenotypic differences in Europe.

Introduction

One of the central goals of human evolutionary genetics is to understand how natural selection has shaped the genomes of present-day people in response to changes in culture and environment. The transition from hunting and gathering, to farming, and subsequently pastoralism, during the Holocene in Eurasia, involved some of the most dramatic changes in diet, health and social organisation experienced during recent human evolution. The dietary changes, and the expansions into new climate zones, represent major shifts in environmental exposure, impacting the evolutionary forces acting on the human gene pool. These changes imposed a series of large-scale heterogeneous selection pressures on humans, beginning around 12,000 years ago and extending to the present-day. As human lifestyles changed, close contact with domestic animals and higher population densities are likely to have increased exposure to, and transmission of, infectious diseases; introducing new challenges to our survival [1,2].

Our understanding of the genetic architecture of complex traits in humans has been substantially advanced by genome-wide association studies (GWAS) of present-day populations, which have identified large numbers of genetic variants associated with phenotypes of interest [3–5]. However, the extent to which these variants have been under directional selection during recent human evolution remains unclear, and the highly polygenic nature of most complex traits makes identifying selection difficult. While signatures of selection can be identified from patterns of genetic diversity in extant populations [6,7], this can be challenging in species such as humans, which show very wide geographic distributions and have thus been exposed to highly diverse and dynamic local environments through time and space. In the complex mosaic of ancestries that constitute a modern human genome, any putative signatures of selection may therefore misrepresent the timing and magnitude of the selective process. Similarly, episodes of admixture between ancestral populations can result in present-day haplotypes which contain no evidence of selective processes occurring further back in time. Ancient DNA (aDNA) provides the potential to resolve these issues, by directly observing changes in trait associated allele frequencies over time.

Whilst numerous prior studies have used ancient DNA to infer patterns of selection in Eurasia during the Holocene (e.g., [8–10]), many key questions remain unanswered. To what extent are present-day differences in human phenotypes due to natural selection or to differing proportions of ancient ancestry? What are the genetic legacies of Stone Age hunter-gatherer groups in present-day complex traits? How has the complicated admixture history of Holocene Eurasia affected our ability to detect natural selection in genetic data? To investigate these questions, and the selective landscape of Eurasian prehistory, we conducted the largest ancient DNA study to date of human Stone Age skeletal material, generating a phased and imputed dataset of >1,600 ancient genomes [11].

To test for traces of divergent selection in health and lifestyle-related genetic variants, we used the imputed ancient genomes to reconstruct polygenic risk scores for hundreds of complex traits in ancient Eurasian populations. Additionally, we reconstructed the allele frequency trajectories and selection coefficients of tens of thousands of trait associated variants through time. We used a novel chromosome painting technique, based on tree sequences, in order to model ancestry-specific allele frequency trajectories through time. This allows us to identify many trait-associated genetic variants with hitherto unknown evidence for positive selection, as well as resolve long-standing questions about the timing of selection for key health, dietary and pigmentation associated loci.

Results/Discussion

Samples and data

Our analyses are undertaken on an unprecedented sample of shotgun-sequenced ancient genomes, presented in the accompanying study 'Population Genomics of Stone Age Eurasia' [11]. This dataset comprises 1664 imputed diploid ancient genomes and more than 8.5 million SNPs. These represent a considerable transect of Eurasia, ranging longitudinally from the Atlantic coast to Lake Baikal, and latitudinally from Scandinavia to the Middle East. Included are many of the key Mesolithic and Neolithic cultures of Western Eurasia, Ukraine, western Russia, and the Trans-Urals, constituting a thorough temporal sequence of human populations from 11,000 cal. BP to 3,000 cal. BP. Additionally, an especially dense sample set from Denmark represents continuous populations from the Mesolithic to the Bronze Age. This dataset allows us to characterise in unprecedented detail the changes in selective pressures exerted by major transitions in human culture and environment.

Ancestry-stratified patterns of natural selection in the last 13,000 years

To account for population structure in our samples, we developed a novel chromosome painting technique that allows us to accurately assign ancestral population labels to haplotypes found in both ancient and present-day individuals. We built a quantitative admixture graph model (Fig. 1; Supplementary Note 1a) that represents the four major ancestry flows contributing to modern European genomes over the last 50,000 years [12]. We used this model to simulate genomes at time periods and in sample sizes equivalent to our empirical aDNA dataset, then inferred tree sequences using Relate [13,14]. We then trained a neural network classifier to estimate the path backwards in time through the population structure taken by each simulated individual, at each position in the genome. We then applied our trained classifier to infer the ancestral paths taken at each site using 1,015 imputed ancient genomes from West Eurasia (Fig. 1; Supplementary Note 1a).

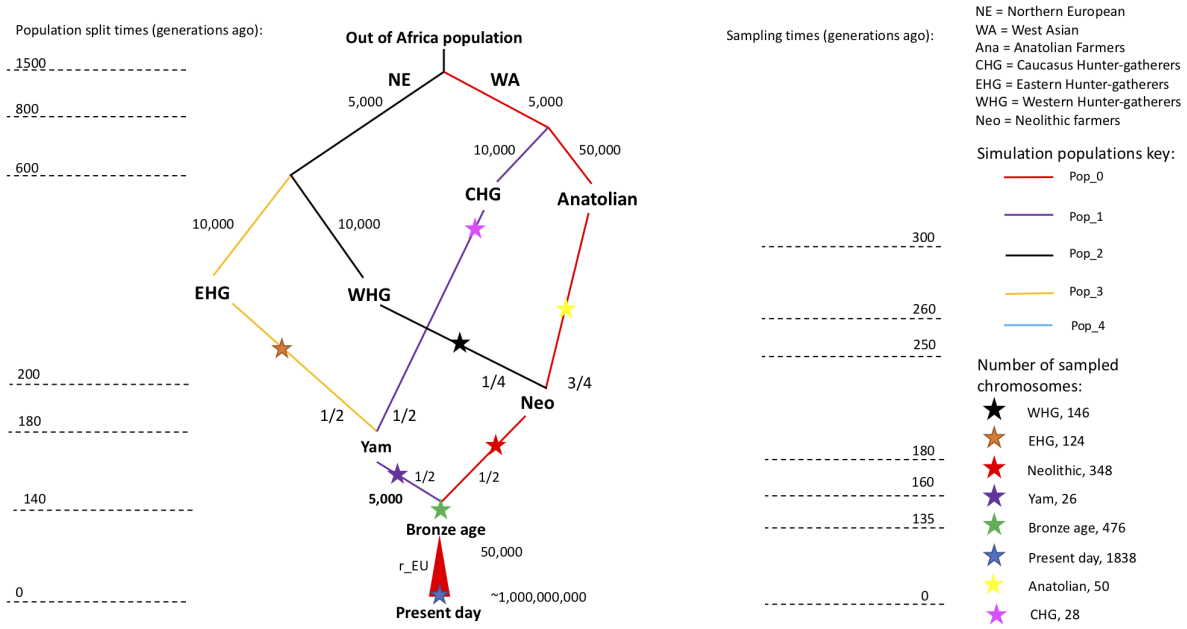


Fig 1. A schematic of the model of population structure in Europe, used to simulate genomes to train the local ancestry neural network classifier. Moving down the figure is forwards in time and the population split times and admixture times are given in generations ago. Coloured lines represent the four populations declared in the simulation that extend through time. Sampled populations and times are marked with a star and the number of chromosomes sampled is given in the key.

We adapted CLUES [15] to model time-series data (Supplementary Note 2a) and used it to infer allele frequency trajectories and selection coefficients for 33,323 quality-controlled phenotype-associated variants ascertained from the GWAS Catalogue [16]. An equal number of putatively neutral, frequency-paired variants were used as a control set. To control for possible confounders, we built a causal model to distinguish direct effects of age on allele frequency from indirect effects mediated by read depth, read length, and/or error rates (Supplementary Note 2b), and developed a mapping bias test used to evaluate systematic differences between data from ancient and present-day populations (Supplementary Note 2a). Because admixture between groups with differing allele frequencies can confound interpretation of allele frequency changes through time, we also applied a novel chromosome painting technique, based on inference of a sample's nearest neighbours in the marginal trees of a tree sequence (Supplementary Note 1a). This allowed us to accurately assign ancestral path labels to haplotypes found in both ancient and present-day individuals. By conditioning on these haplotype path labels, we are able to infer selection trajectories while controlling for changes in admixture proportions through time (Supplementary Note 2a).

Our analysis identified no genome-wide significant ($p < 5e-8$) selective sweeps when using genomes from present-day individuals alone (1000 Genomes Project populations GBR, FIN and TSI), although trait-associated variants were enriched for signatures of selection compared to the control group ($p < 2.2e-16$, Wilcoxon signed-rank test). In contrast, when using imputed aDNA genotype probabilities, we identified 11 genome-wide significant selective sweeps in the GWAS variants, and none in the control group, consistent with selection acting on trait-associated variants (Supplementary Note 2a, Supplementary Figs. S2a.4 to S2a.14). However, when conditioned on one of our four local ancestry paintings—genomic regions arriving in present day genomes through either Western hunter-gatherers (WHG), Eastern hunter-gatherers (EHG), Caucasus hunter-

gatherers (CHG) or Anatolian farmers (ANA)—we identified 21 genome-wide significant selection peaks (including the 11 from the pan-ancestry analysis) (Fig. 2). This suggests that admixture between ancestral populations has masked evidence of selection at many trait associated loci in modern populations.

Selection on diet-associated loci

We find strong changes in selection associated with lactose digestion after the introduction of farming, but prior to the expansion of the Yamnaya pastoralists into Europe around 5,000 years ago [17,18], settling controversies regarding the timing of this selection [19–22]. The strongest overall signal of selection in the pan-ancestry analysis is observed at the *MCM6 / LCT* locus (rs4988235; p=9.86e-31; s=0.020), where the derived allele results in lactase persistence [23,24] (Supplementary Note 2a). The trajectory inferred from the pan-ancestry analysis indicates that the lactase persistence allele began increasing in frequency c. 6,000 years ago, and has continued to increase up to present times (Fig. 2). In the ancestry-stratified analyses, this signal is driven primarily by sweeps in only two of the ancestral backgrounds, EHG and CHG. However, we also observed that many selected SNPs within this locus exhibited earlier evidence of selection than at rs4988235, suggesting that selection at the *MCM6/LCT* locus is more complex than previously thought. To investigate this further, we expanded our selection scan to include all SNPs within the ~2.3 Mbp wide sweep locus (n=5,608), and checked for the earliest evidence of selection in our pan-ancestry analysis (Supplementary Note 2a). To control for potential bias introduced by imputation, we also inferred trajectories using genotype likelihoods, and confirmed that results were consistent between models. We observed that the vast majority of genome-wide significant SNPs at this locus began rising in frequency earlier than rs4988235, indicating that strong positive selection at this locus predates the emergence of the lactase persistence allele by thousands of years. Among the SNPs showing the earliest frequency rises was rs1438307 (p=1.17e-12; s=0.015), which began rising in frequency c. 12,000 years ago (Fig. 2). This SNP, which has been shown to regulate energy expenditure and contributes to metabolic disease, has been hypothesised as an ancient adaptation to famine [25]. The high linkage disequilibrium between rs1438307 and rs4988235 in present-day individuals (R^2 = 0.8943 in GBR) may explain the recently observed correlation between frequency rises in the lactase persistence allele and archaeological proxies for famine and increased pathogen exposure [26].

We also find strong selection in the FADS gene cluster — *FADS1* (rs174546; p=2.65e-10; s=0.013) and *FADS2* (rs174581; p=1.87e-10; s=0.013) — which are associated with fatty acid metabolism and known to respond to changes in diet from a more/less vegetarian to a more/less carnivorous diet [27–32]. In contrast to previous results [30–32], we find that much of the selection associated with a more vegetarian diet occurred in Neolithic populations before they arrived in Europe, but then continued during the Neolithic (Fig. 2). The strong signal of selection in this region in the pan-ancestry analysis is driven primarily by a sweep occurring across the EHG, CHG and ANA haplotypic backgrounds (Fig. 2). Interestingly, we find no evidence for selection at this locus in the WHG background, and most of the allele frequency rise in the EHG background occurs after their admixture with CHG (around 8 Kya, [33]), within whom the selected alleles were already close to present-day frequencies. This suggests that the selected alleles may already have existed at substantial frequencies in early farmer populations in the Middle East and among Caucasus Hunter gatherers (associated with the ANA and CHG and backgrounds, respectively) and were subject to continued selection as eastern groups moved northwards and westwards during the late Neolithic and Bronze Age periods.

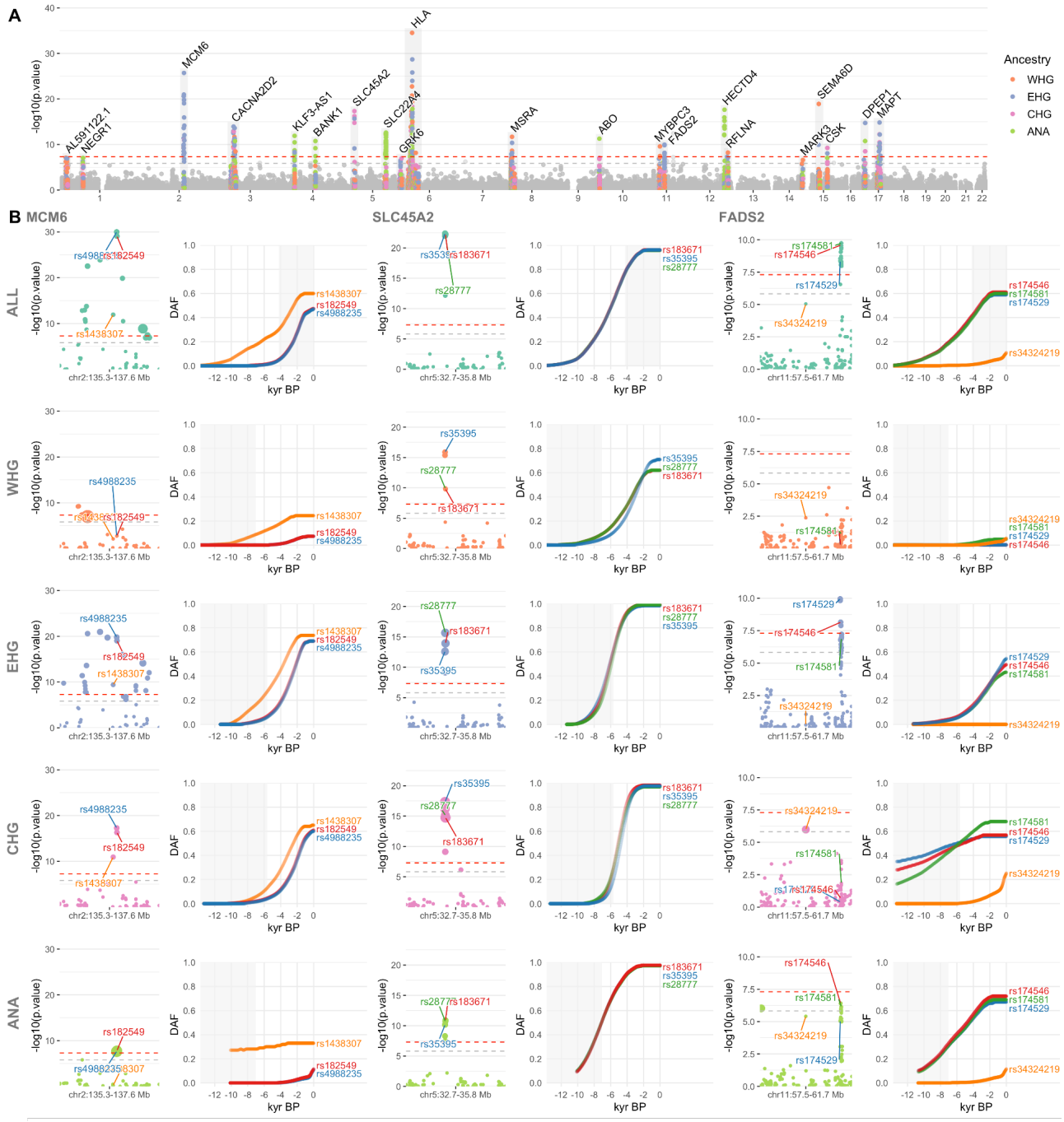


Fig 2. Genome-wide selection scan for trait associated variants. A) Manhattan plot of p-values from selection scan with CLUES, based on a time-series of imputed aDNA genotype probabilities. Twenty-one genome-wide significant selection peaks highlighted in grey and labelled with the most significant gene within each locus. Within each sweep, SNPs are positioned on the y-axis and coloured by their most significant marginal ancestry. Outside of the sweeps, SNPs show p-values from the pan-ancestry analysis and are coloured grey. Red dotted lines indicate genome-wide significance ($p < 5e-8$), while the grey dotted line shows the Bonferroni significance threshold, corrected for the number of tests ($p < 1.35e-6$). B) Detailed plots for three genome-wide significant sweep loci: (i) MCM6, lactase persistence; (ii) SLC45A2, skin pigmentation; and (iii) FADS2, lipid metabolism. Rows show results for the pan-ancestry analysis (ALL) plus the four marginal ancestries: Western hunter-gatherers (WHG), Eastern hunter-gatherers (EHG), Caucasus hunter-gatherers (CHG) and Anatolian farmers (ANA). The first column of each loci shows zoomed Manhattan plots of the p-values for each ancestry (significant SNPs sized by their selection coefficients), and column two shows allele

trajectories for the top SNPs across all ancestries (grey shading for the marginal ancestries indicates approximate temporal extent of the pre-admixture population).

When specifically comparing selection signatures differentiating ancient hunter-gatherer and farmer populations [34], we also observe a large number of regions associated with lipid and sugar metabolism, and various metabolic disorders (Supplementary Note 2d). These include, for example, a region in chromosome 22 containing *PATZ1*, which regulates the expression of *FADS1*, and *MORC2*, which plays an important role in cellular lipid metabolism [35–37]. Another region in chromosome 3 overlaps with *GPR15,* which is both related to immune tolerance and to intestinal homeostasis [38–40]. Finally, in chromosome 18, we recover a selection candidate region spanning *SMAD7,* which is associated with inflammatory bowel diseases such as Crohn's disease [41–43]. Taken together these results suggest that the transition to agriculture imposed a substantial amount of selection for humans to adapt to our new diet and that some diseases observed today in modern societies can likely be understood as a consequence of this selection.

Selection on immunity-associated variants

In addition to diet-related selection, we observe selection in several loci associated with immunity/defence functions and with autoimmune disease (Supplementary Note 2a). Some of these selection events occurred earlier than previously claimed and are likely associated with the transition to agriculture and may help explain the high prevalence of autoimmune diseases today. Most notably, we detect a 33 megabase (Mb) wide selection sweep signal in chromosome 6 (chr6:19.1–50.9 Mb), spanning the human leukocyte antigen (HLA) region (Supplementary Note 2a). The selection trajectories of the variants within this locus support multiple independent sweeps, occurring at different times and with differing intensities. The strongest signal of selection at this locus in the pan-ancestry analysis is at an intergenic variant, located between *HLA-A* and *HLA-W* (rs7747253; p=8.86e-17; s=-0.018), associated with heel bone mineral density [44], the derived allele of which rapidly reduced in frequency, beginning c. 8,000 years ago (Extended Data Fig. 1). In contrast, the signal of selection at *C2* (rs9267677; p= 9.82e-14; s= 0.04463), also found within this sweep, and associated with psoriasis risk in UK Biobank (p=4.1e-291; OR=2.2) [45,46], shows a gradual increase in frequency beginning c. 4,000 years ago, before rising more rapidly c. 1,000 years ago. This locus might provide a good example of the hypothesis that the high prevalence of auto-immune diseases in modern populations may, in part, be due to genetic trade-offs by which selection increasing the defence against pathogens also have the pleiotropic effect of increasing susceptibility to auto-immune diseases [47,48].

These results also highlight the complex temporal dynamics of selection at the HLA locus, which not only plays a role in the regulation of the immune system, but also has association with many non-immune-related phenotypes. The high pleiotropy in this region makes it difficult to determine which selection pressures may have driven these increases in frequencies at different periods of time. However, profound shifts in lifestyle in Eurasian populations during the Holocene, including a change in diet and closer contact with domestic animals, combined with higher mobility and increasing population sizes, are likely drivers for strong selection on loci involved in immune response.

We also identified selection signals at the *SLC22A4* (rs35260072; p=1.15e-10; s=0.018) locus, associated with increased itch intensity from mosquito bites [49], and find that the derived variant has been steadily rising in frequency since c. 9,000 years ago (Extended Data Fig. 2). However, in the same *SLC22A4* candidate region as rs35260072, we find that the frequency of the previously reported SNP rs1050152 plateaued c. 1,500 years ago, contrary to previous reports suggesting a

recent rise in frequency [8]. Similarly, we detect selection at the *HECTD4* (rs11066188; p=3.02e-16; s=0.020) and *ATXN2* (rs653178; p=1.92e-15; s=0.019) loci, associated with celiac disease and rheumatoid arthritis [50], which has been rising in frequency for c. 9,000 years (Extended Data Fig. 3), also contrary to previous reports of a more recent rise in frequency [8]. Thus, several disease-associated loci previously thought to be the result of recent adaptation may have been subject to selection for a longer period of time.

Selection on the 17q21.31 locus

We further detect signs of strong selection in a 12 Mb sweep on chromosome 17 (chr17:36.1–48.1 Mb), spanning a locus on 17q21.3 implicated in neurodegenerative and developmental disorders (Supplementary Note 2a). The locus includes an inversion and other structural polymorphisms with indications of a recent positive selection sweep in some human populations [51,52]. Specifically, partial duplications of the *KANSL1* gene likely occurred independently on the inverted (H2) and non-inverted (H1) haplotypes (Fig. 3B) and both are found in high frequencies (15-25%) among current European and Middle Eastern populations but are much rarer in Sub-Saharan African and East Asian populations. We used both SNP genotypes and WGS read depth information to determine inversion (H1/H2) and *KANSL1* duplication (d) status in the ancient individuals studied here (Supplementary Note 2f).

The H2 haplotype is observed in two of three previously published genomes [53] of Anatolian aceramic Neolithic individuals (Bon001 and Bon004) from around 10,000 BP, but data were insufficient to identify KANSL1 duplications. The oldest evidence for KANSL1 duplications is observed in an early Neolithic individual (AH1 from 9,900 BP [54]) from present-day Iran, followed by two Mesolithic individuals (NEO281 from 9,724 BP and KK1 [55] from 9,720 BP), from present-day Georgia, all of whom are heterozygous for the inversion and carry the inverted duplication. The KANSL1 duplications are also detected in two Neolithic individuals, from present-day Russia (NEO560 from 7,919 BP (H1d) and NEO212 from 7,390 BP (H2d)). With both H1d and H2d having spread to large parts of Europe with Anatolian Neolithic Farmer ancestry, their frequency seems unchanged in most of Europe as Steppe-related ancestry becomes dominant in large parts of the subcontinent (Extended Data Fig. 3D). The fact that both H1d and H2d are found in apparently high frequencies in both early Anatolian Farmers and the earliest Yamnaya/Steppe-related ancestry groups suggests that any selective sweep acting on the H1d and H2d variants would probably have occurred in populations ancestral to both.

We note that the strongest signal of selection observed in this locus is at *MAPT* (rs4792897; p=4.65e-10; s=0.03 (Fig. 3A; Supplementary Note 2a), which codes for the tau protein [56] and is involved in a number of neurodegenerative disorders, including Alzheimer's disease and Parkinson's disease [57–61]. However, the region is also enriched for evidence of reference bias in our dataset—especially around the *KANSL1* gene—due to complex structural polymorphisms (Supplementary Note 2h).

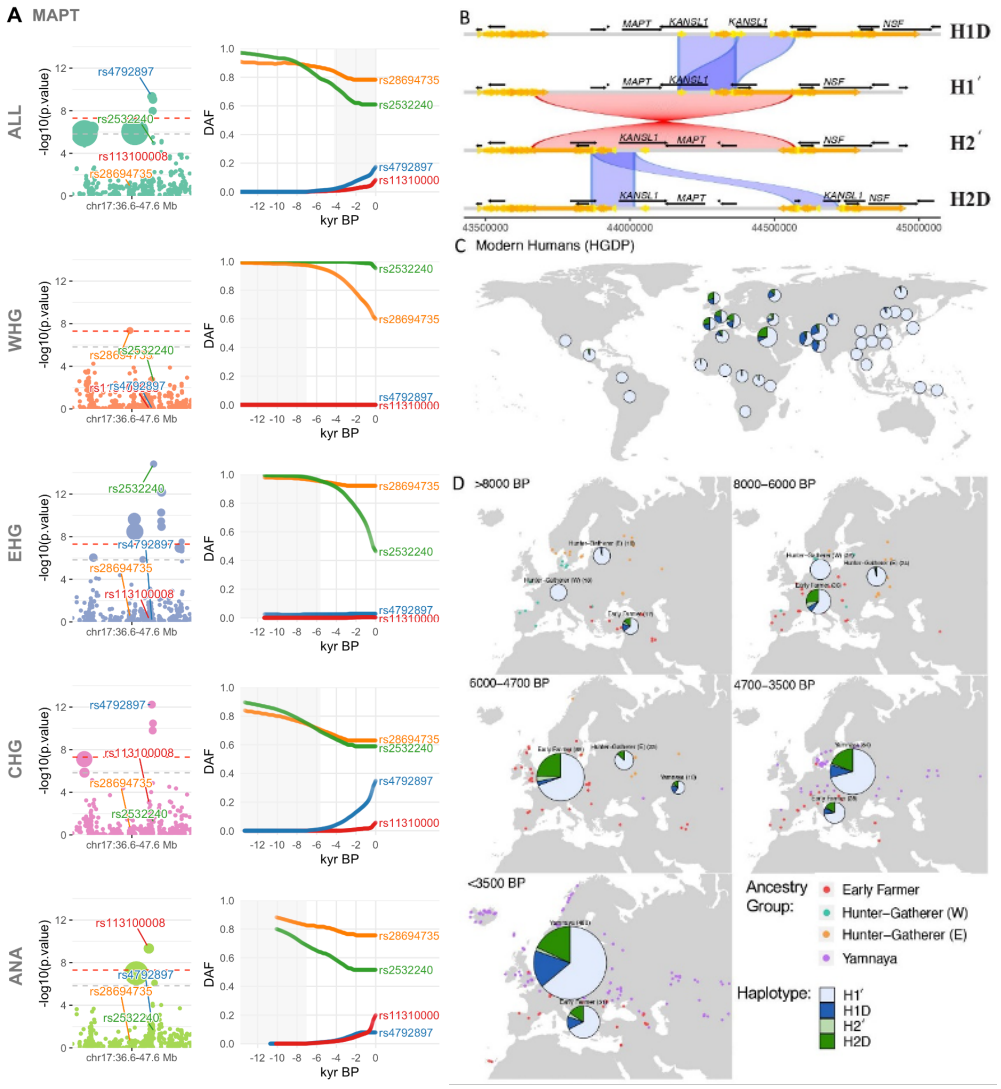


Fig 3. Selection at the *MAPT* / 17q21.31 inversion locus. A) Results for the pan-ancestry analysis (ALL) plus the four marginal ancestries: Western hunter-gatherers (WHG), Eastern hunter-gatherers (EHG), Caucasus hunter-gatherers (CHG) and Anatolian farmers (ANA). Grey shading for the marginal ancestries indicates approximate temporal extent of the pre-admixture population. B) Haplotypes of the 17q21.31 locus: the ancestral (non-inverted) H1 17q21.31 and the inverted H2 haplotype. Duplications of the *KANSL1* gene have occurred independently on both lineages yielding H1D and H2D haplotypes. C) Frequency of the 17q21.31 inversion and duplication haplotypes across modern-day global populations (Human Genome Diversity Project [62]). D) Change in the frequency of the 17q21.31 inversion haplotype through time.

<u>Selection on pigmentation-associated variants</u>

Our results identify strong selection for lighter skin pigmentation in groups moving northwards and westwards, in agreement with the hypothesis that selection is caused by reduced UV exposure and resulting vitamin D deficiency. We find that the most strongly selected alleles reached near-fixation several thousand years ago, suggesting that this was not associated with recent sexual selection as proposed [63,64] (Supplementary Note 2a).

In the pan-ancestry analysis we detect strong selection at the *SLC45A2* locus (rs35395; p=4.13e-23; s=0.022) locus [9,65], with the selected allele (responsible for lighter skin), increasing in frequency from c. 13,000 years ago, until plateauing c. 2,000 years ago (Fig. 2). The dominating hypothesis is

that high melanin levels in the skin are important in equatorial regions owing to its protection against UV radiation, whereas lighter skin has been selected for at higher latitudes (where UV radiation is less intense) because some UV penetration is required for cutaneous synthesis of vitamin D [66,67]. Our findings confirm pigmentation alleles as major targets of selection during the Holocene [8,68,69] particularly on a small proportion of loci with large effect sizes [9].

Additionally, our results provide unprecedentedly detailed information about the duration and geographic spread of these processes (Fig. 2) suggesting that an allele associated with lighter skin was selected for repeatedly, probably as a consequence of similar environmental pressures occurring at different times in different regions. In the ancestry-stratified analysis, all marginal ancestries show broad agreement at the *SLC45A2* locus (Fig. 2) but differ in the timing of their frequency shifts. The ANA ancestry background shows the earliest evidence for selection, followed by EHG and WHG around c. 10,000 years ago, and CHG c. 2,000 years later. In all ancestry backgrounds except WHG, the selected haplotypes reach near fixation by c. 3,000 years ago, whilst the WHG haplotype background contains the majority of ancestral alleles still segregating in present-day Europeans. This finding suggests that selection on this allele was much weaker in ancient western hunter-gatherer groups during the Holocene compared to elsewhere. We also detect strong selection at the *SLC24A5* locus (rs1426654; p=6.45e-09; s=0.019) which is also associated with skin pigmentation [65,70]. At this locus, the selected allele increased in frequency even earlier than *SLC45A2* and reached near fixation c. 3,500 years ago (Supplementary Note 2a). Selection on this locus thus seems to have occurred early on in groups that were moving northwards and westwards, and only later in the Western hunter-gatherer background after these groups encountered and admixed with the incoming populations.

Selection among major axes of ancient population variation

Beyond patterns of genetic change at the Mesolithic-Neolithic transition, much genetic variability observed today reflects high genetic differentiation in the hunter-gatherer groups that eventually contributed to modern European genetic diversity [34]. Indeed, a substantial number of loci associated with cardiovascular disease, metabolism and lifestyle diseases trace their genetic variability prior to the Neolithic transition, to ancient differential selection in ancestry groups occupying different parts of the Eurasian continent (Supplementary Note 2d). These may represent selection episodes that preceded the admixture events described above, and led to differentiation between ancient hunter-gatherer groups in the late Pleistocene and early Holocene. One of these overlaps with the *SLC24A3* gene which is a salt sensitivity gene significantly expressed in obese individuals [71,72]. Another spans *ROPN1* and *KALRN*, two genes involved in vascular disorders [73–75]. A further region contains *SLC35F3*, which codes for a thiamine transport and has been associated with hypertension in a Han Chinese cohort [76,77]. Finally, there is a candidate region containing several genes (*CH25H, FAS*) associated with obesity and lipid metabolism [78–80] and another peak with several genes (*ASXL2, RAB10, HADHA, GPR113*) involved in glucose homeostasis and fatty acid metabolism [81–90]. These loci predominantly reflect ancient patterns of extreme differentiation between Eastern and Western Eurasian genomes, and may be candidates for selection after the separation of the Pleistocene populations that occupied different environments across the continent (roughly 45,000 years ago [91]).

Pathogenic structural variants in ancient vs. modern-day humans

Rare, recurrent copy-number variants (CNVs) are known to cause neurodevelopmental disorders and are associated with a range of psychiatric and physical traits with variable expressivity and incomplete penetrance [92,93]. To understand the prevalence of pathogenic structural variants over time we examined 50 genomic regions susceptible to recurrent CNV, known to be the most

prevalent drivers of human developmental pathologies [94]. The analysis included 1442 ancient shotgun genomes passing quality control for CNV analysis (Supplementary Note 2h) and 1093 modern human genomes for comparison [62,95]. We identified CNVs in ancient individuals at ten loci using a read-depth based approach and digital Comparative Genomic Hybridization [96] (Supplementary Table S2h.1; Supplementary Figs. S2h.1 to S2h.20). Although most of the observed CNVs (including duplications at 15q11.2 and *CHRNA7*, and CNVs spanning parts of the TAR locus and 22q11.2 distal) have not been unambiguously associated with disease in large studies, the identified CNVs include deletions and duplications that have been associated with developmental delay, dysmorphic features, and neuropsychiatric abnormalities such as autism (most notably at 1q21.1, 3q29, 16p12.1 and the DiGeorge/VCFS locus, but also deletions at 15q11.2 and duplications at 16p13.11). An individual harbouring the 16p13.11 deletion, RISE586 [17], a 4,000 BP woman aged 20-30 from the Únětice culture (modern day Czech Republic), had almost complete skeletal remains, which allowed us to test for the presence of various skeletal abnormalities associated with the 16p13.11 microdeletion [97]. RISE586 exhibited a hypoplastic tooth, spondylolysis of the L5 vertebrae, incomplete coalescence of the S1 sacral bone, among other minor skeletal phenotypes. The skeletal phenotypes observed in this individual are relatively common (~10%) in European populations and are not specific to 16p13.1 thus do not indicate strong penetrance of this mutation in RISE586 [98–101]. However, these results do highlight our ability to link putatively pathogenic genotypes to phenotypes in ancient individuals. Overall, the carrier frequency in the ancient individuals is similar to that reported in the UK Biobank genomes (1.25% vs 1.6% at 15q11.2 and *CHRNA7* combined, and 0.8% vs 1.1% across the remaining loci combined) [102]. These results suggest that large, recurrent CNVs that can lead to several pathologies were present at similar frequencies in the ancient and modern populations included in this study.

Genetic trait reconstruction and the phenotypic legacy of ancient Europeans

When comparing modern European genomes in the UK Biobank to ancient Europeans, we find strong differentiation at certain sets of trait-associated variants, and differential contribution of different ancestry groups to various traits. We reconstructed polygenic scores for phenotypes in ancient individuals, using effect size estimates obtained from GWASs performed using the >400,000 UK Biobank genomes [5] (http://www.nealelab.is/uk-biobank) and looked for overdispersion among these scores across ancient populations, beyond what would be expected under a null model of genetic drift [103] (Supplementary Note 2c). We stress that polygenic scores and the Q_X statistic may both be affected by population stratification, so these results should be interpreted with caution [104–107]. The most significantly overdispersed scores are for variants associated with pigmentation, anthropometric differences and disorders related to diet and sugar levels, including diabetes (Fig. 4). We also find psychological trait scores with evidence for overdispersion related to mood instability and irritability, with Western Hunter-gatherers generally showing lower genetic scores for these traits than Neolithic Farmers. Intriguingly, we find highly inconsistent predictions of height based on polygenic scores in western hunter-gatherer and Siberian groups computed using effect sizes estimated from two different - yet largely overlapping - GWAS cohorts (Supplementary Note 2c), highlighting how sensitive polygenic score predictions are to the choice of cohort, particularly when ancient populations are genetically divergent from the reference GWAS cohort [107]. Taking this into account, we do observe that the Eastern hunter-gatherer and individuals associated with the Yamnaya culture have consistently high genetic values for height, which in turn contribute to stature increases in Bronze Age Europe, relative to the earlier Neolithic populations [8,108,109].

We performed an additional analysis to examine the data for strong alignments between axes of trait-association [110] and ancestry gradients, rather than relying on particular choices for population clusters (Supplementary Note 2e). Along the population structure axis separating ancient East Asian and Siberian genomes from Steppe and Western European genomes (Fig. 2), we find significant correlations with trait-association components related to impedance, body measurements, blood measurements, eye measurement and skin disorders. Along the axis separating Mesolithic hunter-gatherers from Anatolian and Neolithic farmer individuals, we find significant correlations with trait-association components related to skin disorders, diet and lifestyle traits, mental health status, and spirometry-related traits (Fig. 4). Our findings show that these phenotypes were genetically different among ancient groups with very different lifestyles. However, we note that the realised value of these traits is highly dependent on environmental factors and gene-environment interactions, which we do not model in this analysis.

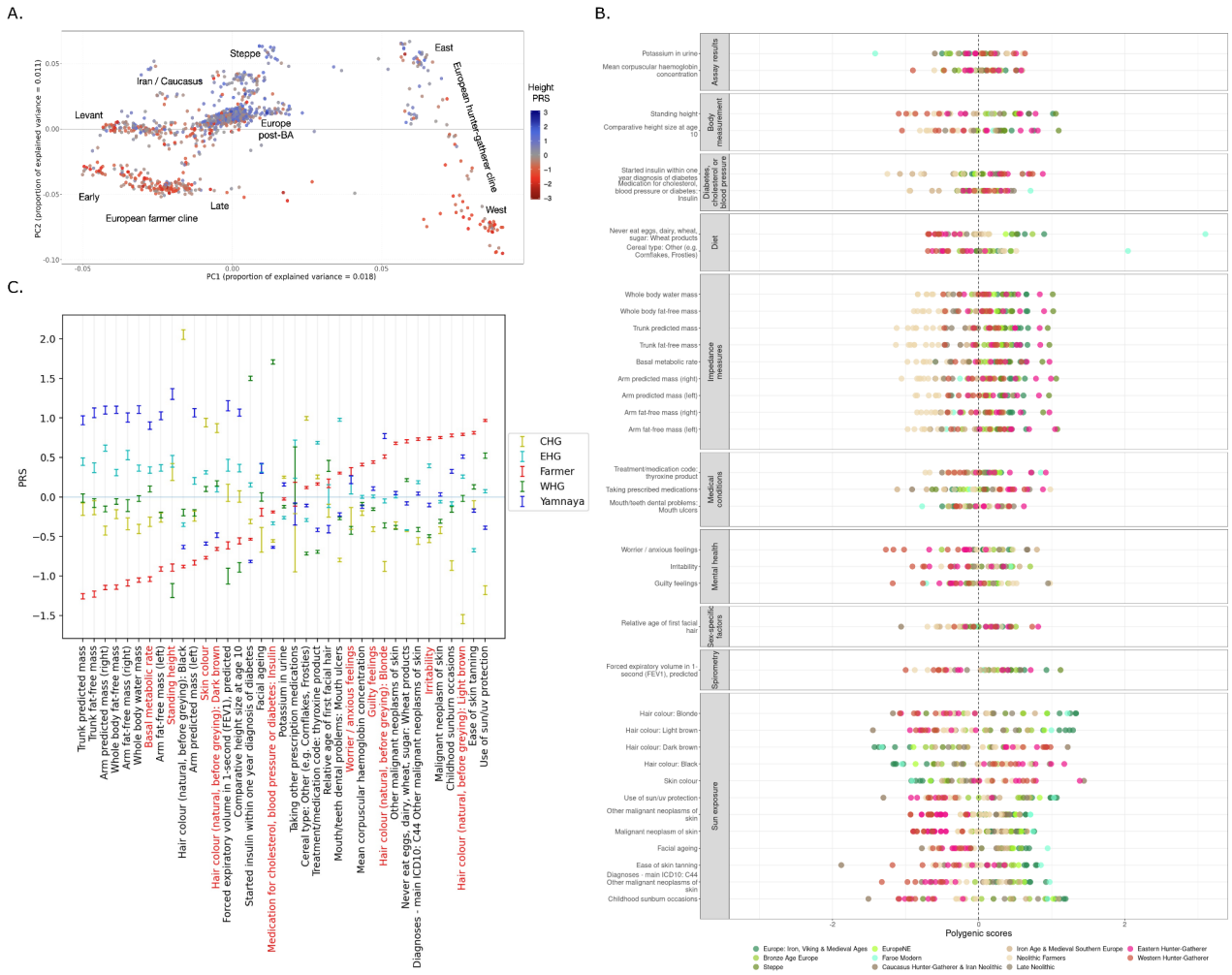


Fig. 4 A. First two principal components of a PCA of ancient Western Eurasian samples, coloured by height polygenic scores. B. Top UK Biobank traits with highest overdispersion in polygenic scores among ancient populations, out of a total of 320 sets of trait-associated SNPs tested ($Q_X > 79.4$, $P < 0.05/320$). C Ancestry-specific polygenic scores based on chromosome painting of the UK Biobank, for traits significantly over-dispersed in ancient populations. Confidence intervals are estimated by bootstrapping modern samples, and traits mentioned in the main text are highlighted in red.

In addition to the above reconstructions of genetic traits among the ancient individuals, we also estimated the contribution from different ancestral populations (EHG, CHG, WHG, Yamnaya and Anatolian farmer) to variation in polygenic phenotypes in present-day individuals, leveraging the exceptional resolution offered by the UK Biobank genomes [5] to investigate this. We calculated ancestry-specific polygenic risk scores based on chromosome painting of the >400,000 UKB genomes, using ChromoPainter [111] (Fig. 4C, Supplementary Note 2g). This allowed us to identify if any of the ancient ancestry components were over-represented in modern UK populations at loci significantly associated with a given trait, and also avoids exporting risk scores over space and time. Working with large numbers of imputed ancient genomes provides high statistical power to use ancient populations as "ancestral sources". We focused on phenotypes whose polygenic scores were significantly over-dispersed in the ancient populations (Supplementary Note 2c), as well as a single high effect variant, ApoE4, known to be a significant risk factor in Alzheimer's Disease ([112,113]). We emphasise that this approach makes no reference to ancient phenotypes but describes how these ancestries contributed to the modern genetic landscape. In light of the ancestry gradients within the British Isles and Eurasia [11], these results support the hypothesis that ancestry-mediated geographic variation in disease risks and phenotypes is commonplace. It points to a way forward for disentangling how ancestry contributed to differences in risk of genetic disease – including metabolic and mental health disorders – between present-day populations.

Taken together, these analyses help to settle the famous discussion of selection in Europe relating to height [8,109,114]. The finding that steppe individuals have consistently high genetic values for height (Supplementary Note 2c), is mirrored by the UK Biobank results, which find that the 'Steppe' ancestral components (Yamnaya/EHG) contributed to increased height in present-day populations (Supplementary Note 2g). This shows that the height differences in Europe between north and south may not be due to selection in Europe, as claimed in many previous studies, but may be a consequence of differential ancestry.

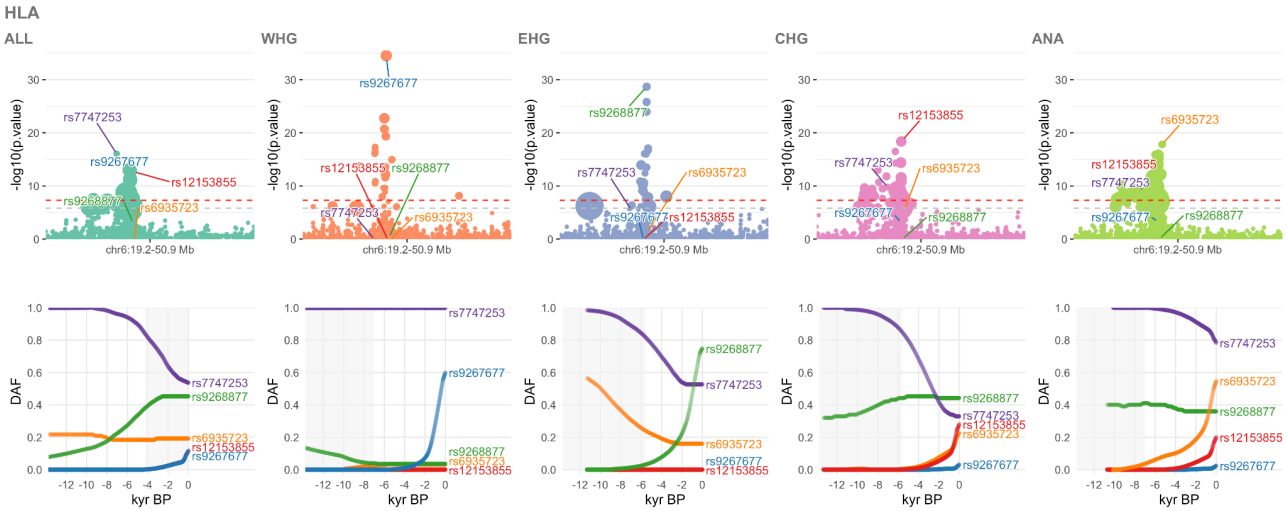
Likewise, European hunter gatherers are genetically predicted to have dark skin pigmentation and dark brown hair [9,10,17,18,115–118], and indeed we see that the WHG, EHG and CHG components contributed to these phenotypes in present-day individuals whereas the Yamnaya and Anatolian farmer ancestry contributed to light brown/blonde hair pigmentation (Supplementary Note 2g). Interestingly, loci associated with overdispersed mood-related polygenic phenotypes recorded among the UK Biobank individuals (like increased anxiety, guilty feelings, and irritability) showed an overrepresentation of the Anatolian farmer ancestry component; and the WHG component showed a strikingly high contribution to traits related to diabetes. We also found that the ApoE4 effect allele (increased risk for Alzheimer's disease) is preferentially found on a WHG/EHG haplotypic background, suggesting it likely was brought to western Europe by early hunter-gatherers (Supplementary Note 2g). This is in line with the present-day European distribution of this allele, which is highest in north-eastern Europe, where the proportion of these ancestries are higher than in other regions of the continent [119].

Conclusions

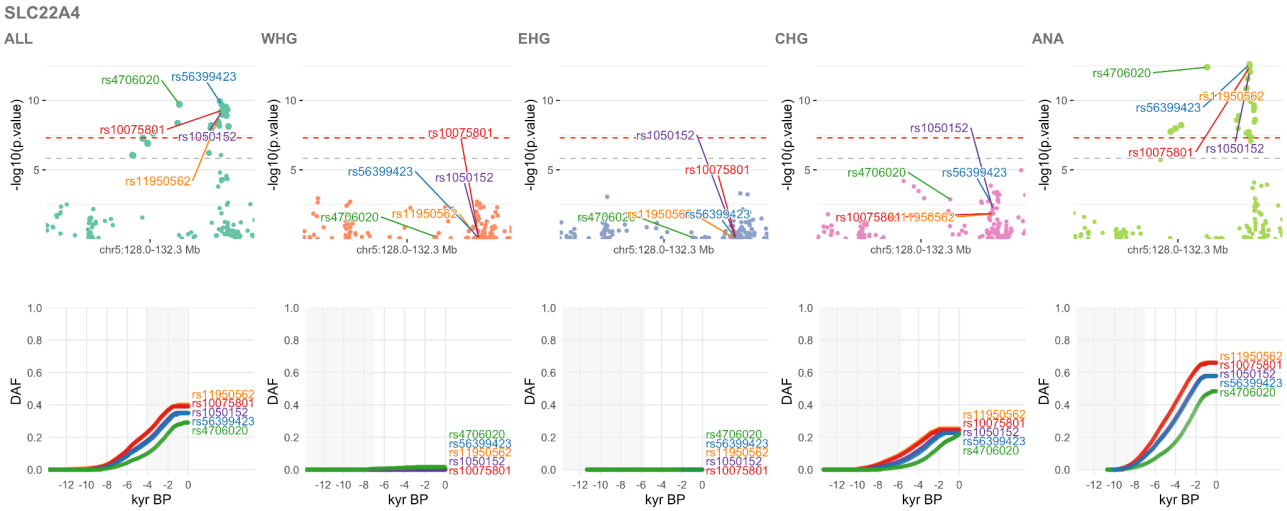
The transition from hunting and gathering, to farming, and subsequently pastoralism, precipitated far-reaching consequences for the diet, and physical and mental health of Eurasian populations. These dramatic cultural changes created a heterogeneous mix of selection pressures. Our analyses revealed that the ability to detect signatures of natural selection in modern human genomes is drastically limited by conflicting selection pressures in different ancestral populations masking the

signals. Developing methods to trace selection in individual ancestry components allowed us to effectively double the number of significant selection peaks, which helped clarify the trajectories of a number of traits related to diet and lifestyle. Furthermore, numerous complex traits thought to have been under local selection are better explained by differing proportions of ancient ancestry in present-day populations. Overall, our results emphasise how the interplay between ancient selection and major admixture events occurring across Europe and Asia in the Stone and Bronze Ages have profoundly shaped the patterns of genetic variation observed in present-day human populations.

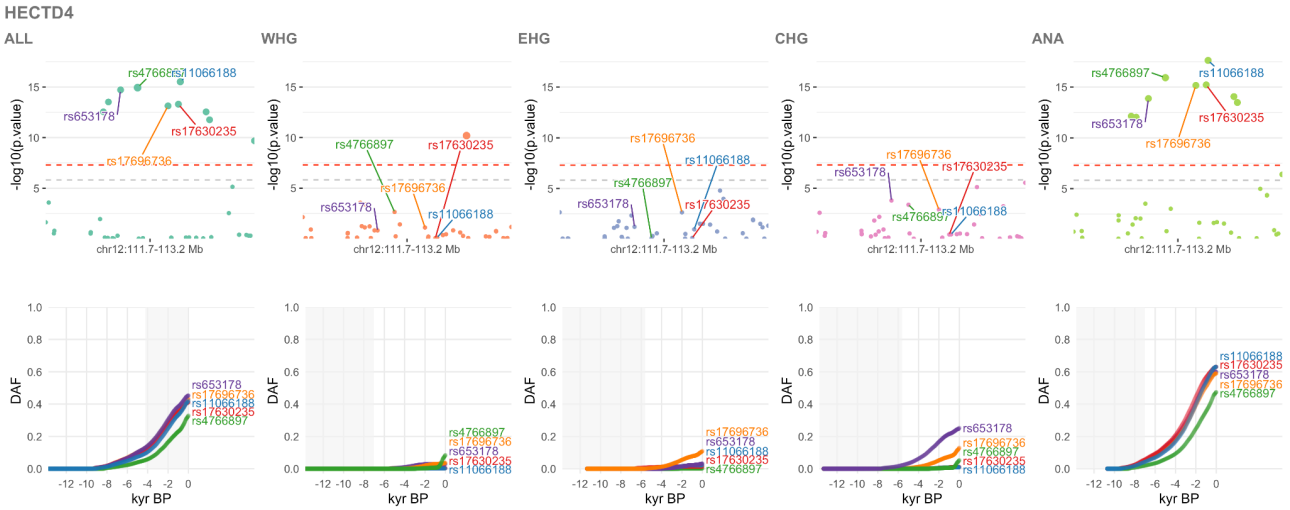
Extended Data Figures



Extended Data Fig. 1. Selection at the HLA locus. Results for the pan-ancestry analysis (ALL) plus the four marginal ancestries: Western hunter-gatherers (WHG), Eastern hunter-gatherers (EHG), Caucasus hunter-gatherers (CHG) and Anatolian farmers (ANA). Row one shows zoomed Manhattan plots of the p-values for each ancestry (significant SNPs sized by their selection coefficients), and row two shows allele trajectories for the top SNPs across all ancestries (grey shading for the marginal ancestries indicates approximate temporal extent of the pre-admixture population).



Extended Data Fig. 2. Selective sweep at the *SLC22A4* locus. Results for the pan-ancestry analysis (ALL) plus the four marginal ancestries: Western hunter-gatherers (WHG), Eastern hunter-gatherers (EHG), Caucasus hunter-gatherers (CHG) and Anatolian farmers (ANA). Row one shows zoomed Manhattan plots of the p-values for each ancestry (significant SNPs sized by their selection coefficients), and row two shows allele trajectories for the top SNPs across all ancestries (grey shading for the marginal ancestries indicates approximate temporal extent of the pre-admixture population).



Extended Data Fig. 3. Selective sweep at the HECTD4 locus. Results for the pan-ancestry analysis (ALL) plus the four marginal ancestries: Western hunter-gatherers (WHG), Eastern hunter-gatherers (EHG), Caucasus hunter-gatherers (CHG) and Anatolian farmers (ANA). Row one shows zoomed Manhattan plots of the p-values for each ancestry (significant SNPs sized by their selection coefficients), and row two shows allele trajectories for the top SNPs across all ancestries (grey shading for the marginal ancestries indicates approximate temporal extent of the pre-admixture population).

Data availability

All collapsed and paired-end sequence data for novel samples sequenced in this study will be made publicly available on the European Nucleotide Archive, together with trimmed sequence alignment map files, aligned using human build GRCh37. Previously published ancient genomic data used in this study are detailed in Supplementary Table VII of [11], and are all already publicly available.

Code availability

The modified version of CLUES used in this study is available from https://github.com/standard-aaron/clues. The pipeline and conda environment necessary to replicate the analysis of allele frequency trajectories of trait-associated variants in Supplementary Note 2a are available on Github at https://github.com/ekirving/mesoneo_paper. The pipeline to replicate the analyses for Supplementary Note 2c-2e can be found at https://github.com/albarema/neo. All other analyses relied upon available software which has been fully referenced in the manuscript and detailed in the relevant supplementary notes.

References

1. Page, A. E. *et al.* Reproductive trade-offs in extant hunter-gatherers suggest adaptive mechanism for the Neolithic expansion. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 4694–4699 (2016).
2. Marciniak, S., Bergey, C., Silva, A. M. & Hałuszko, A. An integrative skeletal and paleogenomic analysis of prehistoric stature variation suggests relatively reduced health for early European farmers. *bioRxiv* (2021).
3. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
4. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
5. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
6. Nielsen, R. Molecular signatures of natural selection. *Annu. Rev. Genet.* **39**, 197–218 (2005).
7. Vitti, J. J., Grossman, S. R. & Sabeti, P. C. Detecting natural selection in genomic data. *Annu. Rev. Genet.* **47**, 97–120 (2013).
8. Mathieson, I. *et al.* Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**, 499–503 (2015).
9. Ju, D. & Mathieson, I. The evolution of skin pigmentation-associated variation in West Eurasia. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
10. Wilde, S. *et al.* Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 4832–4837 (2014).
11. Allentoft, M. E., Sikora, M. & Refoyo-Martínez, A. Population Genomics of Stone Age Eurasia. *bioRxiv* (2022).
12. Jones, E. R. *et al.* Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nat. Commun.* **6**, 8912 (2015).
13. Speidel, L., Forest, M., Shi, S. & Myers, S. R. A method for genome-wide genealogy

estimation for thousands of samples. *Nat. Genet.* **51**, 1321–1329 (2019).
14. Speidel, L. *et al.* Inferring Population Histories for Ancient Genomes Using Genome-Wide Genealogies. *Mol. Biol. Evol.* **38**, 3497–3511 (2021).
15. Stern, A. J., Wilton, P. R. & Nielsen, R. An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. *PLoS Genet.* **15**, e1008384 (2019).
16. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
17. Allentoft, M. E. *et al.* Population genomics of Bronze Age Eurasia. *Nature* **522**, 167–172 (2015).
18. Haak, W. *et al.* Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207–211 (2015).
19. Enattah, N. S. *et al.* Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture. *Am. J. Hum. Genet.* **82**, 57–72 (2008).
20. Itan, Y., Powell, A., Beaumont, M. A., Burger, J. & Thomas, M. G. The origins of lactase persistence in Europe. *PLoS Comput. Biol.* **5**, e1000491 (2009).
21. Ségurel, L. & Bon, C. On the Evolution of Lactase Persistence in Humans. *Annu. Rev. Genomics Hum. Genet.* **18**, 297–319 (2017).
22. Segurel, L. *et al.* Why and when was lactase persistence selected for? Insights from Central Asian herders and ancient DNA. *PLoS Biol.* **18**, e3000742 (2020).
23. Enattah, N. S. *et al.* Identification of a variant associated with adult-type hypolactasia. *Nat. Genet.* **30**, 233–237 (2002).
24. Bersaglieri, T. *et al.* Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* **74**, 1111–1120 (2004).
25. Wang, L. *et al.* A MicroRNA Linking Human Positive Selection and Metabolic Disorders. *Cell* **183**, 684–701.e14 (2020).
26. Evershed, R. P. *et al.* Dairying, diseases and the evolution of lactase persistence in Europe. *Nature* **608**, 336–345 (2022).
27. Willer, C. J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).
28. Gallois, A. *et al.* A comprehensive study of metabolite genetics reveals strong pleiotropy and heterogeneity across time and context. *Nat. Commun.* **10**, 4788 (2019).
29. Ligthart, S. *et al.* Bivariate genome-wide association study identifies novel pleiotropic loci for lipids and inflammation. *BMC Genomics* **17**, 443 (2016).
30. Buckley, M. T. *et al.* Selection in Europeans on Fatty Acid Desaturases Associated with Dietary Changes. *Mol. Biol. Evol.* **34**, 1307–1318 (2017).
31. Ye, K., Gao, F., Wang, D., Bar-Yosef, O. & Keinan, A. Dietary adaptation of FADS genes in Europe varied across time and geography. *Nat Ecol Evol* **1**, 167 (2017).
32. Mathieson, S. & Mathieson, I. FADS1 and the Timing of Human Adaptation to Agriculture. *Mol. Biol. Evol.* **35**, 2957–2970 (2018).

33. Lazaridis, I. The evolutionary history of human populations in Europe. *Curr. Opin. Genet. Dev.* **53**, 21–27 (2018).
34. Luu, K., Bazin, E. & Blum, M. G. B. pcadapt: an R package to perform genome scans for selection based on principal component analysis. *Mol. Ecol. Resour.* **17**, 67–77 (2017).
35. Fairn, G. D. & McMaster, C. R. Emerging roles of the oxysterol-binding protein family in metabolism, transport, and signaling. *Cell. Mol. Life Sci.* **65**, 228–236 (2008).
36. Lehto, M. & Olkkonen, V. M. The OSBP-related proteins: a novel protein family involved in vesicle transport, cellular lipid metabolism, and cell signalling. *Biochim. Biophys. Acta* **1631**, 1–11 (2003).
37. Sánchez-Solana, B., Li, D.-Q. & Kumar, R. Cytosolic functions of MORC2 in lipogenesis and adipogenesis. *Biochim. Biophys. Acta* **1843**, 316–326 (2014).
38. Zhong, X. *et al.* The zinc-finger protein ZFYVE1 modulates TLR3-mediated signaling by facilitating TLR3 ligand binding. *Cell. Mol. Immunol.* **17**, 741–752 (2020).
39. Kim, S. V. *et al.* GPR15-mediated homing controls immune homeostasis in the large intestine mucosa. *Science* **340**, 1456–1459 (2013).
40. Nguyen, L. P. *et al.* Role and species-specific expression of colon T cell homing receptor GPR15 in colitis. *Nature Immunology* vol. 16 207–213 (2015).
41. Monteleone, G., Boirivant, M., Pallone, F. & MacDonald, T. T. TGF-β1 and Smad7 in the regulation of IBD. *Mucosal Immunology* vol. 1 S50–S53 (2008).
42. Kennedy, B. W. C. Mongersen, an Oral SMAD7 Antisense Oligonucleotide, and Crohn's Disease. *The New England journal of medicine* vol. 372 2461 (2015).
43. Garo, L. P. *et al.* Smad7 Controls Immunoregulatory PDL2/1-PD1 Signaling in Intestinal Inflammation and Autoimmunity. *Cell Rep.* **28**, 3353–3366.e5 (2019).
44. Morris, J. A. *et al.* An atlas of genetic influences on osteoporosis in humans and mice. *Nat. Genet.* **51**, 258–266 (2018).
45. Global Biobank Engine. http://gbe.stanford.edu.
46. McInnes, G. *et al.* Global Biobank Engine: enabling genotype-phenotype browsing for biobank summary statistics. *Bioinformatics* **35**, 2495–2497 (2019).
47. Brinkworth, J. F. & Barreiro, L. B. The contribution of natural selection to present-day susceptibility to chronic inflammatory and autoimmune disease. *Curr. Opin. Immunol.* **31**, 66–78 (2014).
48. Fumagalli, M. *et al.* Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genet.* **7**, e1002355 (2011).
49. Jones, A. V. *et al.* GWAS of self-reported mosquito bite size, itch intensity and attractiveness to mosquitoes implicates immune-related predisposition loci. *Hum. Mol. Genet.* **26**, 1391–1406 (2017).
50. Gutierrez-Achury, J. *et al.* Functional implications of disease-specific variants in loci jointly associated with coeliac disease and rheumatoid arthritis. *Hum. Mol. Genet.* **25**, 180–190 (2016).
51. Stefansson, H. *et al.* A common inversion under selection in Europeans. *Nat. Genet.* **37**, 129–137 (2005).
52. Steinberg, K. M. *et al.* Structural diversity and African origin of the 17q21.31 inversion

polymorphism. *Nat. Genet.* **44**, 872–880 (2012).
53. Kılınç, G. M. *et al.* The Demographic Development of the First Farmers in Anatolia. *Curr. Biol.* **26**, 2659–2666 (2016).
54. Broushaki, F. *et al.* Early Neolithic genomes from the eastern Fertile Crescent. *Science* **353**, 499–503 (2016).
55. Jones, E. R. *et al.* The Neolithic Transition in the Baltic Was Not Driven by Admixture with Early European Farmers. *Curr. Biol.* **27**, 576–582 (2017).
56. Andreadis, A., Brown, W. M. & Kosik, K. S. Structure and novel exons of the human tau gene. *Biochemistry* **31**, 10626–10633 (1992).
57. Alonso, A. C., Grundke-Iqbal, I. & Iqbal, K. Alzheimer's disease hyperphosphorylated tau sequesters normal tau into tangles of filaments and disassembles microtubules. *Nat. Med.* **2**, 783–787 (1996).
58. Desikan, R. S. *et al.* Genetic overlap between Alzheimer's disease and Parkinson's disease at the MAPT locus. *Mol. Psychiatry* **20**, 1588–1595 (2015).
59. Satake, W. *et al.* Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease. *Nat. Genet.* **41**, 1303–1307 (2009).
60. Simón-Sánchez, J. *et al.* Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nat. Genet.* **41**, 1308–1312 (2009).
61. Shulman, J. M. & De Jager, P. L. Evidence for a common pathway linking neurodegenerative diseases. *Nature genetics* vol. 41 1261–1262 (2009).
62. Bergström, A. *et al.* Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**, (2020).
63. Aoki, K. Sexual selection as a cause of human skin colour variation: Darwin's hypothesis revisited. *Ann. Hum. Biol.* **29**, 589–608 (2002).
64. Frost, P. The puzzle of European hair, eye, and skin color. *Advances in Anthropology* **2014**, (2014).
65. Lona-Durazo, F. *et al.* Meta-analysis of GWA studies provides new insights on the genetic architecture of skin pigmentation in recently admixed populations. *BMC Genet.* **20**, 59 (2019).
66. Jablonski, N. G. & Chaplin, G. The evolution of human skin coloration. *J. Hum. Evol.* **39**, 57–106 (2000).
67. Engelsen, O. The relationship between ultraviolet radiation exposure and vitamin D status. *Nutrients* **2**, 482–495 (2010).
68. 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).
69. Sabeti, P. C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918 (2007).
70. Martin, A. R. *et al.* An Unexpectedly Complex Architecture for Skin Pigmentation in Africans. *Cell* **171**, 1340–1353.e14 (2017).
71. Wu, H. *et al.* Transcriptome Sequencing to Detect the Potential Role of Long Noncoding RNAs in Salt-Sensitive Hypertensive Rats. *Biomed Res. Int.* **2019**, 2816959 (2019).
72. Logsdon, B. A., Hoffman, G. E. & Mezey, J. G. Mouse obesity network reconstruction with a variational Bayes algorithm to employ aggressive false positive control. *BMC Bioinformatics*

13, 53 (2012).
73. Wang, L. *et al.* Peakwide Mapping on Chromosome 3q13 Identifies the Kalirin Gene as a Novel Candidate Gene for Coronary Artery Disease. *The American Journal of Human Genetics* vol. 80 650–663 (2007).
74. Ikram, M. A., Seshadri, S. & Bis, J. C. Genomewide Association Studies of Stroke. *Journal of Vascular Surgery* vol. 50 467 (2009).
75. Krug, T. *et al.* Kalirin: a novel genetic risk factor for ischemic stroke. *Hum. Genet.* **127**, 513–523 (2010).
76. Zang, X.-L. *et al.* Association of a SNP in SLC35F3 Gene with the Risk of Hypertension in a Chinese Han Population. *Frontiers in Genetics* vol. 7 (2016).
77. Zhang, K. *et al.* Genetic implication of a novel thiamine transporter in human hypertension. *J. Am. Coll. Cardiol.* **63**, 1542–1555 (2014).
78. Russo, L. *et al.* Cholesterol 25-hydroxylase (CH25H) as a promoter of adipose tissue inflammation in obesity and diabetes. *Mol Metab* **39**, 100983 (2020).
79. Zhao, J., Chen, J., Li, M., Chen, M. & Sun, C. Multifaceted Functions of CH25H and 25HC to Modulate the Lipid Metabolism, Immune Responses, and Broadly Antiviral Activities. *Viruses* **12**, (2020).
80. Demir, A., Kahraman, R., Candan, G. & Ergen, A. The role of FAS gene variants in inflammatory bowel disease. *Turk. J. Gastroenterol.* **31**, 356–361 (2020).
81. Izawa, T. *et al.* ASXL2 Regulates Glucose, Lipid, and Skeletal Homeostasis. *Cell Rep.* **11**, 1625–1637 (2015).
82. Zou, W. *et al.* Myeloid-specific Asxl2 deletion limits diet-induced obesity by regulating energy expenditure. *J. Clin. Invest.* **130**, 2644–2656 (2020).
83. Park, U.-H., Yoon, S. K., Park, T., Kim, E.-J. & Um, S.-J. Additional Sex Comb-like (ASXL) Proteins 1 and 2 Play Opposite Roles in Adipogenesis via Reciprocal Regulation of Peroxisome Proliferator-activated Receptor γ. *Journal of Biological Chemistry* vol. 286 1354–1363 (2011).
84. Ponsuksili, S. *et al.* Epigenome-wide skeletal muscle DNA methylation profiles at the background of distinct metabolic types and ryanodine receptor variation in pigs. *BMC Genomics* **20**, 492 (2019).
85. Samad, M. B. *et al.* [6]-Gingerol, from Zingiber officinale, potentiates GLP-1 mediated glucose-stimulated insulin secretion pathway in pancreatic β-cells and increases RAB8/RAB10-regulated membrane presentation of GLUT4 transporters in skeletal muscle to improve hyperglycemia in Leprdb/db type 2 diabetic mice. *BMC Complementary and Alternative Medicine* vol. 17 (2017).
86. Vazirani, R. P. *et al.* Disruption of Adipose Rab10-Dependent Insulin Signaling Causes Hepatic Insulin Resistance. *Diabetes* **65**, 1577–1589 (2016).
87. Hsieh, P. *et al.* Exome Sequencing Provides Evidence of Polygenic Adaptation to a Fat-Rich Animal Diet in Indigenous Siberian Populations. *Mol. Biol. Evol.* **34**, 2913–2926 (2017).
88. Thapa, D. *et al.* The protein acetylase GCN5L1 modulates hepatic fatty acid oxidation activity via acetylation of the mitochondrial β-oxidation enzyme HADHA. *J. Biol. Chem.* **293**, 17676–17684 (2018).

89. Baloni, P. *et al.* Genome-scale metabolic model of the rat liver predicts effects of diet restriction. *Sci. Rep.* **9**, 9807 (2019).
90. Ong, H. S. & Yim, H. C. H. Microbial Factors in Inflammatory Diseases and Cancers. *Regulation of Inflammatory Signaling in Health and Disease* 153–174 (2017) doi:10.1007/978-981-10-5987-2_7.
91. Sikora, M. *et al.* The population history of northeastern Siberia since the Pleistocene. *Nature* **570**, 182–188 (2019).
92. Girirajan, S., Campbell, C. D. & Eichler, E. E. Human copy number variation and complex genetic disease. *Annu. Rev. Genet.* **45**, 203–226 (2011).
93. Weise, A. *et al.* Microdeletion and microduplication syndromes. *J. Histochem. Cytochem.* **60**, 346–358 (2012).
94. Girirajan, S. *et al.* Phenotypic heterogeneity of genomic disorders and rare copy-number variants. *N. Engl. J. Med.* **367**, 1321–1331 (2012).
95. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
96. Sudmant, P. H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330**, 641–646 (2010).
97. Nagamani, S. C. S. *et al.* Phenotypic manifestations of copy number variation in chromosome 16p13.11. *Eur. J. Hum. Genet.* **19**, 280–286 (2011).
98. Cremin, B., Goodman, H., Spranger, J. & Beighton, P. Wormian bones in osteogenesis imperfecta and other disorders. *Skeletal Radiol.* **8**, 35–38 (1982).
99. Foreman, P. *et al.* L5 spondylolysis/spondylolisthesis: a comprehensive review with an anatomic focus. *Childs. Nerv. Syst.* **29**, 209–216 (2012).
100. Vleeming, A. *et al.* The sacroiliac joint: an overview of its anatomy, function and potential clinical implications. *J. Anat.* **221**, 537–567 (2012).
101. Salanitri, S. & Seow, W. K. Developmental enamel defects in the primary dentition: aetiology and clinical management. *Aust. Dent. J.* **58**, 133–40; quiz 266 (2013).
102. Crawford, K. *et al.* Medical consequences of pathogenic CNVs in adults: analysis of the UK Biobank. *J. Med. Genet.* **56**, 131–138 (2019).
103. Berg, J. J. & Coop, G. A population genetic signal of polygenic adaptation. *PLoS Genet.* **10**, e1004412 (2014).
104. Berg, J. J. *et al.* Reduced signal for polygenic adaptation of height in UK Biobank. *Elife* **8**, (2019).
105. Sohail, M. *et al.* Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *Elife* **8**, (2019).
106. Irving-Pease, E. K., Muktupavela, R., Dannemann, M. & Racimo, F. Quantitative Human Paleogenetics: What can Ancient DNA Tell us About Complex Trait Evolution? *Front. Genet.* **12**, 703541 (2021).
107. Refoyo-Martínez, A. *et al.* How robust are cross-population signatures of polygenic adaptation in humans? *bioRxiv* 2020.07.13.200030 (2021) doi:10.1101/2020.07.13.200030.
108. Ruff, C. B. *Skeletal Variation and Adaptation in Europeans: Upper Paleolithic to the Twentieth Century*. (John Wiley & Sons, 2017).

109. Cox, S. L., Ruff, C. B., Maier, R. M. & Mathieson, I. Genetic contributions to variation in human stature in prehistoric Europe. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 21484–21492 (2019).
110. Tanigawa, Y. *et al.* Components of genetic associations across 2,138 phenotypes in the UK Biobank highlight adipocyte biology. *Nat. Commun.* **10**, 4064 (2019).
111. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet.* **8**, e1002453 (2012).
112. Corder, E. H. *et al.* Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261**, 921–923 (1993).
113. Strittmatter, W. J. *et al.* Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proc. Natl. Acad. Sci. U. S. A.* **90**, 1977–1981 (1993).
114. Rosenstock, E. *et al.* Human stature in the Near East and Europe ca. 10,000–1000 BC: its spatiotemporal development in a Bayesian errors-in-variables model. *Archaeol. Anthropol. Sci.* **11**, 5657–5690 (2019).
115. Olalde, I. *et al.* Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. *Nature* **507**, 225–228 (2014).
116. González-Fortes, G. *et al.* Paleogenomic Evidence for Multi-generational Mixing between Neolithic Farmers and Mesolithic Hunter-Gatherers in the Lower Danube Basin. *Curr. Biol.* **27**, 1801–1810.e10 (2017).
117. Brace, S. *et al.* Ancient genomes indicate population replacement in Early Neolithic Britain. *Nat Ecol Evol* **3**, 765–771 (2019).
118. Jensen, T. Z. T. *et al.* A 5700 year-old human genome and oral microbiome from chewed birch pitch. *Nat. Commun.* **10**, 5520 (2019).
119. Belloy, M. E., Napolioni, V. & Greicius, M. D. A Quarter Century of APOE and Alzheimer's Disease: Progress to Date and the Path Forward. *Neuron* **101**, 820–838 (2019).

Acknowledgements

We thank all the former and current staff at the Lundbeck Foundation GeoGenetics Centre and the GeoGenetics Sequencing Core, and to colleagues across the many institutions detailed below. We are particularly grateful to Line Olsen as project manager for the Lundbeck Foundation GeoGenetics Centre project. We thank UK Biobank Ltd. for access to the UK Biobank genomic resource. We are thankful to Illumina Inc. for collaboration. EW thanks St. John's College, Cambridge, for providing a stimulating environment of discussion and learning.

The Lundbeck Foundation GeoGenetics Centre is supported by the the Lundbeck Foundation (R302-2018-2155, R155-2013-16338), the Novo Nordisk Foundation (NNF18SA0035006), the Wellcome Trust (UNS69906), Carlsberg Foundation (CF18-0024), the Danish National Research Foundation (44113220) and the University of Copenhagen (KU2016 programme). This research has been conducted using the UK Biobank Resource and the iPSYCH Initiative, funded by the Lundbeck Foundation (R102-A9118 and R155-2014-1724).

Author Information

These authors contributed equally: Evan K. Irving-Pease, Alba Refoyo-Martínez, Andrés Ingason, Alice Pearson, Anders Fischer and William Barrie

These authors equally supervised research: Peter H. Sudmant, Daniel J. Lawson, Richard Durbin, Thorfinn Korneliussen, Thomas Werge, Morten E. Allentoft, Martin Sikora1, Rasmus Nielsen, Fernando Racimo, Eske Willerslev

Affiliations

Lundbeck Foundation GeoGenetics Centre, Globe Institute, University of Copenhagen, Copenhagen, Denmark
Evan K. Irving-Pease, Alba Refoyo-Martínez, Andrés Ingason, Fabrice Demeter, Rasmus A. Henriksen, Tharsika Vimala, Hugh McColl, Gabriele Scorrano, Abigail Ramsøe, Anders Rosengren, Lei Zhao, Kristian Kristiansen, Thorfinn Korneliussen, Morten E. Allentoft, Martin Sikora, Rasmus Nielsen, Fernando Racimo, and Eske Willerslev

Institute of Biological Psychiatry, Mental Health Services, Copenhagen University Hospital, Roskilde, Denmark
Andrés Ingason, Andrew J. Schork, and Anders Rosengren

Department of Genetics, University of Cambridge, UK
Alice Pearson and Richard Durbin

Department of Zoology, University of Cambridge, UK
Alice Pearson

Department of Historical Studies, University of Gothenburg, 405 30 Gothenburg, Sweden
Anders Fischer

Sealand Archaeology, Gl. Roesnaesvej 27, 4400 Kalundborg, Denmark
Anders Fischer

GeoGenetics Group, Department of Zoology, University of Cambridge, UK
William Barrie, Ruairidh Macleod, and Eske Willerslev

Department of Historical Studies, Gothenburg University, Sweden
Karl-Göran Sjögren

Department of Integrative Biology, University of California Berkeley
Alma S. Halgren and Peter H. Sudmant

Research department of Genetics, Evolution and Environment, University College London
Ruairidh Macleod

National Museum of Natural History, Paris, France
Fabrice Demeter

Center for Computational Biology, University of California, Berkeley, USA

Andrew Vaughn, Aaron J. Stern, and Peter H. Sudmant

Ancient Genomics Laboratory, The Francis Crick Institute, London, UK
Leo Speidel

Genetics Institute, University College London, London, UK
Leo Speidel

Neurogenomics Division, The Translational Genomics Research Institute (TGEN), Phoenix, AZ, USA
Andrew J. Schork

Department of Historical Studies, University of Gothenburg, SE-41255, Gothenburg, Sweden
Kristian Kristiansen

Institute of Statistical Sciences, School of Mathematics, University of Bristol, Bristol, UK
Daniel J. Lawson

Wellcome Sanger Institute, Cambridge, UK
Richard Durbin

Department of Clinical Medicine and Lundbeck Center for Geogenetics, GLOBE Institute, University of Copenhagen
Thomas Werge

Institute of Biological Psychiatry, Mental Health Center Sct Hans, Copenhagen University Hospital, Denmark
Thomas Werge

Trace and Environmental DNA (TrEnD) Laboratory, School of Molecular and Life Science, Curtin University, Australia
Morten E. Allentoft

Departments of Integrative Biology and Statistics, UC Berkeley, Berkeley 94720, USA
Rasmus Nielsen

Contributions

E.K.I-P, A.R-M, A.I., A.P.. A.F., and W.B. contributed equally to this work. P.H.S., D.J.L., R.D., T.S.K., T.W., M.E.A., M.S., R.N., F.R., and E.W. led the study. A.F., T.W., M.E.A., M.S., and E.W. conceptualised the study. P.H.S., D.J.L., R.D., T.S.K., T.W., M.E.A., M.S., R.N., F.R., and E.W. supervised the research. M.E.A., K.K., R.D., T.W., R.N. and E.W. acquired funding for research. E.K.I-P, A.R-M, A.I., A.P., W.B., A.V., L.S., A.J. Stern, K.K., D.J.L., R.D., T.S.K.. M.E.A., M.S., R.N., and F.R. were involved in developing and applying methodology. E.K.I-P, A.R-M, A.I., A.P., W.B., A.S.H., R.A.H, T.V., H.M., A.V., L.S., A. Ramsøe, A.J. Schork, A. Rosengren, L.Z., P.H.S., T.S.K., M.E.A., M.S., and F.R undertook formal analyses of data. E.K.I-P, A.R-M, A.I., A.P., A.F., W.B., K.G.S., A.S.H., R.A.H, T.V., A.J. Stern, A. Ramsøe, A. Rosengren, L.Z., P.H.S., D.J.L., T.S.K., M.S., F.R. and E.W. drafted the main text. E.K.I-P, A.R-M, A.I., A.P., A.F., W.B., K.G.S., A.S.H., R.A.H, T.V., A.J. Stern, G.S., A. Ramsøe, A. Rosengren, L.Z., P.H.S., D.J.L., M.S., and E.W. drafted supplementary notes and materials. E.K.I-P, A.R-M, A.I., A.P., A.F., W.B., K.G.S., A.S.H., R.M., F.D., R.A.H, T.V., H.M., A. Ramsøe, A.J. Schork, L.Z., K.K., P.H.S., D.J.L., R.D., T.S.K., T.W., M.E.A., M.S., R.N., F.R., and E.W. were involved in reviewing drafts and editing. All co-authors read, commented on, and agreed upon the submitted manuscript.

Corresponding authors

Correspondence to Rasmus Nielsen (rasmus_nielsen@berkeley.edu), Fernando Racimo (fracimo@sund.ku.dk) and Eske Willerslev (ew482@cam.ac.uk)

Ethics declarations

Competing interests

The authors declare no competing interests.