Ancient human genomes reveal a hidden history of strong selection in Eurasia

Yassine Souilmi[1,a,*], Raymond Tobler[1,a,*], Angad Johar[1,10], Matthew Williams[1], Shane T. Grey[2,3], Joshua Schmidt[1], João C. Teixeira[1], Adam Rohrlach[4,5], Jonathan Tuke[4,6], Olivia Johnson[1], Graham Gower[1], Chris Turney[7], Murray Cox[8], Christian D. Huber[1,b,*], Alan Cooper[9,b,*]

[1]Australian Centre for Ancient DNA, The University of Adelaide, Adelaide, SA 5005
[2]Transplantation Immunology Group, Immunology Division, Garvan Institute of Medical Research
[3]St Vincent's Clinical School, Faculty of Medicine, University of New South Wales, Darlinghurst, New South Wales, Australia
[4]ARC Centre of Excellence for Mathematical and Statistical Frontiers, The University of Adelaide, Adelaide, South Australia 5005, Australia
[5]Department of Archaeogenetics, Max Planck Institute for the Science of Human History, Jena, Germany
[6]School of Mathematical Sciences, The University of Adelaide, Adelaide, South Australia 5005, Australia
[7]Chronos ^{14}Carbon-Cycle Facility and Palaeontology, Geobiology and Earth Archives Research Centre (PANGEA), University of New South Wales, Sydney, NSW 2052
[8]Statistics and Bioinformatics Group, School of Fundamental Sciences, Massey University, Palmerston North 4410, New Zealand
[9]South Australian Museum, Adelaide, SA 5005, Australia and BlueSky Genetics, PO 287, Ashton, SA 5137, Australia

[a] These authors contributed equally
[b]These authors contributed equally
*Correspondence: yassine.souilmi@adelaide.edu.au (Y.S.), raymond.tobler@adelaide.edu.au (R.T.), christian.huber@adelaide.edu.au (C.D.H.), alanjcooper42@gmail.com (A.C.)

Abstract

The role of selection in shaping genetic diversity in natural populations is an area of intense interest in modern biology, especially the characterization of adaptive loci. Within humans, the rapid increase in genomic information has produced surprisingly few well-defined adaptive loci, promoting the view that recent human adaptation involved numerous loci with small fitness benefits. To examine this we searched for signatures of hard sweeps – the selective fixation of a new or initially rare beneficial variant – in 1,162 ancient western Eurasian genomes and identified 57 sweeps with high confidence. This unexpectedly extensive signal was concentrated on proteins acting at the cell surface, and potential selection pressures include cold adaptation in early Eurasian populations, and oxidative stress from carbohydrate-rich diets in farming populations. Critically, these sweep signals have been obscured in modern European genomes by subsequent population admixture, especially during the Bronze Age (5-3kya) and empires of classical antiquity.

Keywords

Ancient DNA, aDNA, Adaptation, Admixture, Eurasia, Selection, Demography, Population Structure, Sweep, Allele Frequency

INTRODUCTION

The rapidly growing availability of genomic datasets provides a novel ability to address one of the key outstanding questions in evolution, namely the role of natural selection in shaping biological diversity (Jensen et al., 2018). However, most studies to date have inferred evolutionary processes only from modern genetic variation, and such approaches can easily be confounded by complex aspects of population history that remain unknown. For example, signatures of genetic adaptation can be effectively masked by past phases of admixture between populations (Harris et al., 2018; Huber et al., 2014). Recent studies suggest admixture is ubiquitous in natural populations (Mallet et al., 2016) but this factor is not explicitly accounted for in most selection studies, implying that our current understanding of the tempo and mode of selection in natural populations may be quite inaccurate (Harris et al., 2018; Kelley et al., 2006). In humans, where some of the largest genomic datasets are available, population admixture may help explain the surprising rarity of genetic signatures of fixed hard selective sweeps (where a new or initially rare beneficial mutation increases from low to 100% frequency in a population; see Material and Methods; (Hermisson and Pennings, 2017; Hernandez et al., 2011; Pritchard et al., 2010)). While scans of modern human genomes have identified several putative signals of positive selection, only a handful of examples are uncontroversial (Szpak et al., 2019), such as lactase persistence (*LCT (Enattah et al., 2007)*) and skin pigmentation (*SLC24A5 (Marciniak and Perry, 2017)*). Although these selected alleles do appear at high frequencies in modern Europeans (>60% and >95%, respectively; gnomAD (Karczewski et al., 2019)) they remain partial sweeps, as they are not present in all individuals. Indeed, the relative lack of fixed hard sweeps in human genomes has prompted suggestions that most recent selection may have involved numerous genetic variants with small fitness effects, which do not become fixed in a population (i.e. polygenic selection) (Hernandez et al., 2011; Pritchard et al., 2010). However, uncertainties remain about the ubiquity of polygenic selection in human adaptation as even the best example, height differences in Europe, has recently been questioned as a possible artefact (Sohail et al., 2019).

While a handful of ‘real-time’ genetic records of adaptation exist, they are generally limited to either microbial species (Feder et al., 2019) or experiments involving rapidly reproducing organisms in artificially controlled conditions (Lang and Desai, 2014), so that it is uncertain how they relate to the complex ecological and demographic circumstances of natural populations (Bailey and Bataillon, 2016). In contrast, the near-exponential growth in ancient human genomic datasets has created one of the first opportunities to directly observe the tempo and mode of genomic selection in a natural population over more than a thousand generations, effectively making humans a model system for selection. Importantly, this genetic record covers the critical period as humans moved from hunter-gatherer to farming-based lifestyles, involving major changes in diet, sedentary living conditions, epidemiology, and pathogen pressures. By temporally linking selected genes to these changes, ancient human genomes can provide crucial contextual information to identify how selection pressures have shaped modern human phenotypic diversity and disease, and illustrate evolutionary processes more generally.

The majority of currently available ancient human genome-wide datasets, which range in age from 1-45kya, are concentrated in Western Eurasia (Marciniak and Perry, 2017) (Figures 1, S1A; Table S1). Previous population analyses of these ancient genomes have revealed evidence of a complex history of Late Pleistocene European hunter-gatherer populations (Fu et al., 2016), which by the early Holocene (~10-12kya) were distributed in a cline from the west (termed Western Hunter-Gatherers, *WHG*) to the east (Eastern Hunter-Gatherers, *EHG*) (Skoglund and Mathieson, 2018). Around 8.5kya, a large migration of farming populations from the Anatolian region (Anatolian Early Farmers; *Anatolian_EF*) into Europe initiated a major period of genetic admixture with the extant Western Hunter-Gatherers (Skoglund and Mathieson, 2018) (Figure 1). The descendant populations subsequently underwent a further pronounced phase of genetic admixture with pastoralists arriving from the Eurasian Steppe (*Steppe*) at the beginning of the Bronze Age, around 5kya (Skoglund and Mathieson, 2018) (Material

and Methods). Following this, the emergence of successive large territorial empires encompassing much of Europe, the Near East, and north Africa (e.g. Neo-Assyria, Macedon, Roman, Ottoman) during classical antiquity (~3kya) initiated an unprecedented period of human movement (Şeker, 2013; ʿÔdēd, 1979) that facilitated further large-scale admixture across western Eurasia (Haber et al., 2019; Marcus et al., 2020).The disparate population histories of the hunter-gatherers, early farmers, Steppe pastoralists, and subsequent empires provides a unique opportunity to explore genomic signatures of selection, and the consequences of population admixture.

RESULTS

Dataset

We constructed the first publicly accessible Online Ancient Genome Repository (OAGR: www.oagr.org) containing the genomic information and metadata from over 2,200 available ancient human genome-wide datasets (including both shotgun sequenced genomes and high-density SNP scans), ranging in age from 1kya to 45kya (Figures 1, S1A; Table S1). We manually curated metadata for each sample including the archaeological context, molecular preparation methods, and analytical tools used to process the samples, and then reprocessed each genomic dataset with the same bioinformatic pipeline to mitigate potential analytical batch effects. We analyzed 1,162 ancient western Eurasian genome datasets (mostly high-density SNP scans; see Material and Methods) which could be grouped into 18 distinct ancient populations based on their archaeological and genetic relationships, ranging from the Mesolithic (15-8.5kya) to the Bronze Age (5kya) (Figures 1, S1, Table S1; Material and Methods). The 1,162 ancient samples also contained many Late Pleistocene individuals that could not be grouped into distinct populations, including five moderate to high coverage genomes from across Eurasia between 45kya to 20kya (e.g. Ust'-Ishim ~45kya (Fu et al., 2014a), Kostenki14 ~37kya, GoyetQ-116 ~35kya, Věstonice16 ~30kya, and El Mirón ~19kya (Fu et al., 2016)). For comparison, we also analyzed five modern human populations (3 from Europe: northern Europeans, USA (*CEU*); Finnish (*FIN*); Tuscans

(*TSI*); 1 Asian: Han Chinese (*CHB*); and 1 African: Yoruba (*YRI*) (1000 Genomes Project Consortium et al., 2015)). The genomic sequences within each population were aligned and scanned for evidence of hard selective sweeps, which have a well-characterized genomic signature created by the rapid fixation of new or rare beneficial mutations, resulting in genomic regions of anomalously low diversity (see Material and Methods). To do this, we used SweepFinder2 (Huber et al., 2015), which applies a dynamic sliding window approach to examine site frequency spectrum (SFS) patterns at dense 1kb intervals. By contrasting each window against SFS patterns from the same genomic background, SweepFinder2 tests explicitly for the signature of hard selective sweeps while controlling for demographic history and population structure, which often cause false positives in such tests (Huber et al., 2015). Importantly, testing our sweep detection pipeline on simulated genomic datasets modelled upon Eurasian demographic history confirmed that it is robust to false positives resulting from missing data, ascertainment bias, and alignment error (Figure S2; see Material and Methods). Outlier genes were determined for each population through comparison with SweepFinder2 scores for a set of ~21,000 annotated human genes, and subsequently binned into a final set of candidate sweeps (see Material and Methods).

In direct contrast to previous studies of modern human genomes (Szpak et al., 2019) we were able to identify large numbers of hard sweeps in the ancient west Eurasian populations, identifying 57 with high confidence (Figures 2, S3; Table S2 – study-wide false positive rate of <11%; see Material and Methods and Figure 3A). This surprisingly extensive signature of selection is supported by the observation that around 90% of the sweeps show significantly inflated levels of genetic divergence relative to African populations (F_{st}; see Material and Methods and Figure S2F), consistent with the action of strong positive selection since human arrival in Eurasia. For each of the 57 sweeps we inferred the set of alleles most likely carried on the selected haplotype, then scanned all 1,162 ancient samples for the oldest individual that carried at least 95% of these sweep alleles (Material and Methods). The date associated with the oldest sample in

the dataset was used as evidence that the selection pressure was present at that time point, even if the locus was potentially not yet fixed across all individuals in the population(s) undergoing selection.

Ancient sweeps are masked in modern European genomes

Most of the 57 hard sweeps (39, or ~70%) exhibit evidence of selection by 35kya (Material and Methods) although none are present in all of the ancient western Eurasian populations (Figures 2, S3), indicating that they had not been fixed in the initial small founding population that moved out of Africa into Eurasia around 55-50kya (dated via the Neandertal-introgressed DNA within the Ust'-Ishim genome) (Skoglund and Mathieson, 2018). However, the 39 hard sweeps observed by 35kya continue to remain evident amongst the three major founding populations of modern European genetic diversity (i.e. *Western Hunter-Gatherers* (16/39 sweeps, 41%), *Anatolian Early Farmers* (64%), and *Steppe* (62%); Table S2; Figures 2, S3, S5D), so we term these 'Ancestral West Eurasian' sweeps. Given this, it is surprising that only 2 of the 57 hard sweeps we detect (*AGO1* and *ABCF1*; named after the first gene found in each sweep, Figure S3) are still identified as outliers in selection scans of modern European populations, with both having been reported in previous studies (Pritchard et al., 2010). This dramatic decline cannot be readily explained by the degradation of the ancient hard sweep signals over time through random allele frequency changes (i.e. genetic drift) and new mutations, as hard sweeps in humans are expected to remain visible for more than 60,000 years in the absence of other effects (Sabeti, 2006). Indeed, the most recent sweep we detect (*AGAP2*; Figures 2, S3; Table S2), in Anatolian Early Farmers, appears to have only come under selection by 9.5kya.

An alternative explanation is that the signals of hard sweeps have been obscured by population admixture, through the introduction of non-selected haplotypes from the incoming population, creating genomic patterns more compatible with partial sweeps, or other modes of selection such as soft sweeps (where a beneficial variant sweeps to fixation on multiple genomic backgrounds (Material and Methods; e.g. by recombining

with non-sweep haplotypes introduced after the admixture event) (Huber et al., 2014). Indeed, changes in the frequency of the 57 sweep haplotypes across time (see Material and Methods) show that the loss of the ancient sweep signals appears to closely coincide with the timing of known major genetic admixture events in European prehistory (Skoglund and Mathieson, 2018) (Figures 2, S5). Many of the sweeps appear to decrease in frequency at the onset of the expansion of Anatolian Early Farmers into Europe (~8.5kya), and especially in the Bronze Age following the migration of Steppe pastoralists (~5kya) (Figures 2, S5A-C, F, G). The ancient sweep signals become further diluted in the past 2,000 years, corresponding to classical antiquity and the extensive movement of populations associated with western Eurasian empires (Figure 2). The latter involved groups whose ancestry may have contained genetic signatures reflecting a different selective history from the Ancestral West Eurasians, such as northern African and Levantine populations (Botigué et al., 2013; Lazaridis et al., 2017).

To examine this further, we simulated European population genomic datasets using plausible demographic parameters taken from a recent study of ancient Eurasian genomes (Kamm et al., 2019a) and tested our ability to detect sweeps after an admixture event. Briefly, we simulated strong selection (selection strength, *s*, ranging from 1% to 20%, see Figure S6) starting from ~30 to ~200 thousand years ago that resulted in the fixation of an initially new or rare beneficial locus (i.e. a hard sweep), then examined our power to detect the sweep signal in modern populations 4 thousand years after an admixture pulse that decreases the selected haplotype frequency by ~30% (Figures 4, S7). We observe a minimum 2-3 fold reduction in our power to detect hard sweeps that started within the past 30,000 years compared to demographic histories only comprising genetic drift (Figure 4, S7) – despite selection being sufficiently strong for the beneficial mutation to eventually fix again following the admixture event (e.g. the expected fixation time for a beneficial variant starting at 70% frequency with s = 10% is ~3,000 years; see Material and Methods). This intriguing result suggests that hard sweep signals may ‘soften’ after an admixture event even

when the selection pressure persists, potentially because recombination between selected and neutral introgressed haplotypes dilutes the sweep signal. This is consistent with 14 of our 57 sweeps appearing as soft sweeps in a recent study of modern European genomes (Schrider and Kern, 2017).

We also examined whether the impacts of admixture would mean that the age of the sweep itself influences whether it can still be detected in modern European populations (Material and Methods). We expect that the Ancestral West Eurasian hard sweeps, which occurred closer in time to the out-of-Africa migration, should be present at high frequencies in descendant ancient European populations and therefore more likely to survive subsequent admixture events by virtue of being present in both admixing groups. In contrast, more recent sweeps are likely to be local in distribution, increasing the chances that admixture events will erase their selection signal from modern European genomes. Indeed, previous selection scans of the modern European genomes used in our study (Pickrell et al., 2009; Racimo, 2016; Schrider and Kern, 2017) confirm that tests for positive selection are more than twice as likely to detect the Ancestral West Eurasian sweeps (i.e. those arising before 35kya) than those arising afterwards (see Material and Methods; Figures 2, S5H, Table S2). Importantly, these previous selection scans only detected around half (32) of the 57 sweeps we identify, and only 7 of these were characterised as hard sweeps, indicating the power of admixture to obscure signals of selection (Material and Methods).

Adaptation to temperate Paleolithic Eurasian environments

Amongst the 57 sweeps we observe in recent western Eurasian prehistory, only eight sweeps either contain a single gene or a handful of genes that share a common biological process (Table S2), limiting our ability to infer the underlying drivers of adaptation. Hence, we examined if the sweep selection signals distributed throughout the genome might be concentrated within specific biological pathways, as this could indicate potential processes (e.g. cellular, metabolic, or immune-based) targeted by selection (Material and Methods).

We observe a significant enrichment of selection signals in seven unique biological pathways (from ~1,800 tested; see Material and Methods, Table S3), including five metabolic pathways, and one each for cell surface interactions and immunity (Figure 4). Notably, none of the seven pathways showed significant enrichment in modern populations, implying that admixture has also obscured these signals in modern genomes. Four of the metabolic pathways shared across multiple ancient populations (Figure 5) are collagen specific – a metabolic process that has been shown to be a frequent target of selection in mammals living in circumpolar regions (Yudin et al., 2017). In addition, six of the eight sweeps with either a single gene or a single biological process are Ancestral West Eurasian (>35kya), and are either involved in lipid metabolism (two sweeps) or with nervous tissue maintenance (four sweeps) (Material and Methods), both of which have been associated with adaptation to cold environments (Brandefelt et al., 2011; Daló et al., 2007; Nedergaard and Cannon, 2018). A potential association with cold adaptation is interesting as this has been identified as a likely important factor as Anatomically Modern Humans (AMH) first moved into Eurasian environments (Silvert et al., 2019) – which were around 5°C cooler than those of their northern African origins at 55-50kya (Brandefelt et al., 2011) (Material and Methods).

The other metabolic pathway (*Protein digestion and absorption*) was only detected as being under selection in Anatolian Early Farmers and later farming populations. This is also potentially interesting as the pathway is involved in the cellular production of glutathione, which is a key component in removing potentially damaging oxygen free radicals. The adoption of carbohydrate-rich grain-based diets during the Neolithic has been linked with both nutritional and micronutrient stress (Cordain et al., 2005) and metabolic dysregulation (Rani et al., 2016) including the production of oxygen-bearing radicals. Oxidative stress, caused by the inadequate removal of oxygen-bearing radicals, is now a key causative factor in a variety of modern sedentary lifestyle diseases (Rani et al., 2016).

Adaptation was concentrated on proteins located on cell membranes

In order to further investigate the functional significance of the seven biological pathways enriched for selective signals, we tested for specific patterns in the cellular location of the loci with strong selection signals. Interestingly, the seven pathways are over-represented with genes encoding proteins acting at the cell membrane, including the enriched pathway specific to cell-surface interactions itself (*Focal Adhesion*); see Material and Methods, Figure 5, Table S4). This is potentially important as cell surface proteins coordinate appropriate cellular responses to extracellular information, including potential stressors, and are particularly relevant for modern medicine where they represent ~60% of the proteins directly targeted by FDA-approved drugs (www.drugbank.ca). When classifying our genes according to their cellular location in the Human Cell Atlas (HCA; www.humancellatlas.org), two membrane-bound categories (*plasma membrane* and *cell junctions*) were consistently overrepresented across multiple ancient populations (Figure S10). This consistent enrichment of genes across populations suggests the cell-environment interface may be a particularly important selective target for novel evolutionary pressures, such as those faced by the first AMH populations entering Eurasia.

DISCUSSION

The small effective population size and low mutation rates characteristic of humans mean that hard sweeps are expected to be a relatively common form of selection in the recent history of our species (Hermisson and Pennings, 2017). Our results suggest that hard selection has indeed been common in Eurasian populations, but recent admixture, particularly in the past 5kya, appears to have hidden many of the genetic targets of selection in modern European genomes (Figures 2, S5). The associated functions of the sweeps we can identify would suggest at face value that cold temperatures and novel subsistence strategies, more than pathogens, appear to have been important determinants of selection in ancient Eurasia. Either way, these selected loci were

previously associated with a fitness effect and are now segregating in modern Europeans, raising the possibility that they may be contributing to genetic variation in modern diseases. Indeed, we note that genes identified in genome-wide association studies (i.e. GWAS) are significantly more likely to appear amongst the outlier genes we identify here (i.e. those with $q < 0.01$ in at least one population) compared to genes lacking known GWAS associations (Table S5; see Material and Methods). While such patterns may also be influenced by demographic structure, the availability of large numbers of high-coverage ancient human genomes in the next decade will make it possible to improve the resolution of candidate sweeps, potentially to the level of causal mutations (Akbari et al., 2018). In addition, information about the temporal relatedness of past genetic changes could help identify key interacting genes or networks, guiding research efforts investigating target loci. Consequently, the Online Ancient Genome Repository offers a potentially important tool for improving health research in current populations, in addition to providing a window into human history and a uniquely informative view of natural selection in action.

Our findings are not a good match for either of the two major conventional models of selection, which envision adaptation occurring either through occasional sweeps, or via much more subtle and genomically-diffuse polygenic selection. Instead, the sweep signatures in our study indicate that moderate to strong selective pressures were operating over a large number of loci: 41 of the 57 candidate sweeps had inferred selection coefficients > 1% (Figure S5E; see Material and Methods). The evolutionarily short time periods involved were likely insufficient for selection to be acting on *de novo* mutations, and the large number of hard sweep patterns suggest that selection predominantly operated on rare standing genetic variants (Przeworski et al., 2005), with surprisingly swift impacts across multiple loci. While rapid adaptation from standing variation has been reported in domestic species (MacHugh et al., 2017) and within adaptive radiations of natural populations (e.g. cichlid fish (Brawand et al., 2014)), few selected alleles become fixed and most show more moderate frequency changes. The

human genetic data suggests these signals might actually reflect the consequences of admixture with other populations – e.g. the frequent admixture between domesticates and their wild progenitors (MacHugh et al., 2017) – and demonstrate the need for admixture to be explicitly modelled in future selection studies. Given the growing evidence that admixture is frequent in other species (Mallet et al., 2016), our results suggest the extent of past hard sweep events have been underestimated, biasing our understanding of adaptation in natural populations, including humans.

ACKNOWLEDGMENTS

We thank three anonymous reviewers for their suggestions which helped to improve the manuscript. We also thank Fernando Racimo, Iain Mathieson, Nigel Bean, Wolfgang Haak, Bastien Llamas, Laura Weyrich, Jimmy Breen, Julien Soubrier, Christopher Franks, Cathy Miller, Michael Brooks, and Ryk Ward (deceased). We gratefully acknowledge the museums, archaeological collections, collectors and curators responsible for the original skeletal material and making samples available for genomic analysis; **Funding:** the Australian National Database Service for funding the creation of the OAGR database, and Nectar Cloud for hosting the database server. Funding was provided by the University of Adelaide, Bioplatforms Australia, and Australian Research Council Grants to A.C. (FL140100260), C.D.H. (DE180100883), and R.T. (DE190101069).

AUTHOR CONTRIBUTIONS

A.C., Y.S., R.T. and C.D.H. conceived the study, Y.S., R.T., M.W. assembled the dataset, Y.S., R.T., A.J., C.D.H., G.G., M.C., A.R., O.J., and J.T. performed analyses, A.C. and C.D.H supervised analyses, Y.S., R.T., A.J., C.D.H., S.G., C.T., A.R., O.J.,

J.S., and A.C. interpreted the results, A.C., Y.S., R.T., A.J., and C.D.H. wrote the paper with input from all co-authors.

DECLARATION OF INTERESTS

The authors declare no competing financial interests.

MAIN FIGURE TITLES AND LEGENDS

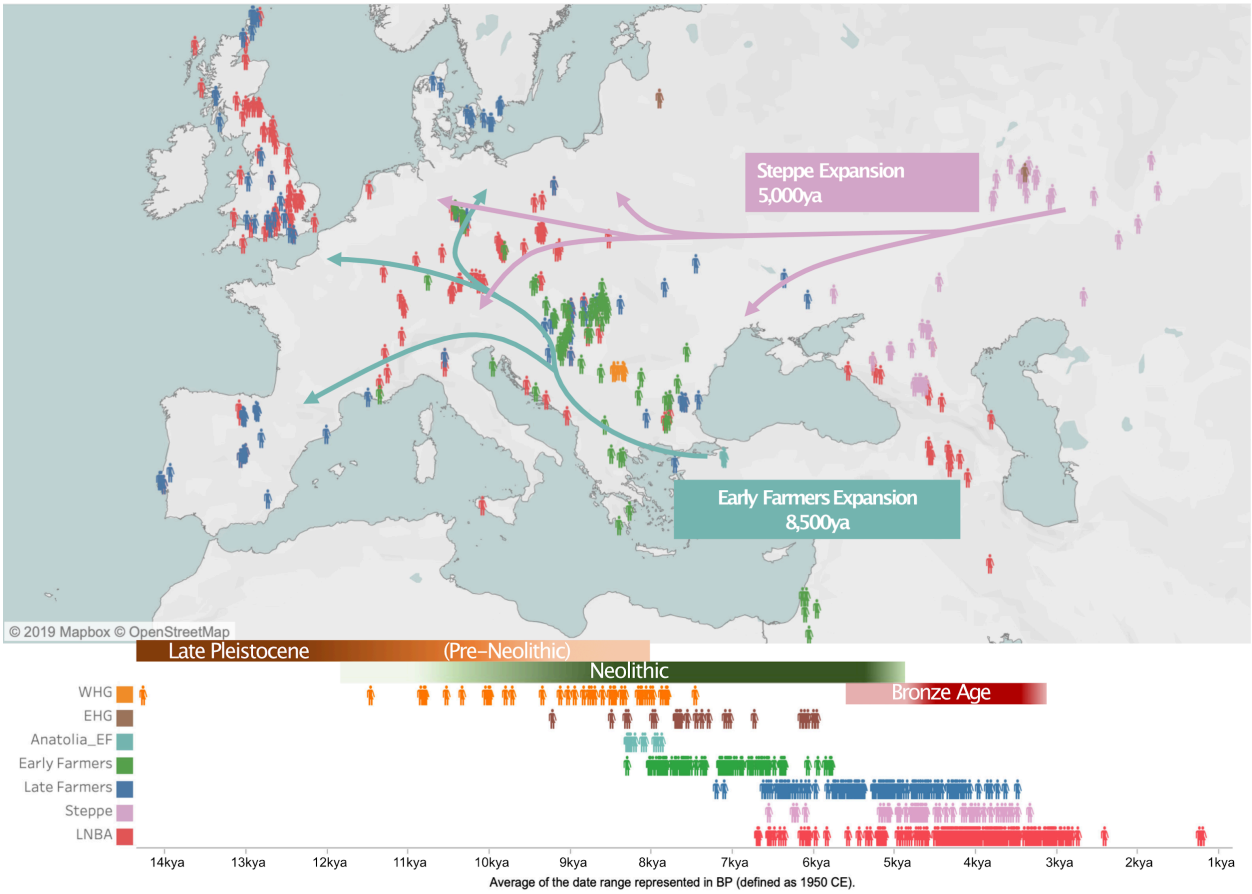

Figure 1. Geographic and temporal distribution of 1,162 ancient Eurasian samples used in this study.

Each human symbol represents a sample and the colors indicate different populations groups classified into broad groupings according to archaeological records of material culture and lifestyle (colors indicated at the bottom right-hand side; see (Material and Methods). The green lines depict the generalized migration route of Anatolian farmers (*Anatolia_EF*) into Europe ~8.5kya, where they mixed with Western Hunter-Gatherers (*WHG*; *EHG* refers to the contemporaneous Eastern Hunter Gatherers) to create the Early European Farmers (*EF*). Similarly, the pink arrows represent the generalized movement of Steppe pastoralists (*Steppe*; samples east of the Ural Mountains not

shown) which resulted in admixture with Late European Farmers (*LF*) ~5kya, giving rise to Bronze Age (*LNBA*) societies.

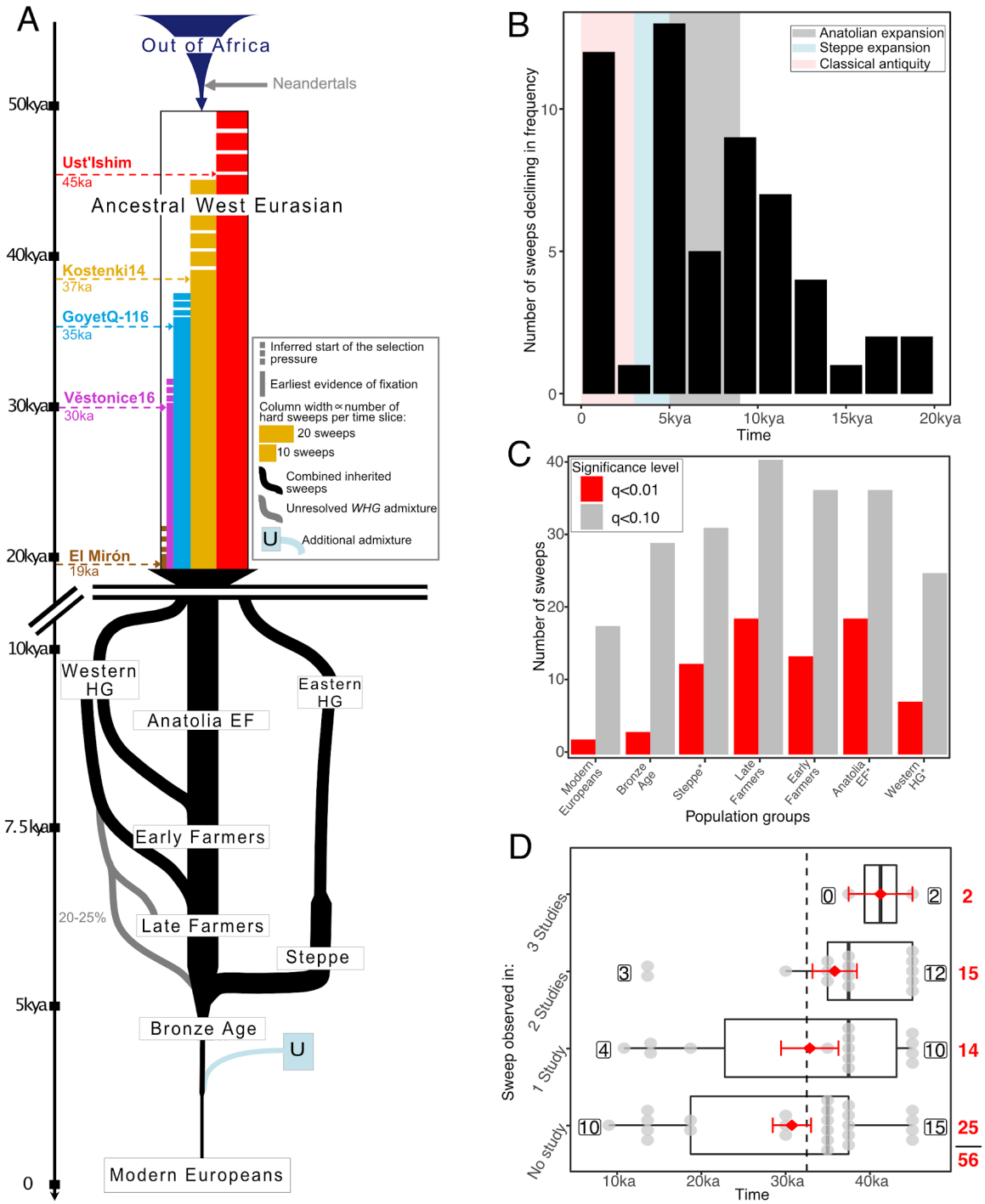
A
Out of Africa
Neandertals
50kya
Ust'Ishim
45ka
Ancestral West Eurasian
40kya
Kostenki14
37ka
GoyetQ-116
35ka
Věstonice16
30ka
30kya
El Mirón
19ka
20kya
10kya
Western HG
Eastern HG
Anatolia EF
7.5kya
Early Farmers
20-25%
Late Farmers
Steppe
5kya
Bronze Age
U
0
Modern Europeans
Inferred start of the selection pressure
Earliest evidence of fixation
Column width ∝ number of hard sweeps per time slice:
20 sweeps
10 sweeps
Combined inherited sweeps
Unresolved WHG admixture
Additional admixture
B
Anatolian expansion
Steppe expansion
Classical antiquity
Number of sweeps declining in frequency
Time
0
5kya
10kya
15kya
20kya
C
Significance level
q<0.01
q<0.10
Number of sweeps
Population groups
Modern Europeans
Bronze Age
Steppe*
Late Farmers
Early Farmers
Anatolia EF*
Western HG*
D
Sweep observed in:
3 Studies
2 Studies
1 Study
No study
Time
10ka
20ka
30ka
40ka
2
15
14
25
56

Figure 2. A history of hard sweeps in west Eurasia and the dilution of selection signals in modern populations through admixture.

(A) Schematic representation of the temporal distribution of the top 57 outlier hard sweeps (*q*-value < 0.01) in Western Eurasian populations, with vertical line width proportional to the number of sweeps inferred to exist in the medium to high coverage Late Pleistocene samples Ust'-Ishim: 45kya, Kostenki14: 37kya, GoyetQ-116: 35kya, Věstonice16: 30kya, El Mirón: 19kya , and 12 Holocene population groups (8 shown) in the lower portion (black lines). The temporal origin of the sweeps (first detected presence: solid line, suspected presence: dashed line) and distribution are shown through time. Population admixture during the Bronze Age, and subsequently during classical antiquity (represented by the light-blue "U" for unspecified), progressively introduces genetic diversity that obscures evidence of the 57 sweeps in modern populations. (B) The time interval when the selected haplotype first drops below its maximum historical frequency for each of the sweeps (time binned into 2,000 year intervals; sweep *10:122.9-123* omitted due to having insufficient data for inference; see Material and Methods). Marked declines in sweep frequency closely coincide with historic periods of extensive population movements and admixture in west Eurasia; i.e. Anatolian farmer migration (~8.5kya; blue shading), Steppe migration (~4.5kya; blue shading), classical antiquity (~3kya; red shading). (C) The number of the 57 sweeps (through time from the right) that were either significant (*q* < 0.01, red bars) or evident (*q* < 0.10, grey bars) in ancient west Eurasian populations. Sweeps were tallied separately for European Early Farmers (EF, 2 populations), Late Farmers (LF, 3 populations), and the Bronze Age (LNBA, 2 populations; excluding *South_Caucasus_LNBA* which has distinct ancestry from other LNBA populations; see Material and Methods). (D) Boxplots showing that older sweeps are more likely to be reported as selection targets in genetic screens of modern Europeans (sweep *10:122.9-123* omitted; see Material and Methods). Sweeps are shown as grey dots, with inferred age of the sweep on the x-axis, and total number of sweeps detected in 3 previous studies of selection in modern Europeans shown on the right of plot. Of the 25 sweeps that were not detected in any of

the three studies (`No studies` on the y-axis), 15 (60%) were older than 35kya (text box left of dashed line; text box on right indicates sweeps older than 35kya). In contrast, sweeps reported as significant in either 1, 2, or all of the 3 previous studies become increasingly likely to arise before 35kya (70%, 80% and 100%, respectively; red diamonds and bars show the associated mean sweep ages and standard deviations). This is consistent with older sweeps being more broadly distributed, and therefore less susceptible to signal dilution from subsequent admixture events.

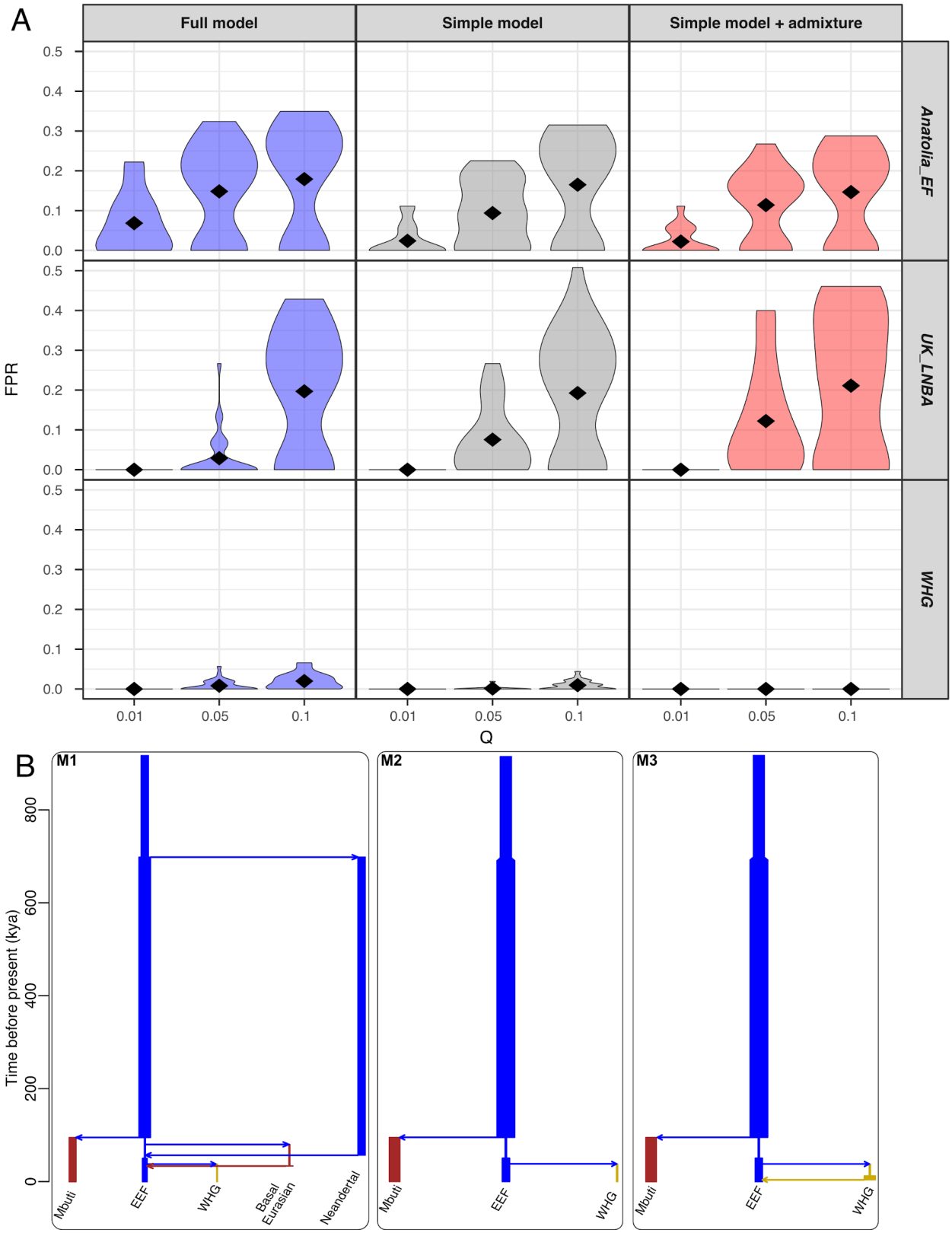
A
Full model
Simple model
Simple model + admixture
Anatolia_EF
UK_LNBA
WHG
FPR
0.0
0.1
0.2
0.3
0.4
0.5
0.01
0.05
0.1
Q
B
M1
M2
M3
Time before present (kya)
0
200
400
600
800
Mbuti
EEF
WHG
Basal Eurasian
Neandertal

Figure 3. Inferring the false positive rate for sweep detection.

(A) Violin plots of the false positive rate (FPR; y-axis) of our sweep detection pipeline estimated on simulated genomic data for three ancient western Eurasian populations – Western Hunter Gatherers (*WHG*), Anatolian farmers (*Anatolian_EF*), and a European Early Farmer population (*WesternEurope_LNBA*) (row panels). Simulations were performed for (B) three different west Eurasian demographic scenarios: 1. a model that includes historical admixture with Neanderthal and Basal Eurasians (M1; blue violins); 2. a simplified version of model M1 that lacks the Neanderthal and Basal Eurasian admixture events (M2; gray violins); 3. the same as model M2, but including a 30% admixture event between Western Hunter Gatherers and Early European Farmer populations ~4kya (M3; red violins). For each demographic model, 30 simulated genomic datasets were generated for each of the three ancient western Eurasian populations that replicated their specific data missingness and sample size. The FPR was estimated by dividing the number of outlier sweeps detected in the simulations by the number of empirically observed outlier sweeps at three different detection thresholds ($q < 0.01$, $q < 0.05$, $q < 0.10$; x-axis), averaged across the 30 simulated genomic datasets (violins show the distribution of FPR estimates for the 30 simulated genomic datasets, diamonds indicate the mean FPR). For all simulated populations and demographic models, the FPR was less than 7% at $q < 0.01$ (the threshold used to identify the 57 candidate sweeps in this study), which extrapolates to a conservative study-wide FPR of ~11% (see Material and Methods).

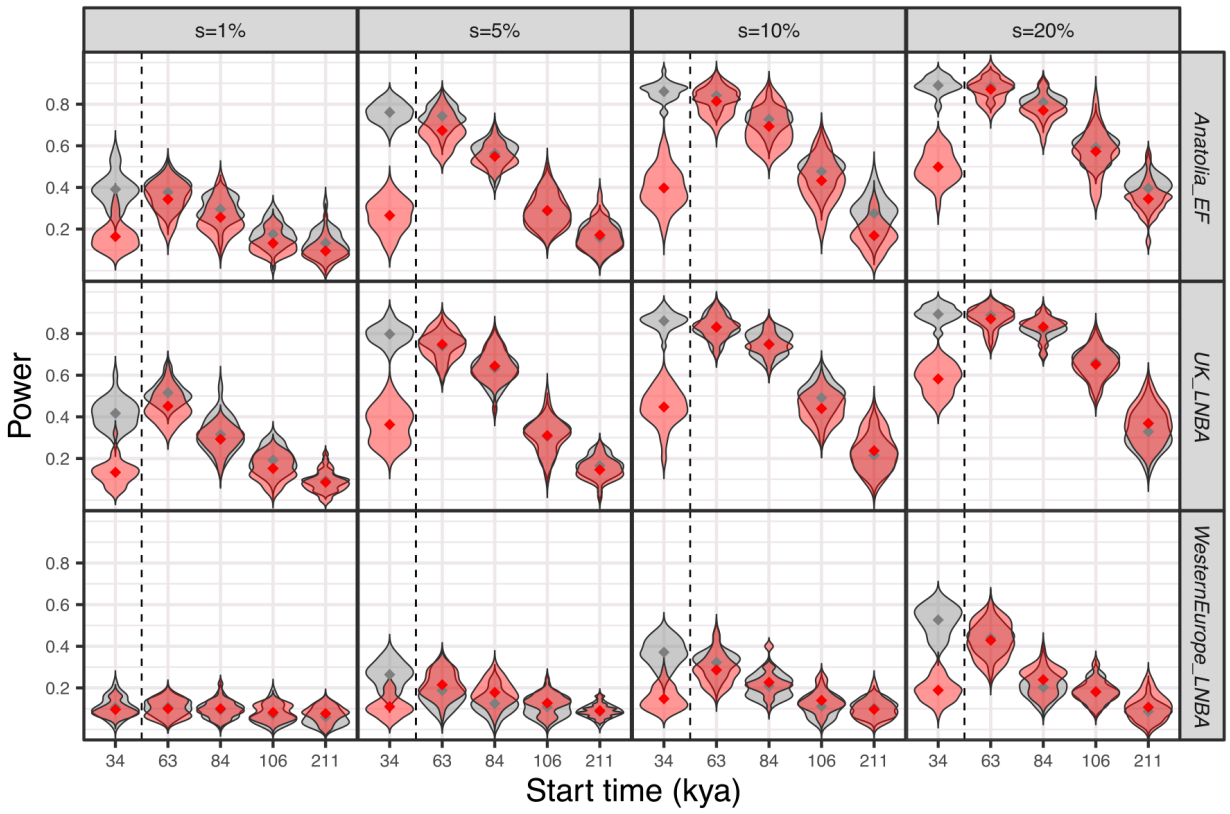

Figure 4. The impact of admixture on sweep detection in modern populations. To quantify the impact of admixture on sweep signals in modern European genomes, we inferred the power of our analytical pipeline to detect sweeps in simulated genomic datasets under two demographic scenarios: a model that included a 30% admixture event between Western Hunter Gatherers (WHG) and Early European Farmer (EEF) populations ~4kya (red violins) prior to genome sampling, and a model without this admixture event (gray violins) (i.e. models M3 and M2 shown in Figure 2). For both demographic models, we simulated 30 full-genome datasets that each contained 50 hard sweeps with four different selection strengths (*s,* column panels) and starting at different time points in Eurasian history (x-axis labels). This included sweeps starting after the genetic separation between WHG and EEF populations (violins on left of dashed line), where selection was confined to one of three historical populations (column panels), or in the common ancestral population of western Eurasians prior to

human departure from Africa (violins on right of dashed line). Notably, admixture has no impact on detection power when the sweep arose prior to separation between WHG and EEF populations (violins show distribution of power estimates for the 30 simulated genomic datasets, mean power indicated by diamonds; power measured at a 0.1% FPR – see Material and Methods). However, if the sweep started after this historical split the detection power declined at least two-fold following the admixture event relative to simulations without admixture – even though selection often fixed the beneficial haplotype again following the admixture event.

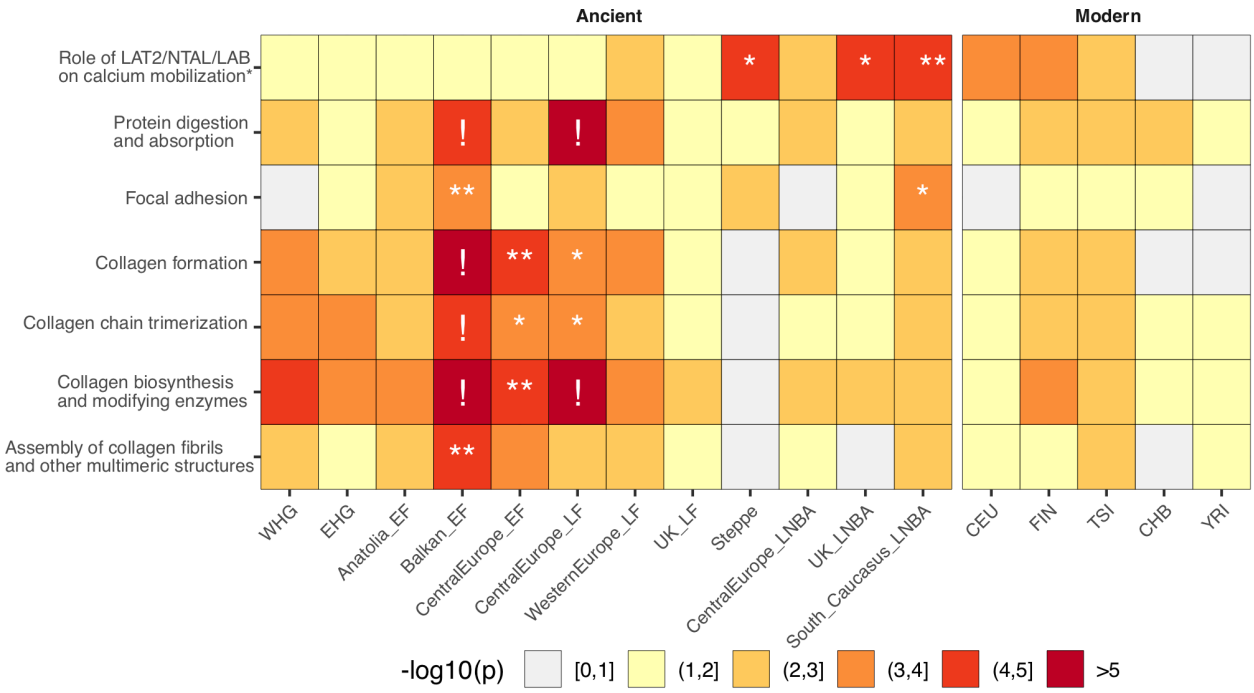

Figure 5. Biological pathways enriched for sweep signals across ancient and modern western Eurasian populations.

To determine if any human biological pathway was a particular target of strong positive selection in west Eurasian history, we applied a novel enrichment method, PolyLink (https://github.com/ACAD-UofA/PolyLink) to the complete set of transformed gene scores in each ancient and modern west Eurasian population. Of the 1,800 annotated biological pathways tested for enriched selection signals, seven pathways with general immune, cellular, and metabolic functions were significantly enriched in at least one population after correcting for multiple testing (q-value < 0.05). Consistent with our observations for the 57 candidate sweeps, the enriched pathway signals are absent from modern populations and suggest that admixture has also obscured this information. Different levels of significance are indicated after multiple testing correction (!: q <0.01, **: 0.01≤ q <0.05, *: 0.05≤ q <0.1), while colors indicate -$\log_{10}$ transformed p-values.

Material and Methods

1. Population designation

The assignment of individuals to populations was governed by minimizing temporal and spatial variability amongst samples that share genetic affinities defined in the ancient DNA literature (Table S1, Figures 1, S1), while also preserving the archaeological context and retaining enough individuals to ensure that our analyses were sufficiently powerful to detect selection (more details on pipeline power in Material and Methods section 4). The reported genetic relationships between samples were characterised using standard genetic clustering methods such as principal component analysis (PCA) and ancestry decomposition techniques (e.g. ADMIXTURE; (Alexander et al., 2009)). This classification process resulted in the assignment of 1,162 ancient human genomes into 18 unique populations (Table S1). We use a population nomenclature that follows the guidelines recently proposed for ancient DNA research (Eisenmann et al., 2018). Importantly, the shared ancestry and selection history of the populations in this study (Figures S1C-S1G) meant that the candidate sweeps were quite robust to different sampling choices for particular populations (Figures S3, S4; see Material and Methods section 4.5).

2. Data collection and processing

To produce a robust dataset and avoid potential bioinformatic batch effects, the raw sequence read data for 1,162 ancient genomic datasets (Table S1) was retrieved from the European Nucleotide Archive and processed through a standardized pipeline. To minimize the risk of modern contamination, the forward and reverse reads of the paired-end reads were merged (collapsed) using fastp (Chen et al., 2018), and only merged reads were retained (modern data is more likely to comprise large DNA fragments that do not collapse). All collapsed reads were filtered for potential residual adaptor sequences and chimaeras using Poly-X with fastp (Chen et al., 2018). The retained filtered set of sequence reads were aligned to the human reference genome (h37d)

using the Burrows-Wheeler Aligner v0.7.15 (Li, 2013). All mapped reads were sorted using SAMtools v1.3 (Li et al., 2009) and then realigned around insertions and deletions and potential PCR duplicate reads marked and removed using the Genome Analysis ToolKit (GATK) v3.5 (McKenna et al., 2010).

Prior to variant calling, all remaining aligned reads were screened and base-calls recalibrated for aDNA postmortem damage using mapDamage2 (Jónsson et al., 2013). To further limit the impact of postmortem damage on variant calling (Kircher, 2012), bamUtil (Jun et al., 2015) was used to trim 3 base-pairs from each of the 5' and 3' ends of each mapped read. From the resulting set of reads, pseudo-haploid variants were called at the set 1240k capture SNPs (Mathieson et al., 2015) found on the 22 autosomes, using a combination of SAMtools mpileup (Li, 2011) and sequenceTools (https://github.com/stschiff/sequenceTools). Pseudo-haploidization of read data is a standard strategy in aDNA analyses, whereby a single read is randomly sampled at each prespecified SNP position (Mathieson et al., 2015) in order to mitigate potential biases introduced by differences in coverage or post-mortem damage between samples (Haak et al., 2015). The 1240k capture was developed to minimize ascertainment in non-African populations and was used to generate data for most samples used in the study, whereby concentrating on the 1240k variants ensured a common and robust set of variants for the subsequent analyses. The pseudo-haploid variant calls were converted from EIGENSOFT format (Patterson et al., 2006; Price et al., 2006) to binary Plink format using EIGENSOFT. Plink v1.9 (Chang et al., 2015; Purcell et al., 2007) was used to assign samples to the predefined populations (Table S1) and convert the variants to reference polarized VCF files, with correct polarization being checked using BCFtools (Li, 2011). Finally, a custom Python script was used to generate the site frequency spectrum (SFS) input files for SweepFinder2 analysis (https://github.com/ACAD-UofA/OAGR_Selection_Paper_Helper_Scripts).

3. Selection pipeline

3.1. Sweep terminology

The classical definition of a hard sweep is based on the fixation of new beneficial allele and linked neutral variants (*sensu* (Smith and Haigh, 1974)). In contrast, soft sweeps capture scenarios where the beneficial allele is present prior to the onset of selection, leading to multiple haplotypes that can potentially be picked up by selection. In this study, however, we are primarily interested in the characteristic genetic signatures left by the fixation of a single sweeping haplotype – which is still possible when the beneficial variant is rare – rather than the initial frequency of the beneficial allele *per se* (i.e. a *de novo* mutation vs standing variation). Hence, a hard sweep in this study refers to positive selection leading to the fixation of a single haplotype, a definition that also encompasses selection from rare standing variation (Hermisson and Pennings).

3.2. Sweep scans

For all 18 ancient and 5 modern human populations, we computed the SweepFinder2 CLR (composite likelihood ratio) statistic (DeGiorgio et al., 2016; Huber et al., 2015) across the genome in successive 1kb intervals. The CLR statistic evaluates the evidence for hard selective sweeps in dynamically sized windows, by comparing the distribution of allele frequencies expected under a mathematical model of a hard sweep with the expectation under neutral evolution (larger CLR scores indicate more evidence for selection). The expected site frequency spectrum (SFS) under the hard selective sweep model is computed conditional on the neutral SFS, assuming a certain selection coefficient and recombination rate. The neutral SFS is based on the background SFS calculated from the whole genome, assuming that the influence of sweeps on the SFS on a genome-wide scale is negligible.

SweepFinder2 is robust to genome-wide effects such as ascertainment bias and demography (Nielsen et al., 2005) by allowing these processes to affect the expected

SFS under neutrality (i.e. the background SFS). Further, unlike many other selection methods, the assumptions on the input data for SweepFinder2 are suitable for the low coverage and ascertained nature of ancient DNA datasets (also see Material and Methods section 4 for explicit tests of the statistical properties of our full analytical pipeline). Since SweepFinder2 is only based on the spatial (genomic) pattern of allele frequencies but not on haplotype homozygosity or population differentiation, it is possible to detect selection without reference to a second population, calling genotypes, or phasing haplotypes. The reliance on an empirically estimated null model (i.e. the background SFS) and the model-based alternative hypothesis makes it both more powerful and more robust compared to alternate test statistics that are based on deviations of the SFS from expectations under the standard neutral model (e.g. Tajima's *D*, Fay and Wu's *H*).

Note that SweepFinder2 also has an option to detect sweeps based on local genomic reductions in diversity. However, we did not calculate this diversity-based metric since the accurate and unbiased estimation of diversity requires full genome data, whereas our dataset consists of an ascertained set of SNPs.

3.3. Outlier gene detection

Human gene annotations were obtained from the ENSEMBL database (Kinsella et al., 2011) (genome reference version GRCh37), which was accessed using the R biomaRt package (Durinck et al., 2009; Storey, 2003) (version 2.36.1). Of the 24,554 annotated 'genes' on the biomaRt database, we removed any that were not annotated in the NCBI database (ftp://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/Mammalia/Homo_sapiens.gene_info.gz) and also excluded those that lacked specific protein and RNA based annotations (in the biomaRt transcript_biotype field). This resulted in a list of 19,603 genes, from which we removed 26 additional genes that did not contain any polymorphic sites in our datasets (all being situated in the most terminal areas of chromosomes), leaving 19,577 genes

that were used in the subsequent analyses. For each population, all SweepFinder2 CLR scores were log_{10} transformed and assigned to one of 19,577 genes by binning the transformed scores within the genomic boundaries of each gene. The gene boundaries were extended by 50kb on either side in order to also capture *cis*-regulatory regions. Because this typically resulted in several scores being assigned to each gene, we took the maximum score to represent the evidence for a sweep involving that gene. Each gene score was corrected for gene-length using a non-parametric standardization algorithm (Daub et al., 2013; Iglewicz and Hoaglin, 1993), resulting in the gene scores having an approximately standard normal distribution. To do this, all genes were split into 10 approximately equally sized bins according to gene length. Within each bin, scores were adjusted using the following non-parametric standardization function (Iglewicz and Hoaglin, 1993):

$$Z_i = 0.675 \times \frac{y_i}{median(|y|)}$$

where $y_i = CLR_i$ - median(CLR) and i is an index for each gene in a given gene-length bin. The resulting set of adjusted gene scores approximated a standard Gaussian distribution – i.e. were Z scores, with mean = 0 and standard deviation = 1 – whereby we used standardized Gaussian quantiles to compute p-values for all genes for each population. Finally, to account for the expected inflation of false positives due to multiple testing, for each population we applied q-value correction (Storey, 2003) to the p-values. A q-value of 0.01 implies a false discovery rate of 1% per population.

Notably, when applying a one-sided test to determine outlier genes (i.e. those with scores lying in the upper tail), for a handful of populations the distribution of the resulting p-values was U-shaped rather than a J-shaped distribution (i.e. a combination of non-selected genes that have a standard uniform distribution, plus a set of selected genes that mostly include low p-values) that is assumed by the q-value correction method. Hence, p-values were calculated by applying a two-tailed test to genes with $Z > 0$ (since selected genes will have larger positive Z scores). This ensured that the expected *J*-

shaped *p*-value distribution was obtained for each population (Figure 2B). Importantly, this approach is not unnecessarily conservative, since our significance cutoff is based on *q*-values, i.e. the estimated false discovery rate (FDR) conditional on a specific the *p*-value threshold. This ensures that our *q*-values will still be accurately estimated and correctly assign the FDR rate to the desired value.

3.4. Candidate sweep classification

Sweeps were identified by determining a set of outlier genes across all populations, which were classified into sweep regions according to 1) the distance between the mid-point of neighbouring pairs of outlier genes (i.e. inter-gene distance) and 2) overlapping sweep regions between populations. Step 1 was performed independently for each population, whereby all outlier genes with midpoints that were less than a specific distance apart from the midpoint of a neighbouring outlier gene were collapsed into a single category. After generating the collapsed categories for each population, step 2 was applied to ensure that the sweep categories sharing at least one gene across different populations were considered as a single historical sweep (a parsimonious expectation based on shared genetic history of the populations).

We ran our sweep quantification pipeline at three *q*-value thresholds (i.e. $q < 0.01$, 0.05 or 0.10; which imply false discovery rates of 1%, 5% and 10% per population, respectively) and the three different inter-gene distances (midpoint distances less than 250kb, 500kb or 1Mb). As expected, changing the *q*-value had a large impact on the number of sweeps (ranging from ~50 for $q < 0.01$ to ~400 for $q < 0.1$), whereas changing the inter-gene midpoint distance had comparatively little impact overall, particularly at more stringent *q*-value cutoffs (Figure S3). Based on these results, we decided to use the most stringent *q*-value cutoff and the most liberal inter-gene distance to define a robust set of candidate sweeps for all further analyses. However, because this stringent cutoff might lead to the removal of potentially causal genes in a sweep (which could have values slightly lower than 0.01), we first defined our sweeps based

on the more permissive $q < 0.1$ threshold and then removed all sweeps that did not have at least one gene with $q < 0.1$. To further improve sweep determination, we removed populations with small sample sizes from the sweep determination process, as our own analyses based on simulated ancient human genomes (see Material and Methods section 4) and previous results of modern genomes suggest that SweepFinder2 has poor power to detect sweeps when the number of haploid genome copies being analyzed is 10 or less (Ferrer-Admetlla et al., 2014). Finally, to ensure that the sweeps being defined were all relevant to western Eurasian history, the two modern populations from East Asia (*CHB*) and Africa (*YRI*) were also excluded from the sweep classification process. This strategy resulted in a total of 57 candidate sweeps that were used in subsequent analyses. The Material and Methods contain additional technical details of the sweep detection pipeline, and plots showing the distribution of CLR scores and gene scores across all candidate sweeps are included in supplementary data S1-S57 (https://adelaide.figshare.com/s/20c0d1786a7ae8d3d45f).

4. Testing the analytical pipeline power and FDR

We assessed the statistical properties of our analytical pipeline by conducting extensive coalescent simulations using MSMS (Ewing and Hermisson, 2010). Of particular interest for the present study is measuring how the particular demographic history of Early European Farmers (*EEF*) and Western Hunter-Gatherers (*WHG*) – which includes the Out-of-Africa (OOA) expansion and admixture of migrant farmers from Antolian and *WHG* populations to create the *EEF* – impacts the false positive rate (FDR) and power of our analytical pipeline to detect hard sweeps.

To produce demographic histories relevant to the ancient Eurasian populations in this study, we used MSMS to simulate population genetic data using parameter values from the recently published demographic model of Kamm and colleagues (Kamm et al., 2019b). This study explicitly incorporated ancient Eurasian genomic data to model the demographic history of Eurasian hunter gatherers and early farmers, thereby providing

essential parameter estimates (i.e. the timing and intensity of population bottlenecks and admixture events) needed to accurately simulate genomic datasets for the populations in this study. Specifically, we conducted simulations under the following three demographic scenarios:

Model 1. The full model of Kamm et al. that includes the population bottleneck after the OOA expansion and ancient admixture events from a Neanderthal and a basal Eurasian population into the Out-of-Africa population (Figure 4). The effective population size in the common African ancestral population is fixed at 18,200, which is preserved across all three models; msms command:

```
msms -ms 31 1 -t 3000 -r 5265 -I 3 2 29 0 0 -n 1 0.95 -n 2 0.66
-n 3 0.11 -es 0.016 2 0.906 -en 0.016001 4 0.13 -ej 0.018 3 2 -
en 0.024 2 0.13 -es 0.027 2 0.97 -en 0.027001 5 1 -ej 0.038 4 2
-ej 0.045 1 2 -en 0.045001 2 1.6 -ej 0.33 5 2 -en 0.330001 2 1
```

Model 2. A simplified version of model 1 that ignores ancient introgression events from Neanderthals and basal Eurasians (Figure 4). msms command:

```
msms -ms 31 1 -t 3000 -r 5265 -I 3 2 29 0 0 -n 1 0.95 -n 2 0.66
-n 3 0.11 -ej 0.018 3 2 -en 0.024 2 0.13 -ej 0.045 1 2 -en
0.045001 2 1.6 -en 0.330001 2 1
```

Model 3. Uses the same simplified demography as model 2, but adds a 30% pulse of admixture from *WHG* into *EEF* at 4kya (Figure 4). The 30% admixture pulse is based on previous estimates of *WHG* genetic contributions into *EEF* ancestry (Lazaridis et al., 2016; Mathieson et al., 2015); msms command:

```
msms -ms 31 1 -t 3000 -r 5265 -I 2 2 29 0 -n 1 0.95 -n 2 0.66 -
es 0.0019 2 0.66 -en 0.006001 3 0.11 -ej 0.018 3 2 -en 0.024 2
0.13 -ej 0.045 1 2 -en 0.045001 2 1.6 -en 0.330001 2 1
```

For each of these three different demographic scenarios, we used MSMS (Ewing and Hermisson, 2010) to simulate 1,000 neutrally evolving 5Mb sequences for four ancient Eurasian populations: Western Hunter-Gatherers (*WHG*), Anatolia Early Farmers (*Anatolia_EF*), and two admixed Bronze Age EEF populations (*WesternEurope_LNBA* & *UK_LNBA*). In addition to capturing key branches of Eurasian history, these populations were also chosen to quantify the impact of sample size and data missingness in our ancient genomic datasets on the FDR and power of our analytical pipeline by incorporating the population-specific empirical patterns for both factors into our simulated datasets.

While SweepFinder2 is fairly robust to recombination rate variation (Nielsen et al., 2005), we performed separate simulations under three different recombination rates, including an 'average' rate of 1.45 cM/Mb (estimated by averaging across 1,000 randomly placed 1Mb windows using data from (Bhérer et al., 2017) and also 2- and 10-fold lower rates (0.72 cM/Mb and 0.145 cM/Mb). The two lowest rates were included to specifically test if the false positive rate (FPR) for sweep detection was particularly problematic in genomic regions with relatively low recombination rates.

Finally, our analyses were limited to a set of ~1.1 million variant positions, which were selected based on heterozygosity in a set of sequenced reference individuals from both African and non-African populations (Lu et al., 2011). Thus, our empirical data shows ascertainment bias towards more common genetic variation, which could lead to unexpected behaviour of the SweepFinder2 statistic. To incorporate this ascertainment into our simulations, for each replicate we sampled one modern African and one EEF individual and limited our analyses to SNPs that were heterozygous in either or both individuals. We simulated under a sufficiently high mutation rate such that we could select exactly 1,200 SNPs per replicated simulation after accounting for the ascertainment and missing data, which approximates the average number of SNPs found in a random 5Mb region in our datasets.

4.1. Analytical pipeline FDR under neutral demographic histories

Our analytical pipeline involves a series of steps that may lead to an inflated false positive rate (FPR) if some of our implicit underlying assumptions are violated. In particular, we have assumed that the gene scores used in this study will follow a standard Gaussian distribution. To test if our assumptions were met and investigate the statistical robustness of our pipeline, we created 30 artificial population genomic datasets for each simulated population and demographic scenario. Each artificial dataset was created by linking 565 randomly selected 5Mb sequences into full 'genomes'. We then applied our statistical pipeline to each artificial genomic dataset – i.e. overlapped the simulated CLR scores with the ~19k annotated genes arranged linearly along the simulated genomic dataset, calculated the gene scores and determined outlier genes using our *q*-value thresholds, and finally merged all outliers into sweeps by combining outlier genes less than 1Mb apart.

For all four simulated populations, the distribution of *p*-values for simulated genes closely match the expectations of the standard Gaussian distribution and show a good correspondence to the empirical datasets – regardless of which demographic model or recombination rate was used (Figures 2C, 2D). For all populations and models, we then calculated FPR for each *q*-value threshold by dividing the number of sweeps detected in each simulated genomic dataset by the number of sweeps observed in the empirical data at the same *q*-value threshold. Notably, at the $q < 0.01$ threshold (which was used to determine the 57 candidate genes), false positive sweeps were only observed in the *Anatolian_EF* population, where the mean FPR was always <7%. To estimate the study wide false positive rate, we created a multiset of all outlier sweeps in which each of the 57 sweeps appeared at the same frequency that it was observed in the ancient populations. This resulted in a multiset with 99 instances of the candidate 57 sweeps, from which we then drew 100,000 binomial samples with p=0.07 (mimicking our most conservative FDR estimate). To estimate the study-wide FPR, we counted the number of unique sweeps for each of the 100,000 random samples and calculated the average

number of unique sweeps. This resulted ~6.3 false positive sweeps on average, providing a conservative FPR of 6.3/57 of ~11%.

4.2. Analytical pipeline power

To test the power of our approach to detect selective sweeps in European history, we augmented demographic models 2 and 3 by simulating positive selection under four different selection coefficients (*s* = 1%, 5%, 10%, and 20%; covering a similar range to empirical estimates from our candidate sweeps), and four distinct times of origin for the beneficial mutation (211.1kya, 105.6kya, 84.4kya, 63.3kya, 33.8kya; i.e. ranging from the approximate formation of Anatomically Modern Humans to the split between *EEF* and *WHG* populations ~38kya, Figure 4; (Kamm et al., 2019b)).

Selection was included in a simulation by adding the following MSMS parameter settings to a given demographic scenario:

```
-SFC -SI $StartTime 3 0 $StartingFrequency 0 -SA
$SelectionStrength -Sp 0.5
```

To model *de novo* mutations, the variable `$StartingFrequency` was set to one over twice the current population size (i.e. $1/2N_e$). We also simulated selection on rare standing variation – a scenario consistent with the large number of candidate sweeps observed in our study – by including starting frequencies of 0.1% at the onset of selection. The `-SFC` option conditions on the beneficial mutation being present at the time of sampling; i.e. simulations, where the beneficial allele is lost are discarded. Each selection scenario was repeated 200 times for each demographic model and simulated population. For all scenarios other than the 1% selection coefficient, the selected mutation became fixed or was close to fixation (> 99% frequency) at the time of sampling (Figures S6A, S6B) and selection was allowed to persist after an admixture event. All simulations were performed at with a fixed genome-wide recombination rate of 0.72 cM/Mb.

For both models, we then created 30 population genomic datasets that included selection, by randomly replacing 50 of the 5Mb regions from each of the 30 previously generated neutral datasets with 50 randomly sampled 5Mb selected regions. This was repeated separately for each combination of selection strength, starting time, and sweep mode, and gene scores were recalculated for all genes falling in selected regions. So that the selected genes did not impact the gene score standardization, we reused the standardization parameters previously estimated from the complete neutral genome. We then counted the proportion of the 50 selected regions that had at least one gene score exceeding a specific outlier threshold evaluated from the full neutral genome (calculated for 0.95, 0.99 and 0.999 gene score quantiles, approximating 5%, 1% and 0.1% false positive rates). Finally, the power was calculated by averaging the number of selected windows exceeding a given FPR threshold across the 30 simulated population genomic datasets.

The detection power for our analytical pipeline was consistently high for sweeps arising in the last 80kya, particularly for strong selection acting on *de novo* mutations (Figure 4, S7; see more discussion in Material and Methods section 4.3). Additionally, at the most stringent (0.1% FPR) threshold – which results in the closest match of the minimum *Z*-scores observed for outlier genes in the present study – the detection power declined more than 2-fold across all populations for sweeps arising at 34kya for simulations with the admixture event compared to simulations lacking admixture – even though the beneficial mutation is still selected at the same strength after the admixture event (Figure 4). Moreover, other than for the weakest selection coefficient (s = 1%), selection was strong enough to re-fix the beneficial allele in present-day European populations in several simulations (for instance, the expected fixation time for a beneficial variant starting at ~70% frequency is ~3,000 years, given a selection coefficient of 10%).

Taken together, our simulations imply that genetic drift likely had negligible impact on detection power for sweeps arising around the time of the diversification of western Eurasian populations, but that subsequent admixture greatly impacts power even if the

selection pressure persists after the admixture event. We note further reductions in sweep detection power are expected following additional admixture events in Eurasian history, and changes in lifestyle may resulted in relaxed the selection pressures in some cases (e.g. improved housing and clothing for cold adaptation), whereby our simulations likely underestimate the true impact of admixture on masking historical hard sweep signals in modern European genomes.

4.3. Impact of small sample size and missing data on detection power

Previous results suggest that SweepFinder2 maintains high power to detect true sweeps when the number of haplotypes being analyzed is 10 or more for modern genomes (Ferrer-Admetlla et al., 2014). To examine the impact of sample size on SweepFinder2 estimation on our ancient populations, we introduce a measure of sample size that incorporates both pseudo-haploidy and different levels of data missingness in our ancient samples, called the 'effective' sample size, n_{eff}. We calculated n_{eff} for each population as $k * n * (1 - M)$, where k is the ploidy of each sample, n is the number of samples and M is the average proportion of missing sites at informative SNPs in that population. Informative SNPs are those that are segregating in each population, as only such polymorphic SNPs were used to detect sweeps using SweepFinder2.

Consistent with previous results (Ferrer-Admetlla et al., 2014), SweepFinder2 has consistently high power to detect hard sweeps in simulated populations with reasonable sample sizes (i.e. $n_{eff} \geq 10$; Figure 4, S8A, S8B). For the two 'large' ancient Eurasian farming populations (i.e. $n_{eff} \geq 10$; *Anatolian_EF* and *WesternEurope_LNBA*), detection power was reasonable for sweeps with $s > 1\%$ (> 40%), and excellent for those with $s > 5\%$ (> 80%). The detection power continually decreases for sweeps starting more than 63 ky ago, dropping below 40% for sweeps > 200kya, in line with previous work that suggests that the power to detect sweeps begins to decay ~0.2 N_e generations after

fixation (Huber et al., 2015) when assuming an effective population size on the order of 10,000 (which is consistent with N_e estimates for European populations; Lohmueller et al., 2009). Sweep detection power was consistently poor for the *WesternEurope_LNBA* population (which had an n_{eff} ~4; Figure 4), and the simulated *p*-value distribution strongly deviates from the expected uniform distribution (Figures S2C, S2D) – a feature also observed amongst the ancient populations in this study with similarly low n_{eff} values (Figures 2A, 2B). Based on these results, we decided to only use ancient populations with $n_{eff} \geq 10$ in the present study.

Interestingly, the simulated *WHG* population displayed consistently lower power to detect sweeps arising from *de novo* mutations than the two large ancient Eurasian farming populations – being 20% lower on average in the *WHG* population compared to the two 'large' ancient farming populations (Figure S7). Notably, demographic parameters inferred by Kamm et al. and adopted in our simulations suggest that the *WHG* population experienced a prolonged and severely reduced N_e compared to Early European Farming populations following the Out-of-Africa bottleneck (EEF: N_e = 12k; *WHG*: N_e = 2k). The subsequent inflation in genetic drift caused by the lower N_e in *WHG* lineages appears to have led to more rapid erasure of fixed sweep signals and a subsequent reduction of detection power relative to the *Anatolian_EF* and EEF populations. This may also partly explain the lower number of candidate sweeps observed in *WHG* compared to Anatolian Farmer populations in the present study (Figure S8).

4.4. Impact of sample sizes on sweep detection in modern populations

Modern populations tended to have systematically fewer sweeps than ancient populations in this study, a result that did not change when altering the *q*-value threshold used to determine outlier genes that define the sweeps (Figures S3, S8). To test if smaller samples size were more likely to produce more false positives (for example, by inflating the number of sweep haplotypes in a partial sweep by chance), we

resampled one modern (*FIN*) and one ancient population with large sample numbers (*CentralEurope_LNBA*) at lower sample sizes (25, 10 and 5) and reran the complete detection pipeline to see if there was a systematic increase in the number of detected sweeps. We used an ancient population in addition to a modern population to take into account the impact that missing data and pseudo-haploidy might have on sweep detection. Notably, reduction in sample size did occasionally result in an increase in the number of detected sweeps for both ancient and modern populations; however, this effect was not systematic and in most cases the number of sweeps decreased in number – particularly at $n_{eff} < 10$ (Figure S8A, S8B). Further, the number of sweeps detected in the modern population was always lower than that detected for the ancient population at comparable effective sample sizes. Our results indicate that while there is some stochasticity in the number of reported sweeps, the overall pattern of decreased sweeps in modern populations is unlikely to be an artefact of larger sample sizes.

4.5. Impact of population sample composition on power to detect sweeps

To estimate the impact of our population assignment on the robustness of our candidate sweep detection, we reran the sweep detection pipeline on alternate sample groupings for both the *WHG* and *Steppe* populations and compared these to the original sample groupings. The original sample assignments for each population was governed by minimizing the temporal and spatial variability within each population while maintaining a homogeneous archaeological context. The alternate population groupings preserve the archaeological context but use a coarser geographical and temporal context and/or include samples with additional ancestry components not present in the original sample groupings (Table S1). Specifically, for the *WHG* population – which initially comprised 44 samples largely derived from the Balkan region – we added 12 samples mostly sourced from the Gravettian culture from western Europe (sourced from France, Germany, and Luxembourg) that were ~1kya older on average (mean sample ages: 10kya vs 9ya for original samples). The 75 samples from the *Steppe* population were supplemented by 15 samples of similar antiquity and provenance (Maykop Culture

samples; mean age ~4.5kya for both groupings), but which draw ~4% of their ancestry from Siberian hunter-gatherers (i.e. Eastern Hunter-Gatherers with Siberian genetic affinity) that is absent in the samples in the original grouping (Allentoft et al., 2015; Wang et al., 2019).

The two sample groupings had significantly correlated gene scores for each population (Pearson's r = 0.82 and 0.64 for WHG and Steppe, respectively; Figure S4B). Notably, whilst changing the sample composition of the *WHG* and *Steppe* populations resulted in some different sweeps being detected for each (Figure S4), 19 of the 27 sweeps (4/7 and 15/20 for *WHG*- and *Steppe*-specific sweeps, respectively) were observed regardless of which samples were used. Amongst these 19 retained sweeps, 7 were observed in a population other than the alternate *WHG* or *Steppe* population (e.g. sweep *4:71.5-72.4* is significant in only one of the two *Steppe* sample groupings, however it is also significant in *Anatolia_EF* and *CentralEurope_LNBA*; Figure S3, S4). Overall, 22 of the 57 candidate sweeps appear in two or more ancient Eurasian populations, which increases to 49 out of 57 candidate sweeps when sweeps are defined by containing at least one gene with a minimum *q*-value < 0.05 (instead of the *q* < 0.01). Our results imply that the shared genetic history and large number of ancient populations used in this study improve the chance that a sweep will be observed and thereby impart a degree of robustness to the sweep detection process.

4.6. Testing if sweeps were artefacts of low genome complexity

To test for potential false-positive sweep signals arising in genomic regions of low sequence complexity, we partitioned the genome into areas of high and low complexity based on CRG100 scores (Derrien et al., 2012) and checked if the proportion of mappable nucleotides were systematically reduced in the 57 candidate sweep regions. Mappability is a metric that indicates how readily reads can be aligned to a particular genomic region, which has a positive association with sequence complexity. Hence, consistent observation of a low proportion of mappable nucleotides relative to the genome-wide average would suggest that many sweeps were false positives due to

mapping artefacts. The CRG100 values were downloaded from the UCSC Genome Browser (Kent et al., 2002) and windows with a mean value less than 0.9 were defined as low complexity regions, and the remaining windows defined as high complexity regions. For each of the 57 candidate sweeps we calculated the proportion of ~1240k SNPs found within each sweep, that also lie within high complexity regions. Notably, the SNP probes were designed to target genome regions with high mappability (Lu et al., 2011), and our results confirm this with at least 96.7% of the SNPs in each sweep lying in high mappability regions (Table S6). Taken together, our results imply that our candidate sweeps were not enriched with artefacts caused by low genome complexity, and likely represent true historical selection patterns.

5. Inference of sweep dynamics

To learn more about the evolutionary dynamics of each of the 57 sweeps, we inferred the most likely selected haplotype and then traced the frequency changes of this haplotype across time using several different approaches. To reconstruct the sweep haplotype, we reasoned that the alleles carried on the selected haplotype would be at high frequencies in the ancient populations where the sweep was significant. Accordingly, for each sweep, we identified a subset of alleles that are strongly associated with the sweep, using the following two statistics:

$$\sum_i \quad (1\text{-}4pq) \times ln(CLR_i) \qquad (1)$$

$$ln[\sum_i \quad p \times ln(CLR_i) \,/\, \sum_i \quad q \times ln(CLR_i)] \qquad (2)$$

For each sweep, equation 1 was evaluated for every SNP lying in the sweep region to identify a subset of SNPs that act as reliable markers for the sweep. Equation 1 provides a weighted score for each SNP, which is the product of the log-transformed SweepFinder2 *CLR* score (i.e. the *CLR* score overlapping the SNP in question) and the standardised allele frequency, which is scaled to fall between 0 (when both alleles are at 50%) and 1 (when either allele is fixed). This value is calculated for each of the 18 ancient populations used in the study (indexed by *i*) and then summed together.

Accordingly, equation 1 returns large weighted scores for SNPs where one allele is consistently near fixation in populations where the sweep also exhibits evidence for selection.

Once a subset of marker SNPs were identified for each sweep, equation 2 was used to evaluate the allele that is most strongly associated with the selected haplotype. Following the logic of equation 1, we calculated the weighted score separately for each allele in a given SNP, where p and q denote the frequency of each allele in population i, then calculated the log transformed ratio of these two values. The sign of the resulting value indicates which of the two alleles is most likely linked with the haplotype (+ve = p; -ve = q), with larger magnitudes providing stronger evidence for association.

To test how the number of SNPs chosen impacted our analyses using the sweep haplotypes, all such analyses were performed using sets of 30, 50 and 100 marker SNPs for each of the 57 sweeps. For cases where a sweep had fewer SNPs than the number of marker SNPs used in the analysis, all of the SNPs found in the sweep were used. Notably, one of the sweeps (*10:122.9-123*) only contained six SNPs overall (the next lowest had 23 SNPs in total; see Figure S5A). SNP haplotype inference was reliable for this sweep and it was discarded from all subsequent haplotype frequency analyses.

The marker SNPs and associated weights are shown with respect to the distribution of CLR scores and gene scores across all candidate sweeps in supplementary data S1-S57 (https://figshare.com/s/20c0d1786a7ae8d3d45f).

5.1. Quantifying the impact of admixture on selected haplotype frequencies

For each sweep, we measured the change of the selected haplotype frequencies following the arrival of humans in Eurasia using two separate approaches. Both methods involve comparing the marker alleles on the selected haplotype for each sweep with alleles present in the samples withinin the OAGR database. In order make

these comparisons meaningful for all samples, a random allele was chosen at each SNP for all diploid samples (i.e. Ust'-Ishim and modern European samples). This approximates the sampling involved in generating the pseudo-haploid allele calls for the ancient samples. We caution that the random sampling of alleles means that our haplotype frequency estimates are likely to underestimate the true frequency of the selected haplotype. This is because samples that are heterozygous for the selected haplotype are more likely to be called as non-selected 'haplotypes', a probability that will likely increase with local levels of genome diversity and also the number of marker SNPs that are used. Accordingly, our estimated selected haplotype frequencies should be regarded as approximate lower bounds for the true frequency.

Our first approach to measure selected haplotype frequency changes for each sweep calculates the proportion of ancient and modern European individuals that carried the selected haplotype within successive time intervals covering Eurasian human history. To determine if a sample carried the selected haplotype, we required that at least 95% of the alleles were present in the sample, to allow for the possibility that some of the inferred alleles on the selected haplotype were incorrectly called. We also repeated these analyses using a less stringent threshold of 90%, to see how this impacted subsequent analyses. To improve the robustness of our estimates, all samples with less than 10 marker SNPs were not considered in the analysis of that given sweep.

After determining if the selected haplotype was present for each sample, we calculated the proportion of samples carrying the selected haplotype across seven consecutive time intervals (see Figures S5A-S5C). We chose the following intervals – 50-20kya [13], 20-10kya [43], 10-8kya [94], 8-6kya [272], 6-4kya [546], 4-2kya [183], and 2-0kya [305] – with the interval choice being largely motivated by the desire to have reasonable number of samples in each to improve our estimates (the number of samples falling in each interval are shown in square brackets). After estimating the frequency changes across successive time intervals, we examined if these selected haplotypes tend to decrease at specific times in history. Indeed, we observe that successive 2,000 year

periods starting at 10kya and continuing up to the present were marked by an average decrease in selected haplotype frequencies, with the last 6kya being particularly prominent (Figure S5G). This general pattern was consistent when performing this analysis with different numbers of SNP markers (i.e. 30, 50, and 100) and also using different thresholds for the proportion of marker alleles present in a sample (i.e. 90% and 95%; Figure S5G).

To improve our resolution for the timing of decrements in the selected haplotype frequency, we used a second approach based on regressing the proportion of selected haplotype alleles present in each sample (which we term $prop_h$) against the age of the sample, using the default LOESS function implemented in R (Team, 2012). In order to take missing SNP information for each sample into account, we weighted the contribution of each sample to the fit of the LOESS curve using the natural logarithm of the number of SNPs in a given sweep region carried by each sample. Further, all samples with less than 10 marker SNPs in a given sweep region were excluded from the analysis of that sweep.

After fitting the LOESS curve for each sweep, we determined when the predicted $prop_h$ value rose to its maximum, which approximates the time-point when the selected haplotype achieved its historical peak value. Next, we inferred the subsequent point in time when the predicted $prop_h$ value first dropped below this maximum, which we took as the time when the point estimate for $prop_h$ plus the error estimated at that point first dropped below the maximum $prop_h$ value. This value provides a rough estimate of when the selected haplotype was first appreciably impacted by admixture for each sweep (Figures S5A-S5C). The estimated times of first frequency decrease for each sweep were then binned into 1,000 year increments over the past 50,000 years and the resulting distribution was used to investigate if there were any particularly intense periods of admixture. The subsequent distributions were reasonably consistent for the different number of marker SNPs used (i.e. 30, 50, and 100), and the patterns closely matched our results from the binned haplotype frequency changes. We observe that the

selected haplotypes are more likely to show their first decrease at three sets of time-points that coincide with known periods of extensive admixture in European history (see Figures 2, S5F; Table S2): 1) around 8-9kya when Anatolian farmers spread into Europe, 2) following the westward migration of Steppe pastoralists around 5kya, and 3) from ~2.5kya to the present which marks the transition into the classical period which saw large scale human movements following the emergence of various large territorial empires, kingdoms, and nation-states.

5.2. Estimating the earliest evidence for selection

To estimate when the selection pressure(s) underlying each of the 57 sweeps may have first arisen, we searched all samples in the OAGR repository for the oldest sample that had the selected haplotype. Chronological information for the earliest period of human occupation of Eurasia are provided by five Late Pleistocene Eurasian human samples with moderate to high coverage genomes: Ust'-Ishim (Fu et al., 2014b), Kostenki 14 (Seguin-Orlando et al., 2014), GoyetQ116-1 (Fu et al., 2016), Věstonice (Fu et al., 2016), El Mirón (Fu et al., 2016). Each of these samples is drawn from a different time period in Eurasian prehistory, and detection of the sweep haplotype in a particular sample suggests that the selection pressure was present at the time the sample was deposited. Overall, the probability of observing the haplotype is proportional to its frequency in a population that the sample was drawn from, whereby it is unlikely that the sweep haplotype is absent from an individual from a population where it is at high frequencies, though this becomes more likely if the frequency of the haplotype is rare (i.e. it is not under selection, or only recently started coming under selection). Conversely, for selection acting on standing variation – which is likely to be prominent for the sweeps identified in the present study – there is some chance of detecting the sweep haplotype prior to it becoming beneficial that depends on the frequency of the haplotype when selection begins. This would lead to overestimating the age of origin of the selection pressure. These factors suggest that our dates should be regarded as an approximate upper (i.e. older) boundary of the onset of selection.

To test the robustness of this estimate, we used two different $prop_h$ thresholds (i.e. 0.90 and 0.95) and three different numbers of marker SNPs (i.e. 30, 50, and 100). Estimates of the timing of selection for each sweep for the different combinations of $prop_h$ values and number of marker SNPs are shown in Figures S5A-S5C (dates for the 100 marker SNP analyses are also shown in Table S2). Further, pseudo-haplotype plots for all 57 candidate sweeps are provided in supplementary data S58-S114 (https://adelaide.figshare.com/s/f00443ef535a8af2a61d). Each image depicts the allele found at each SNP in a given sweep window plus the immediate flanking regions for all populations and the five Late Pleistocene samples used in this study. Individuals with more than 50% missing data in this window are omitted to help improve visualisation of the haplotype frequencies.

5.3. Estimating selection strength and sweep size

After classifying sweeps by likely starting time according to the earliest evidence of the sweep haplotype amongst the five Late Pleistocene individuals (Figure 2; Table S2), we tested if there were any differences in the strength of selection, the size of the sweep, or the number of genes contained in a sweep according to their starting time. We inferred the selection strength, s, for each of our 57 candidate sweeps using $s = r \ln(2N)/\alpha$ (Durrett and Schweinsberg, 2004), where N is the population size, r is the recombination rate, and α is a composite selection parameter estimated for each sweep region in SweepFinder2 (Nielsen et al., 2005). For each candidate sweep, we took α at the position of the largest CLR value from all ancient Eurasian populations. Recombination rates were estimated for each sweep using information from (Bhérer et al., 2017), and assumed a population size of 10,000 (noting that the estimation of s is robust to changes in population size due to the logarithmic transformation). Notably, estimating s based on SweepFinder2 α estimates tended to result in slightly lower values than expected (Figure S6C, S6D), suggesting that we have likely systematically underestimated the strength of selection for the candidate sweeps.

Using the five Late Pleistocene individuals to time the candidate sweeps suggests that the majority of candidate sweep haplotypes were at high frequencies by 35kya, suggesting frequent episodes of surprisingly strong selection following the arrival of Anatomically Modern Humans in Eurasia. Our estimates of *s* confirm that selection was strong, with 53 sweeps having $s > 0.5\%$, 41 having $s > 1\%$, with the largest value of *s* nearing 10%. The selection coefficients we observe are strong enough to drive new or initially rare mutations to a detectable frequency (i.e. > 90%) in only 3-4 thousand years for the strongest selection coefficients observed in our data, and within 10-20 thousand years for sweeps with the weakest selection coefficients (Figure S6A, S6B). Importantly, these estimates are likely lower than their true *s* values – the SFS-based approach of SweepFinder2 tends to underestimate *s* by about 20-30% (Vy and Kim, 2015), and selection from standing variation can lead to substantially narrower sweep regions also resulting in underestimation of *s*. Indeed, when comparing the inferred selection coefficients against the simulated values for the *WHG* and *Anatolian_EF* populations, *s* was consistently underestimated across the three different demographies and selection regimes, particularly when acting upon more recent standing variation (Figures S6A, S6B). Thus, our data are consistent with the observation of multiple sweep haplotypes at high frequencies within 20,000 years of the OOA dispersal.

We also estimated the size of the sweep using $d = -ln(1 - P_e)/\alpha$, where P_e is the probability of escaping the selective sweep at distance *d* (Nielsen et al., 2005). We assumed a probability of escape of 0.75 for all sweeps to compute the sweep sizes (Figure S5E). We did not observe any significant difference in sweep size or estimated selection coefficient across these age groups (ANOVA, p=0.9528 & p=0.6803, respectively), nor did the number of genes within each sweep region or subsequent gene density vary by age group (Kruskal-Wallis rank-sum test, p=0.6909 & ANOVA, p=0.4023, respectively). Thus, we do not observe any systematic change of selection pressure over time that would lead to different sweep sizes, nor do we observe any

change in the properties of the regions that contain the selected mutation (i.e. low gene density vs. high gene density regions).

5.4. Testing if sweep timing is associated with reported evidence of positive selection

Assuming that sweeps occurring more recently in time would also be more likely to be limited to a subset of ancient Eurasian populations, we tested if any of the sweeps that had been previously reported as regions showing evidence of positive selection were more likely to be older on average than sweeps that had not. For this analysis, we chose three studies with well defined and robust outlier detection methods that use different methods to search for evidence of positive selection: a study using a combination statistics using haplotype and site frequency spectra (Pickrell et al., 2009), a study testing allele frequency differentiation across different branches of the Eurasian ancestral tree (Racimo, 2016), and a recent machine-learning-based method (Schrider and Kern, 2017). Importantly, all studies were performed using the same 1000 genomes dataset (*CEU*) that was used as one of the modern European populations in the present study. For each of the three studies, we used simple permutation tests to examine if the age of the sweeps showing evidence selection are older than those without evidence for selection (permuting inferred sweep ages). We also tested if the dates are older for sweeps appearing in at least one of the three studies against those that did not. This was repeated for both of the 90% and 95% $prop_h$ thresholds for the top 100 SNP markers (using lower numbers of SNP markers resulted in consistently older ages (Figures S5A-S5C), suggesting that they were too permissive and limiting the power of these analyses, and are not considered here). We observe that sweeps with evidence for positive selection from one or more of these three studies are consistently significantly older than those that are not (Figure S5H). Moreover, this pattern becomes more apparent when partitioning the sweeps by the number of studies in which the sweep has been reported, with the inferred age increasing according to the number of studies in which the sweep was observed (Figure 2).

5.5. Sweep sharing across populations

We quantified the extent to which sweeps were shared across different populations by performing principal coordinate analysis (PCoA) on the set of 57 candidate sweeps. The PCoA analyses used a distance matrix that measures the amount of signal sharing by computing the Jaccard similarity coefficient between all pairs of populations. The resulting similarity coefficient, *S*, is defined as the number of significant sweep regions that overlap between two populations, divided by the total number of significant sweep regions that are significant in at least one of the two populations. A distance measure for each pair of populations is then defined as 1-*S*, whereby a distance of zero means that all the sweep signals in one of the populations overlap with signals in the other population (and vice versa) and a distance of one means that none of the signals overlap. We excluded sweeps that are entirely private to only one population since these signals are not informative about sweep sharing. The PCoA analysis was conducted using the *pcoa* function from the ape package in R (Paradis and Schliep, 2019).

To compare the clustering of candidate sweeps with patterns of population genetic differentiation, we performed a PCA analysis of the SNP data for each sample in the 21 ancient and modern Eurasian populations. We used Plink v1.9 (Chang et al., 2015) to prune the SNP set by removing closely linked variants (settings: --indep-pairwise 500 kb 5 0.5), and converted the resulting plink file to Eigenstrat *PACKEDANCESTRYMAP* format (Patterson et al., 2006; Price et al., 2006). The PCA was performed using the *smartpca* function with the *noxdata, inbreed, numoutlier,* and *autoshrink* options all enabled. The *CEU*, *TSI*, and *FIN* modern populations were used for the projection, with Eigenvectors defined. To visualize the population sweep and genetic relationships, the first two principal coordinates from the PCoA and the first two principal components of the PCA are plotted in Figure S1B-S1C.

The PCoA and PCA results show very similar clustering patterns, suggesting that the sweeps are more likely to be shared if the populations have close genetic relationships. In particular, we see three basic clusters forming amongst the ancient populations: 1) the two Hunter-Gatherer populations, 2) Early and Late Farmers (including *Anatolia_EF*) and 3) the *Steppe* and Late Neolithic/Bronze Age (*LNBA*) populations. Similar clustering patterns also appear when the sweeps are determined using less stringent *q*-value cutoffs (Figure S1D), suggesting that this is a robust signal. Further, the consistent population clustering observed between the PCoA and PCA remains when rerunning the latter after removing all SNPs contained in the 57 candidate sweeps (Figure S1E-S1G), indicating that this pattern is independent of the sweeps. The clustering of Early and Late Farmers with *Anatolia_EF,* and *Steppe* with *LNBA* populations, both closely match previously described major colonization and genetic mixing events in European prehistory – i.e. the movement of Anatolian farmers ~8.5kya and Steppe pastoralists around 3kya (Skoglund and Mathieson, 2018) – and show that the resulting admixed European populations take a large proportion of their genetic ancestry, and their sweep signals, from the newly colonizing group. Further, the *Caucasus_LNBA* sit between the *Anatolia_EF/*Early Farmer/Late Farmer and *Steppe*/*LNBA* clusters, which is consistent with their likely descent from Caucasian Hunter-Gatherers (which are also ancestors of Steppe pastoralists) and the Late Farmer groups. Finally, the *CEU* populations also sit closer to the *Steppe* and *LNBA* populations, reflecting a strong genetic ancestry from this group. In contrast, both *TSI* and *FIN* form outliers in the PCA plot since they have very few of the 57 sweeps overall, whereby the PCoA results shown in Figures S1C and S1D excluded these two populations. Notably, both *FIN* and *TSI* populations have been shown to have ancestry from non-European groups that is not found in *CEU* individuals, which have a mixture of Northern and Western European ancestry (e.g. Near Eastern gene flow into ancestral Tuscans around 3kya (Fiorito et al., 2016; Pardo-Seco et al., 2014); and Siberian gene flow into ancestral Finns around ~3.5kya (Lamnidis et al., 2018)). Thus, our results suggest that many of the sweep signals that have been preserved in *CEU* individuals have been eroded in *TSI* and *FIN* populations via these

additional pulses of admixture from populations where the sweep haplotype was uncommon or absent.

5.6. GWAS enrichment in selected genes

Information from the genome wide association study (GWAS) catalogue (EMBL-EBI, https://www.ebi.ac.uk/gwas/; (Buniello et al., 2019)) was used to test for potential associations between putatively selected genes and modern human diseases (Table S5). We created five different sets of genes by binning those with *q*-values less than 0.01, 0.05, 0.1, 0.2, and 0.5 in at least one population. For these five gene sets, we tested if each was enriched for genes from GWAS catalogue relative to the remaining genes using a one-sided Fisher's Exact Test. Gene sets with $q < 0.01$ and $q < 0.05$ had significantly more GWAS genes than expected (p = 8e-6 and 0.016, respectively, Table S5), although only the $q < 0.01$ gene set remained marginally insignificant ($p = 0.077$) when excluding genes found in the MHC region (Table S5).

6. F_{st}-based selection tests

To further investigate the validity of the 57 candidate sweeps, we tested if our sweeps were enriched with highly divergent SNPs amongst the 12 ancient Eurasian populations (i.e. those with $n_{eff} \geq 10$) and a modern African population (i.e. *YRI*). We calculated F_{st} for each of the ~1.1M ascertained SNPs using the Weir-Cockerham estimator (Weir and Cockerham, 1984), then used OutFLANK to estimate the probability that each SNP was more divergent than expected under neutrality (based on fitting a χ^2 distribution to F_{st} values from putatively neutral SNPs (Whitlock and Lotterhos, 2015)). Importantly, OutFLANK is robust to non-equilibrium demographic models (Whitlock and Lotterhos, 2015), including rapid range expansions that are thought to have resulted in highly divergent alleles observed in modern human populations (i.e. through allele surfing; (Excoffier and Ray, 2008)). Applying the *q*-value correction to the *p*-values to control for multiple testing resulted in 29 of the 57 candidate sweeps having one or more SNPs with a q-value < 0.05, and 55 candidate sweeps having at least one SNP with a q-value < 0.20 (Figure S2F). We then tested if the SNPs found in each of the 57 candidate sweeps had significantly higher F_{st} values than background genome levels. 49 of the 57 sweeps had significantly elevated F_{st} values relative to the remaining background genome ($p < 0.05$; Wilcoxon Rank Sum Test; Figure S2F), confirming that our sweeps were unlikely to be an artefact caused by non-equilibrium demographic processes such as allele surfing.

7. Pathway enrichment analysis

To determine if positively selected genes were enriched within particular biological pathways in historical or modern populations, we developed a novel enrichment method, PolyLink (https://github.com/ACAD-UofA/PolyLink), and applied this to the log_{10}-transformed, length-corrected gene scores. Unlike other gene set and pathway enrichment methods that require a predefined set of outlier genes – and therefore are contingent on an arbitrary threshold to determine outlier genes – our method follows the approach of a recent enrichment test, PolySel (Daub et al., 2013), by using the scores from all genes in quantifying evidence for pathway enrichment. The standard PolySel algorithm contrasts the sum of all genes scores within a given pathway to a null distribution that is generated by repeatedly summing the scores from the same number genes randomly sampled across the genome, with scores above a pre-specified threshold value taken as evidence for pathway enrichment. However, this method does not account for genomic linkage between genes within the same pathway, which could result in false positives; for instance, pathway enrichment may occur in situations when selection has targeted a single gene located amongst a cluster of genes belonging to the same pathway, resulting in the neighbouring genes also having larger scores despite not being under direct selection themselves. Hence, to generate a null distribution that explicitly accounts for genomic linkage we use the following modified randomization procedure:

1. Gene scores are arranged linearly in each chromosome according to their chromosomal position, and chromosomes are randomly joined together (preserving 5' to 3' orientation). Note that the gene-poor Y chromosome is not used here (nor in any other test reported in this study). Specifically, the random joining of chromosomes is achieved by permuting the set of chromosome numbers (1 to 22) and then connecting the chromosomes according to this permutation.

2. Despite the large number of possible combinations of chromosomes – there are *22!*, or greater than 10^{21}, unique combinations of chromosomes – a large number of combinations preserve the assignment of gene scores to the same genes (e.g. wherever the chromosome permutation has at least one chromosome maintained in its original numerical position). To maximize the chance that unique gene assignments are achieved during each iteration of the randomization procedure, we perform a further 'rotation' of the gene scores subsequent to joining the chromosomes in step 1. Assuming that there are N genes, we randomly sample a number, x, between 1 and N. The gene score initially in position x is re-assigned to position 1, the score initially in position $x+1$ is shifted to position 2, and so on until the score initially in position N takes position $N-x+1$; similarly the gene score initially in position 1 is now re-assigned to position $N-x+2$, the score initially in position 2 goes to position $N-x+3$ and so on until the score initially in position $x-1$ takes position N.

This two-step randomization procedure can generate up to $22! \times N$ ($>10^{25}$ in the present case) unique mappings of gene scores to genes. In contrast to other algorithms that sample statistics at random from the genome, our null estimation procedure preserves the physical linkage inherent in each population and explicitly accounts for the genomic clustering specific to each pathway (i.e. by computing a null for each pathway rather than a single null for all pathways with the same number of genes). The trade-off is that this results in a considerable reduction in the speed of computation for the test, though optimization of the new algorithm was not a specific goal of the current study.

7.1. Application of enrichment test to annotated pathways / gene sets

To generate a comprehensive set of human biological pathways, we combined annotated pathways from four different widely-used databases, BIOCYC (Romero et al., 2005), KEGG (Kanehisa and Goto, 2000), PID (Schaefer et al., 2009), and REACTOME (Fabregat et al., 2018; Schaefer et al., 2009), resulting in 1,859 unique pathways in total. To minimize redundancy amongst these pathways, we combined all pathways that

shared more than 95% of the same genes, resulting in a final set of 1,681 tested pathways. All references to combined pathways are highlighted in the text with an asterisk. We applied PolyLink to each population, repeating the randomization algorithm 1,000,000 times to generate an empirical null distribution for each of the 1,681 pathways, from which *p*-values were calculated. Following estimation of *p*-values for each pathway, we applied a *q*-value correction to control for multiple testing in each population (Storey, 2003), and report pathways that have a *q*-value < 0.05 in at least one of the 12 ancient populations with $n_{eff} \geq 10$ or one of the three modern European populations. This resulted in a total of 7 enriched pathways (Figure 5).

Amongst the seven enriched pathways detected in the PolyLink analyses were a subset of genes that were shared across several enriched pathways (Figure S9A). Notably, after removing genes that were shared across multiple pathways, the distribution of *Z* scores for the remaining genes that were unique to one of the seven pathways still exhibited an inflation of large *Z* scores relative to null expectations (Figure S9B). This result suggests that pathway enrichment signal was at least partially driven by genes that are unique to particular pathways.

7.2. Testing PolyLink performance

We assessed the efficacy of the PolyLink permutation algorithm to control for false positives due to the genomic clustering of genes in specific pathways, which is the major difference from similar pathway enrichment methods, *e.g.* PolySel (Daub et al., 2013). Specifically, we quantified the level of gene clustering in each pathway using a simple ratio $(N_{genes} - N_{fictive}) / N_{genes}$, where the N_{genes} is the total number of genes in a pathway and $N_{fictive}$ is the number of 'fictive' genes in a pathway after accounting for clustering, and call this metric the pathway clustering index (*PCI*). $N_{fictive}$ explicitly quantifies the reduction in the number of genes in a pathway caused by clustering. This is achieved by first ordering all genes in each pathway by their chromosomal position. For the first gene in each pathway, all neighbouring genes situated within 1 cM of this

gene are combined with the focal gene into a single fictive gene, and this process is repeated for all successive downstream genes, resulting in $N_{fictive} \leq N_{genes}$.

Consequently, for a given pathway, $N_{fictive}$ obtains a minimum value of 1 when all genes in the pathway lie within 1 cM of the first gene (based on distance between gene midpoints) and a maximum value of N_{genes} when no gene lies with 1 cM of any other gene in the pathway. Accordingly, pathways with no clustering have a *PCI* of 0, and those with successively more clustering approach a maximum *PCI* value of ~1.

Notably, the PolyLink randomisation procedure produces highly similar *p*-values to a standard algorithm permuting gene scores randomly across the genome, particularly when gene clustering is low ($0 \leq PCI < 0.1$; Figure S9A). However, the standard randomisation algorithm tends to result in increasingly more outliers than the PolyLink procedure for increasing *PCI* values, consistent with an elevated FPR in the former due to gene clustering in pathways. The fact that the two randomisation approaches have a 1:1 correspondence in their *p*-values for low *PCI* levels indicates that PolyLink procedure is effectively equivalent to randomly permuting gene scores when gene clustering is negligible. As clustering levels increase, however, the PolyLink procedure imposes an increasingly hefty penalty to large pathway scores, limiting the potential for false positives due to gene clustering.

8. Inferring putative selection pressures

Many of the sweeps in the Palaeolithic period carry genes that suggest humans faced a host of new metabolic and physiological challenges, caused by sustained low temperatures which were encountered following the peopling of Eurasia. Two of the single function sweeps that arose by 35kya contain genes are both involved with fat metabolism, which plays an essential role in thermogenesis (*WWOX* and *CAND1*). Functional evidence suggests that *WWOX* is directly involved in metabolising brown fat, high-density lipoprotein (HDL), and glycogen (Chang, 2016; Teng et al., 2013), whilst

CAND1 has been previously shown to regulate lipid droplet formation and preadipocyte differentiation (Dubiel et al., 2015). Experiments have demonstrated that brown fat metabolism and adipogenesis are increased under cold conditions (van den Berg et al., 2017; Bourque et al., 2008).

Previous studies have also shown that cold environments can lead to neuropathy through a number of different mechanisms, some of which involve genes identified in our sweeps. For example, *in vivo* experiments in amphibians and mammals show that hypothermic conditions trigger excessive electrochemical signaling in spinal cord neurons, leading to imbalances in neuronal ionic composition and cell death (Daló et al., 2007; Li et al., 2015). *NFASC* and *DOK5* may have preventative roles in this context: *NFASC* produces an adhesion molecule required to maintain tight junctions with neighbouring cells that form a barrier, whilst *DOK5* regulates neuronal differentiation. These processes are required for protection against and repair from hypothermic-related nerve damage, respectively (Ango et al., 2004; Goudriaan et al., 2014; Grimm et al., 2001; Spencer et al., 2019). Similar protective effects may also be conferred by *TMEM100* and *MPP6,* which are needed to maintain interactions between temperature-dependent ion channels (such as *TRPA1* and *TRPV1,* which together influence sensitivity to heat and cold) and facilitate myelination plus cytoskeletal maintenance of neuronal cells, respectively (Chung and Wang, 2011; Yu et al., 2019).

Taken together, our overall interpretation of these results suggests that climatic factors encountered by humans during the Upper Palaeolithic triggered a range of adaptive responses for core metabolic and neuronal functions. However, we emphasise that these interpretations remain speculative and further functional studies are needed to validate our findings.

8.1. Palaeolithic temperature gradient

To investigate the magnitude of the temperature gradient Palaeolithic populations would have experienced as they migrated into Eurasia during Stage 3 – the period between

60kya and 27kya during the last glacial cycle – we interrogated a number of proxy-based and climate modelling datasets (Brandefelt et al., 2011; Huntley et al., 2013; Kjellström et al., 2010; Van Meerbeeck et al., 2011; Voelker, 2002). Most relevant here are fully coupled atmosphere-ocean global climate simulations using the Community Climate System Model (or CCSM3) incorporating land-surface model and sea-ice components, which have been used to generate global and regional reconstructions (Brandefelt et al., 2011; Kjellström et al., 2010). Here, globally mean annual mean surface temperatures have been found to be 5˚C cooler during Stage 3 (centered on 44kya) compared to present-day but warmer than the Last Glacial Maximum, consistent with other modelling (Van Meerbeeck et al., 2011) and proxy (Voelker, 2002) studies. Mean annual mean surface temperatures (January-December) were compared over north Africa (15-30˚N) and central Europe (40-50˚N) and found to be ~5˚C cooler in the north during Stage 3. Similar temperature differences were experienced during the boreal summer (July-September) and winter (January-March). These temperature estimates, however, should be considered conservative, with a larger gradient likely between 55-50kya during the time Anatomically Modern Humans migrated out of Africa (O'Connell et al., 2018; Timmermann and Friedrich, 2016).

8.2. Adaptation of proteins found the cellular membrane

The cellular membrane is the location of many proteins targeted by pharmacological treatments, whereby we tested if there is an overrepresentation of selected genes coding for proteins located in this region. We determined the cellular location of each protein using information downloaded from the Human Cell Atlas Database (Thul et al., 2017) (HCAD; dataset: https://www.proteinatlas.org/search/has_protein_data_in:cell). After assigning out a gene to one of these categories, we performed PolySel analyses on the 31 named pathways that contained at least 10 genes. Amongst the three cellular membrane based categories – *Plasma membrane*, *Cell junctions*, and *Focal adhesion* – the first two were enriched consistently ($p < 0.05$) across multiple populations (Figure S10). Further, when ranking the cell location categories according to which had the

highest -log_{10} transformed *p*-values summed across all populations, the top-ranked category was Cell junctions, with Plasma membrane ranked 4th. Surprisingly, the Focal adhesion category defined in HCAD did not show any enrichment across populations; however, the Focal adhesion pathway defined in KEGG (Kanehisa and Goto, 2000) was amongst the outliers in our pathway enrichment analysis. Only a handful of genes were shared between the two pathways, suggesting that this discrepancy was due to differences in the annotation process between the two databases.

SUPPLEMENTARY FIGURE TITLES AND LEGENDS

Figure S1. Selective sweeps mirror western Eurasian population history. (A) Full geographic distribution of ancient Eurasian samples, including the far eastern Eurasian samples omitted from Figure 1, with different populations (indicated by colored symbols) classified into broader groupings according to archaeological records of material culture and lifestyle. (B) PCA analysis showing the genetic relationships between the different populations used in this study; (C) PCoA analysis depicting how the candidate sweeps ($q < 0.05$) are shared across the different ancient populations and one modern population (*CEU*); (D) PCoA analysis depicting how the candidate sweeps ($q < 0.1$), the close match between the genetic affinities and sweep clusters indicate that admixed populations (i.e. Early Farmers: *EF,* Late Farmers: *LF*, Late Neolithic/Bronze Age: *LNBA*, and Modern Europeans: *CEU*) retain sweeps reflecting their population ancestry. Populations that lack sufficient data to reliably detect sweeps (i.e. those with $n_{eff} < 10$) are not shown. The results are highly consistent with patterns observed for the 57 candidate sweeps (panel C), indicating that the sweep sharing patterns are robust to *q*-value threshold. (E) PCA using the full autosomal genomic data (~1.1M SNPs), using modern European samples to define the PCA space and projecting ancient samples after accounting for PCA shrinkage; (F) PCA using all SNPs other than those SNPs that overlap any sweep region; (G) The two PCAs with and without sweep SNPs projected onto the same plane defined by the first two PCs, with arrows signifying the spatial shift of each sample due to the removal of the sweep regions. The consistency between the two PCAs indicates that population patterns are independent of the candidate sweep signals.

Figure S2. Sweep detection pipeline quality controls. (A) Distribution of gene scores across all populations. The standardisation process results in gene scores that are approximately standard normal distributed (*i.e.* Z scores, expected distribution = red line), though deviations become more evident as the effective sample size, n_{eff} (a measure of sample size that accounts for pseudo-haploidy and missing data; see

Material and Methods), declines below 10 (n_{eff} values shown in panel header in square brackets). (B) Distribution of *p*-values for each gene, across all populations (see Material and Methods). The observed J-shaped distribution is consistent with populations containing a mixture of unselected genes, which have a standard uniform distribution, and selected genes which cluster at low *p*-values. Deviations from expectations become more notable as the effective sample size, n_{eff}, declines below 10 (see panel labels). (C) Quantile-Quantile plots of simulated gene scores vs. expectations from a standard normal distribution for four different populations (panels) and three different demographic scenarios (coloured lines, see panel key; demographic models in Figure 3). (D) Same as panel C, but showing results for simulated genetic datasets at three different recombination rates. Echoing our results in panels A and B, the simulation population with low n_{eff} (i.e. *WesternEurope_LNBA*) had drastically reduced detection power. Importantly, the simulated gene score distributions closely fit theoretical expectations indicating that our analytical pipeline is robust to different demographies, recombination rates, and other potential biases introduced by ancient genomic data properties. Power was also lower for *WHG* despite having a reasonably large n_{eff}, potentially as a result of the lower effective population size and elevated drift that has been inferred for *WHG* populations relative to Early Farmers. (E) The frequency of sweeps across populations according to different *q*-value thresholds. Sweeps were defined by grouping together genes within 250kb, 500kb, or 1Mb (i.e. join width) of each other that fell under a specific *q*-value threshold (see key). The 'x' indicates how many of the sweeps defined at $q < 0.10$ and join width of 1Mb overlapped with the 57 candidate sweeps. Populations are split into two groups based on whether they had sufficient data to make robust inferences ($n_{eff} \geq 10$ = 'Large') or not ($n_{eff} < 10$ = 'Small') (see Material and Methods). (F) Distribution of F_{st} scores – evaluated for the 13 ancient western Eurasian populations (those with $n_{eff} \geq 10$) and a modern African population (*YRI*) – at each SNP within the 57 candidate sweeps (red or grey violins) vs. the remaining genomic background (blue violin). More than half (29/57) of the candidate

sweeps contained significantly divergent SNPs ($q < 0.05$, large black dots; see Material and Methods), and nearly all (55/57) contained strongly divergent SNPs ($q < 0.20$, small black dots). Further, 49 of the 57 sweeps had significantly higher F_{st} scores than the remaining genomic background (Wilcoxon Rank Sum Test $p < 0.05$; significant = red violins; not significant = grey violins), consistent with elevated divergence in the candidate sweep regions resulting from positive selection.

Figure S3. The *Z*-scores for all outlier genes in candidate sweeps. Gene scores observed amongst the 57 candidate sweeps (y-axis), for all modern and ancient populations (x-axis). Populations are organized chronologically according to broad archaeological categorizations, with populations that had insufficient data for robust inferences labelled as 'Small' (i.e. $n_{eff} < 10$). Sweeps and genes are organized by chromosomal position. Outlier genes are indicated with a specific symbol according to their q-value ($0.05 \leq q < 0.10$ = *; $0.01 \leq q < 0.05$ = **; $q < 0.01$ = !).

Figure S4. Candidates sweep robustness under different population sample groupings. (A) The *Z*-score for each gene amongst the 27 sweeps (y-axis) observed for *WHG* and *Steppe* populations, using two different sample groupings for each population (see Table S1 and Material and Methods). Sweeps and genes are organized by chromosomal position. Outlier genes are indicated with a specific symbol according to their q-value ($0.05 \leq q < 0.10$ = *; $0.01 \leq q < 0.05$ = **; $q < 0.01$ = !). Importantly, while changing the sample grouping of the *WHG* and *Steppe* populations resulted in some different sweeps being detected for each, 19 of the 27 sweeps were observed regardless of which sample grouping was used – 12 were observed in both sample groupings and the remaining 7 were observed in one or more population other than the *WHG* or *Steppe*. (B) Distribution of gene scores for two alternate sample groupings for the *WHG* and *Steppe* populations (see Table S1 and Material and Methods). The gene scores are strongly correlated (Pearson's $r = 0.82$ and 0.64 for *WHG* and *Steppe*, respectively), though outliers ($q < 0.01$) differ between the alternate sample groupings.

Nonetheless, the shared ancestry amongst the ancient western Eurasian populations means that a similar set of outlier sweeps were observed using either grouping, as the sweep signals tend to be observed across multiple ancient populations (see Figures S5, S6, S24). Our results imply that the shared genetic history and large number of ancient populations used in this study improve the chance that a sweep will be observed and thereby impart a degree of robustness to the sweep detection process.

Figure S5. Candidate sweep history and properties. (A) Scatter plot showing the proportion of marker alleles characterising each sweep (separate panels) that are carried by each of the ancient samples (dots). Samples are coloured by the sweep significance in their designated population (see key; 'Other' indicates that the sample was not one of the 18 ancient populations, '<50%' means that the sample had less than 50% of the marker SNPs and was subsequently excluded from the further analyses of that sweep). African (*YRI*) and East Asian (*CHB*) populations are included for comparison. For each sweep, the proportion of selected alleles present in each sample was regressed against sample age (grey line, shading denotes the estimation error). The vertical grey dashed line indicates the time that the fitted regression curve decreases below the maximum proportion – approximating when admixture first impacts the selected haplotype. The average frequency of the haplotype is also plotted across successive time intervals (green dots and lines; see Material and Methods). The selected haplotype was considered to be present if a sample had at least 90% (light green) or 95% (dark green) of the 100 marker SNPs characterizing the selected haplotype (number of marker SNPs in each sweep shown in square brackets in the panel label). The green dashed line indicates the age of the oldest sample observed to have 95% of the sweep haplotype alleles, which approximates when the sweep began. The shaded intervals indicate the approximate onset of three different time periods of intensive admixture in European history (blue = Anatolian farmer migration, gray = Steppe migration, red = Classical era). (B) Identical to panel A, but now using 50 or (C) 30 marker alleles to characterize the selected haplotype for each sweep. (D) The number of sweeps that are shared between the three major historical groups that

contribute ancestry to modern Europeans (i.e. *WHG*, *Anatolia_EF*, and *Steppe*), organised by the inferred origin of the selection pressure. Dates were determined by observation of the oldest sample having at least 90% (top panel) or 95% (bottom panel) of (up to) 100 marker SNPs carried by the selected haplotype. Most dates first occur in one of five moderate to high coverage genomes dating to the early Paleolitic (see *Material and Methods*). No sweeps are observed in all three of the ancestral populations after 35ka. Sweeps occuring in all three populations suggest that the underlying selection pressure may have arisen in the common ancestor(s) for all three populations. Accordingly, sweeps occurring prior to 35ka are described throughout the text as 'Ancestral Western Eurasian' sweeps. (E) The relationship between the properties of a sweep and the earliest evidence for selection in that region from haplotype analysis of five Late Pleistocene individuals (see Figure 2). The size and strength of selection for each sweep region was determined using the compound selection parameter α inferred by SweepFinder2 (see Material and Methods). Neither (a) the size of the sweep, (b) the number of significant genes in the sweep region, (c) the gene density, nor (d) the selection coefficient, change significantly with the earliest evidence of selection (all $p > 0.05$). (F) Same as Figure 2B, but results are shown for the time of decrease estimated using three different marker SNP counts – 30, 50, and 100; see panels A to C. (G) Same as Figure 2C, but results are shown for dates derived from different combinations marker SNP count (30, 50, 100) and proportion of inferred selected alleles present (>90% and >95%). (H) Same as 2D, however the sweep presence is shown for all three studies (0 = not reported in any study, 1 = reported in one study, >2 = reported in two or more studies) and also each individual study separately (0 = sweep not significant, 1 = sweep significant) and for 90% and 95% sweep haplotype affinity (see Material and Methods).

Figure S6. Estimating sweep strength and fixation time properties. (A) Time to fixation for a beneficial *de novo* mutation for 1000 forward-in-time simulations at each of four different selection coefficients, *s*, ranging between 1% and 10% (see Material and Methods). The beneficial *de novo* mutation arises at time zero and the simulations are

conditioned on the fixation of this allele. We assume a generation time of 30 years and a single randomly mating population of 10,000 diploid individuals. The red dashed line indicates the expected time to fixation for a beneficial *de novo* using an analytical formula. (B) Same a panel A, but the beneficial allele starts at a frequency of 0.1% at time zero. (C) Estimation of the selection coefficient, *s*, for simulated sweeps from *de novo* mutations and (D) rare standing variation. Simulations were performed with two different selection start times (see key). The simulated *s* is shown as a triangle and the inferred values depicted as coloured circles (see Material and Methods). Estimated *s* values behave reasonably well for simulated *s* = 1%, though tend to be systematically downwardly biased for *s* values of 5% and larger – particularly for recently selected standing variants – suggesting that the selection coefficients reported for most of the 57 candidate sweeps are also probably underestimated.

Figure S7. Inferring sweep detection pipeline statistical power. Same as Figure 4, but showing results for selection acting on (A) *de novo* and (B) rare beneficial variants initially at 0.1% frequency and now including power estimates for Western Hunter Gatherers (*WHG*) for a demographic model without any admixture events (see Material and Methods). Power tends to be lower when selected acts on standing variation (though not at 1% starting frequency, where the selected allele is not always fixed at the time of sampling), and is systematically decreased for the for simulations that include a 30% admixture pulse 4kya before sampling relative to those lacking this admixture event (i.e. models M2 and M3 in Figure 3B, respectively).

Figure S8. Effect of small sample size and missing data on sweep detection. (A) The frequency of sweeps across populations relative to the effective sample size for each population. Horizontal panels show different *q*-values used to determine the outlier sweeps. Notable decreases in the number of observed sweeps occur for small populations (i.e. $n_{eff} < 10$; gray dots) and also modern European populations (blue dots). (B) The frequency of sweeps relative to effective sample size, n_{eff}, across an ancient (red dots) and a modern (blue dots) population resampled to smaller effective sample

sizes (the number of samples per population, *n*, is indicated in labels next to each point, with the resulting n_{eff} on x-axis). Horizontal panels show different *q*-values used to determine the outlier sweeps. There is a general tendency for the number of detected sweeps to decrease with smaller n_{eff}, particularly for less stringent *q*-value cutoffs where stochasticity is less likely to be a factor. Importantly, the number of detected sweeps always remains smaller for the modern population relative to the ancient population for comparable n_{eff} values, indicating that the lack of sweeps in modern populations is unlikely to have been a sample size artefact.

Figure S9. Evaluating the Performance of the PolyLink enrichment algorithm and the contribution of mutually exclusive gene sets to the enriched pathways. (A) Performance of the PolyLink randomisation method used to control for genomic clustering of pathway genes (y-axis) relative to a standard permutation algorithm (x-axis). Clustering is measured via an index that ranges from 0 (i.e. no gene in the pathway sits within 1cM of another gene) to 1 (i.e. all genes are within 1cM from the first gene ordered by chromosome position). When gene clustering is low, both algorithms perform similarly; however, the standard permutation algorithm tends to result in increasingly more outliers than the clustering-sensitive algorithm as clustering increases. Importantly, the *p*-values are linearly correlated between the two approaches at low clustering levels, indicating that clustering-sensitive randomization method is not underpowered when clustering is low. (B) Upset plot showing the number of genes in each mutually exclusive combination of the seven enriched pathways identified using the PolyLink pathway enrichment algorithm (see Material and Methods). In contrast with the other enriched pathways, all four collagen-related pathways lack pathway specific genes, and also share genes with multiple other collagen pathways, whereby they can be considered as a single pathway. (C) Distribution of the *Z*-scores for all mutually exclusive gene sets from the enriched pathways (red bars), with the four collagen metabolism pathways considered as a single pathway (CM; other pathways also referred to by the initial letter in the first two words in each pathway name), relative to expectations under the null (i.e. standard normal distribution; grey bars). Note that

observed gene scores are consistently skewed toward larger values across all mutually exclusive gene sets, including gene sets unique to each specific pathway (i.e. CM, FA, PD, RL). This suggests that the enrichment signal for each pathway comprises a component that is based on pathway-specific genes, and is not solely driven by selection on genes that are shared across pathways.

Figure S10. Enrichment analysis of selected genes based on the location of the associated protein product within the cell. Genes were assigned into categories according to information from Human Cell Atlas Database, and tested for enrichment using PolyLink. The categories were ranked according to which had the highest $-\log_{10}$ transformed *p*-values summed across all populations, with the 1st and 4th ranked categories both being situated at the cell membrane.

SUPPLEMENTARY TABLES AND DATA FIGURES TITLES AND LEGENDS

Table S1. Full set of metadata for all ancient western Eurasian sampled considered in this study. Data includes sample ages, geographic locations, as well as associated cultural and historical affinities for each individual. This information was combined with genetic relationships between the samples in order to categorise the populations used in this study.

Table S2. List of 57 outlier sweeps at *q*-value < 0.01. Each sweep is labeled according to its genomic location and by the first gene present within the selected haplotype. Additional information for each sweep includes the likely origin of selection (according to their earliest presence in ancient west Eurasian samples), the inferred timing of the first impact of admixture on the selected haplotype, and previous published evidence of the sweeps in three recent studies of positive selection on modern European genomes. The full list of outlier genes present in each sweep are also included in the last column. Sweeps listed in bold italics contain either a single gene or multiple genes of the same function, which were used to tentatively infer potential historical selection pressures in western Eurasia.

Table S3. Full list of the 1,800 annotated pathways tested using the PolyLink enrichment method. The dataset was compiled from annotated pathways obtained from BIOCYC, KEGG, REACTOME, and the Pathway Interaction Database (PID). Enrichment tests were performed on genes from each of these pathways to help infer potential functions coming under selection in western Eurasian prehistory.

Table S4. Protein cellular location enrichment. All genes were assigned to one of 33 categories based on the cellular location of their protein product as reported in the Human Cell Atlas Database (HCAD). The number and proportion of genes found in each cellular location are shown for each of the seven enriched pathways, and for the combined genes from all enriched pathways, the combined genes found in candidate sweeps, and all genes shared with the HCAD.

Table S5. Testing for enrichment of GWAS loci in selected genes. Fisher's Exact Test was used to test if the proportion of GWAS hits in each gene set was significantly larger than for the remaining genomic background. Separate tests were conducted for each gene set with or without genes from the MHC locus, for q-value thresholds ranging from 0.01 to 0.5 for the observed sweep signals.

Table S6. Mappability test of sweep regions. The proportion of mappable nucleotides falling within each of the 57 candidate sweep windows relative to the rest of the genome. All sweeps contain at least 96% mappable SNPs, indicating that they are unlikely to be alignment artefacts.

Supplementary Data Figures 1-57. Composite likelihood ratio (CLR) scores (grey line) as a function of genomic coordinates (x-axis) for the top 57 candidate sweeps. Genes within the sweep region are represented by colored boxes, with the color intensity indicating their Z-scores. For each sweep, up to 10 genes with *q*-values less than 0.01 are named. If more than 10 genes have *q*-values < 0.01, then the 10 genes with the highest aggregate *Z* score across all populations are reported. The SNP marker weights used to determine the selected haplotype for each sweep are shown in the top panel, with the SNPs used for the top 30, 50, and 100 markers indicated in the key.

Supplementary Data Figures 58-114. HaploImages for each of the 57 candidate sweeps (*q*-value < 0.01). The images depict the pseudo-haplotype calls at each locus in the sweep window and flanking region, for each population used in this study. Majority alleles, *i.e.* those alleles most likely to be associated with the sweep, are shown in yellow, and minority alleles in black, with missing data being white. Only individuals with no more than 50% missing data are shown. In addition to the populations scanned for sweeps using SweepFinder2, haplo-images for five Late Pleistocene Eurasian AMH samples (Ust'-Ishim, Kostenki14, GoyetQ-116, and El Mirón) were also included to help determine the earliest timing of the selection pressure for each sweep. Note that only Ust'-Ishim had sufficient coverage to make diploid calls (*i.e.* both alleles are present at each locus), whereas the four remaining samples use pseudo-haploid calls.

REFERENCES

1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Akbari, A., Vitti, J.J., Iranmehr, A., Bakhtiari, M., Sabeti, P.C., Mirarab, S., and Bafna, V. (2018). Identifying the favored mutation in a positive selective sweep. Nature Methods *15*, 279–282.

Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664.

Allentoft, M.E., Sikora, M., Sjögren, K.-G., Rasmussen, S., Rasmussen, M., Stenderup, J., Damgaard, P.B., Schroeder, H., Ahlström, T., Vinner, L., et al. (2015). Population genomics of Bronze Age Eurasia. Nature *522*, 167–172.

Ango, F., di Cristo, G., Higashiyama, H., Bennett, V., Wu, P., and Huang, Z.J. (2004). Ankyrin-based subcellular gradient of neurofascin, an immunoglobulin family protein, directs GABAergic innervation at purkinje axon initial segment. Cell *119*, 257–272.

Bailey, S.F., and Bataillon, T. (2016). Can the experimental evolution programme help us elucidate the genetic basis of adaptation in nature? Mol. Ecol. *25*, 203–218.

van den Berg, S.M., van Dam, A.D., Rensen, P.C.N., de Winther, M.P.J., and Lutgens, E. (2017). Immune Modulation of Brown(ing) Adipose Tissue in Obesity. Endocr. Rev. *38*, 46–68.

Bhérer, C., Campbell, C.L., and Auton, A. (2017). Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales. Nat. Commun. *8*, 1–9.

Botigué, L.R., Henn, B.M., Gravel, S., Maples, B.K., Gignoux, C.R., Corona, E., Atzmon, G., Burns, E., Ostrer, H., Flores, C., et al. (2013). Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. Proc. Natl. Acad. Sci. U. S. A. *110*, 11791–11796.

Bourque, S.L., Komolova, M., Nakatsu, K., and Adams, M.A. (2008). Long-Term Circulatory Consequences of Perinatal Iron Deficiency in Male Wistar Rats. Hypertension *51*, 154–159.

Brandefelt, J., Kjellström, E., -O. Näslund, J., Strandberg, G., Voelker, A.H.L., and Wohlfarth, B. (2011). A coupled climate model simulation of Marine Isotope Stage 3 stadial climate. Climate of the Past *7*, 649–670.

Brawand, D., Wagner, C.E., Li, Y.I., Malinsky, M., Keller, I., Fan, S., Simakov, O., Ng, A.Y., Lim, Z.W., Bezault, E., et al. (2014). The genomic substrate for adaptive radiation in African cichlid fish. Nature *513*, 375–381.

Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics

2019. Nucleic Acids Res. *47*, D1005–D1012.

Chang, N.-S. (2016). Bubbling cell death: A hot air balloon released from the nucleus in the cold. Exp. Biol. Med. *241*, 1306–1315.

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 559.

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890.

Chung, M.-K., and Wang, S. (2011). Cold suppresses agonist-induced activation of TRPV1. J. Dent. Res. *90*, 1098–1102.

Cordain, L., Boyd Eaton, S., Sebastian, A., Mann, N., Lindeberg, S., Watkins, B.A., O'Keefe, J.H., and Brand-Miller, J. (2005). Origins and evolution of the Western diet: health implications for the 21st century. The American Journal of Clinical Nutrition *81*, 341–354.

Daló, N.L., Bracho, G.A., and Piña-Crespo, J.C. (2007). Motor impairment and neuronal damage following hypothermia in tropical amphibians. Int. J. Exp. Pathol. *88*, 1–7.

Daub, J.T., Hofer, T., Cutivet, E., Dupanloup, I., Quintana-Murci, L., Robinson-Rechavi, M., and Excoffier, L. (2013). Evidence for Polygenic Adaptation to Pathogens in the Human Genome. Molecular Biology and Evolution *30*, 1544–1558.

DeGiorgio, M., Huber, C.D., Hubisz, M.J., Hellmann, I., and Nielsen, R. (2016). S weep F inder 2: increased sensitivity, robustness and flexibility. Bioinformatics *32*, 1895–1897.

Derrien, T., Estellé, J., Marco Sola, S., Knowles, D.G., Raineri, E., Guigó, R., and Ribeca, P. (2012). Fast Computation and Applications of Genome Mappability. PLoS One *7*, e30377.

Dubiel, D., Ordemann, J., Pratschke, J., Dubiel, W., and Naumann, M. (2015). CAND1 exchange factor promotes Keap1 integration into cullin 3-RING ubiquitin ligase during adipogenesis. Int. J. Biochem. Cell Biol. *66*, 95–100.

Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nature Protocols *4*, 1184–1191.

Durrett, R., and Schweinsberg, J. (2004). Approximating selective sweeps. Theor. Popul. Biol. *66*, 129–138.

Eisenmann, S., Bánffy, E., van Dommelen, P., Hofmann, K.P., Maran, J., Lazaridis, I., Mittnik, A., McCormick, M., Krause, J., Reich, D., et al. (2018). Reconciling material cultures in archaeology with genetic data: The nomenclature of clusters emerging from archaeogenomic analysis. Sci. Rep. *8*, 13003.

Enattah, N.S., Trudeau, A., Pimenoff, V., Maiuri, L., Auricchio, S., Greco, L., Rossi, M., Lentze, M., Seo, J.K., Rahgozar, S., et al. (2007). Evidence of still-ongoing convergence evolution of the

lactase persistence T-13910 alleles in humans. Am. J. Hum. Genet. *81*, 615–625.

Ewing, G., and Hermisson, J. (2010). MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics *26*, 2064–2065.

Excoffier, L., and Ray, N. (2008). Surfing during population expansions promotes genetic revolutions and structuration. Trends Ecol. Evol. *23*, 347–351.

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., et al. (2018). The Reactome Pathway Knowledgebase. Nucleic Acids Res. *46*, D649–D655.

Feder, A.F., Pennings, P.S., and Petrov, D.A. (2019). The clarifying role of time series data in the population genetics of HIV.

Feldman, M., Master, D.M., Bianco, R.A., Burri, M., Stockhammer, P.W., Mittnik, A., Aja, A.J., Jeong, C., and Krause, J. (2019). Ancient DNA sheds light on the genetic origins of early Iron Age Philistines. Sci Adv *5*, eaax0061.

Ferrer-Admetlla, A., Liang, M., Korneliussen, T., and Nielsen, R. (2014). On detecting incomplete soft or hard selective sweeps using haplotype structure. Mol. Biol. Evol. *31*, 1275–1291.

Fiorito, G., Di Gaetano, C., Guarrera, S., Rosa, F., Feldman, M.W., Piazza, A., and Matullo, G. (2016). The Italian genome reflects the history of Europe and the Mediterranean basin. Eur. J. Hum. Genet. *24*, 1056–1062.

Fu, Q., Li, H., Moorjani, P., Jay, F., Slepchenko, S.M., Bondarev, A.A., Johnson, P.L.F., Aximu-Petri, A., Prüfer, K., de Filippo, C., et al. (2014a). Genome sequence of a 45,000-year-old modern human from western Siberia. Nature *514*, 445–449.

Fu, Q., Li, H., Moorjani, P., Jay, F., Slepchenko, S.M., Bondarev, A.A., Johnson, P.L.F., Aximu-Petri, A., Prüfer, K., de Filippo, C., et al. (2014b). Genome sequence of a 45,000-year-old modern human from western Siberia. Nature *514*, 445–449.

Fu, Q., Posth, C., Hajdinjak, M., Petr, M., Mallick, S., Fernandes, D., Furtwängler, A., Haak, W., Meyer, M., Mittnik, A., et al. (2016). The genetic history of Ice Age Europe. Nature *534*, 200–205.

Goudriaan, A., Camargo, N., Carney, K.E., Oliet, S.H.R., Smit, A.B., and Verheijen, M.H.G. (2014). Novel cell separation method for molecular analysis of neuron-astrocyte co-cultures. Front. Cell. Neurosci. *8*, 12.

Grimm, J., Sachs, M., Britsch, S., Di Cesare, S., Schwarz-Romond, T., Alitalo, K., and Birchmeier, W. (2001). Novel p62dok family members, dok-4 and dok-5, are substrates of the c-Ret receptor tyrosine kinase and mediate neuronal differentiation. J. Cell Biol. *154*, 345–354.

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G.,

Nordenfelt, S., Harney, E., Stewardson, K., et al. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. Nature *522*, 207–211.

Haber, M., Doumet-Serhal, C., Scheib, C.L., Xue, Y., Mikulski, R., Martiniano, R., Fischer-Genz, B., Schutkowski, H., Kivisild, T., and Tyler-Smith, C. (2019). A Transient Pulse of Genetic Admixture from the Crusaders in the Near East Identified from Ancient Genome Sequences. Am. J. Hum. Genet. *104*, 977–984.

Harris, R.B., Sackman, A., and Jensen, J.D. (2018). On the unfounded enthusiasm for soft selective sweeps II: Examining recent evidence from humans, flies, and viruses. PLoS Genet. *14*, e1007859.

Hermisson, J., and Pennings, P.S. (2017). Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. Methods in Ecology and Evolution *8*, 700–716.

Hermisson, J., and Pennings, P.S. Soft sweeps and beyond: Understanding the patterns and probabilities of selection footprints under rapid adaptation.

Hernandez, R.D., Kelley, J.L., Elyashiv, E., Melton, S.C., Auton, A., McVean, G., 1000 Genomes Project, Sella, G., and Przeworski, M. (2011). Classic selective sweeps were rare in recent human evolution. Science *331*, 920–924.

Hofmanová, Z., Kreutzer, S., Hellenthal, G., Sell, C., Diekmann, Y., Díez-Del-Molino, D., van Dorp, L., López, S., Kousathanas, A., Link, V., et al. (2016). Early farmers from across Europe directly descended from Neolithic Aegeans. Proc. Natl. Acad. Sci. U. S. A. *113*, 6886–6891.

Huber, C.D., Nordborg, M., Hermisson, J., and Hellmann, I. (2014). Keeping it local: evidence for positive selection in Swedish Arabidopsis thaliana. Mol. Biol. Evol. *31*, 3026–3039.

Huber, C.D., DeGiorgio, M., Hellmann, I., and Nielsen, R. (2015). Detecting recent selective sweeps while controlling for mutation rate and background selection. Mol. Ecol.

Huntley, B., Allen, J.R.M., Collingham, Y.C., Hickler, T., Lister, A.M., Singarayer, J., Stuart, A.J., Sykes, M.T., and Valdes, P.J. (2013). Millennial climatic fluctuations are key to the structure of last glacial ecosystems. PLoS One *8*, e61963.

Iglewicz, B., and Hoaglin, D.C. (1993). How to Detect and Handle Outliers (Asq Press).

Jensen, J.D., Payseur, B.A., Stephan, W., Aquadro, C.F., Lynch, M., Charlesworth, D., and Charlesworth, B. (2018). The importance of the neutral theory in 1968 and 50 years on: a response to Kern & Hahn 2018. Evolution *112*, 2109.

Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P.L.F., and Orlando, L. (2013). mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics *29*, 1682–1684.

Jun, G., Wing, M.K., Abecasis, G.R., and Kang, H.M. (2015). An efficient and scalable analysis framework for variant extraction and refinement from population scale DNA sequence data.

Genome Res. *25*, gr.176552.114–925.

Kamm, J., Terhorst, J., Durbin, R., and Song, Y.S. (2019a). Efficiently Inferring the Demographic History of Many Populations With Allele Count Data. Journal of the American Statistical Association 1–16.

Kamm, J., Terhorst, J., Durbin, R., and Song, Y.S. (2019b). Efficiently Inferring the Demographic History of Many Populations With Allele Count Data. J. Am. Stat. Assoc. 1–16.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30.

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2019). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes.

Kelley, J.L., Madeoy, J., Calhoun, J.C., Swanson, W., and Akey, J.M. (2006). Genomic signatures of positive selection in humans and the limits of outlier approaches. Genome Res. *16*, 980–989.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, a. D. (2002). The Human Genome Browser at UCSC. Genome Research *12*, 996–1006.

Key, F.M., Abdul-Aziz, M.A., Mundry, R., Peter, B.M., Sekar, A., D'Amato, M., Dennis, M.Y., Schmidt, J.M., and Andrés, A.M. (2018). Human local adaptation of the TRPM8 cold receptor along a latitudinal cline. PLoS Genet. *14*, e1007298.

Kinsella, R.J., Kahari, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A., et al. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database *2011*, bar030–bar030.

Kircher, M. (2012). Analysis of High-Throughput Ancient DNA Sequencing Data. In Ancient DNA: Methods and Protocols, B. Shapiro, and M. Hofreiter, eds. (Totowa, NJ: Humana Press), pp. 197–228.

Kjellström, E., Brandefelt, J., Näslund, J.-O., Smith, B., Strandberg, G., Voelker, A.H.L., and Wohlfarth, B. (2010). Simulated climate conditions in Europe during the Marine Isotope Stage 3 stadial. Boreas *39*, 436–456.

Lamnidis, T.C., Majander, K., Jeong, C., Salmela, E., Wessman, A., Moiseyev, V., Khartanovich, V., Balanovsky, O., Ongyerth, M., Weihmann, A., et al. (2018). Ancient Fennoscandian genomes reveal origin and spread of Siberian ancestry in Europe. Nat. Commun. *9*, 5018.

Lang, G.I., and Desai, M.M. (2014). The spectrum of adaptive mutations in experimental evolution. Genomics *104*, 412–416.

Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P.H.,

Schraiber, J.G., Castellano, S., Lipson, M., et al. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature *513*, 409–413.

Lazaridis, I., Nadel, D., Rollefson, G., Merrett, D.C., Rohland, N., Mallick, S., Fernandes, D., Novak, M., Gamarra, B., Sirak, K., et al. (2016). Genomic insights into the origin of farming in the ancient Near East. Nature *536*, 419–424.

Lazaridis, I., Mittnik, A., Patterson, N., Mallick, S., Rohland, N., Pfrengle, S., Furtwängler, A., Peltzer, A., Posth, C., Vasilakis, A., et al. (2017). Genetic origins of the Minoans and Mycenaeans. Nature *548*, 214–218.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics *27*, 2987–2993.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Li, H., Jia, J.-P., Xu, M., and Zhang, L. (2015). Changes in the blood-nerve barrier after sciatic nerve cold injury: indications supporting early treatment. Neural Regeneration Res. *10*, 419–424.

Lipson, M., Szécsényi-Nagy, A., Mallick, S., Pósa, A., Stégmár, B., Keerl, V., Rohland, N., Stewardson, K., Ferry, M., Michel, M., et al. (2017). Parallel palaeogenomic transects reveal complex genetic history of early European farmers. Nature *551*, 368–372.

Lohmueller, K.E., Bustamante, C.D., and Clark, A.G. (2009). Methods for human demographic inference using haplotype patterns from genomewide single-nucleotide polymorphism data. Genetics *182*, 217–231.

Lu, Y., Patterson, N., Zhan, Y., Mallick, S., and Reich, D. (2011). Technical design document for a SNP array that is optimized for population genetics.

MacHugh, D.E., Larson, G., and Orlando, L. (2017). Taming the Past: Ancient DNA and the Study of Animal Domestication. Annual Review of Animal Biosciences *5*, 329–351.

Mallet, J., Besansky, N., and Hahn, M.W. (2016). How reticulated are species? BioEssays *38*, 140–149.

Marciniak, S., and Perry, G.H. (2017). Harnessing ancient genomes to study the history of human adaptation. Nat. Rev. Genet. *18*, 659–674.

Marcus, J.H., Posth, C., Ringbauer, H., Lai, L., Skeates, R., Sidore, C., Beckett, J., Furtwängler, A., Olivieri, A., Chiang, C.W.K., et al. (2020). Genetic history from the Middle Neolithic to present on the Mediterranean island of Sardinia. Nat. Commun. *11*, 939.

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S.A., Harney, E., Stewardson, K., Fernandes, D., Novak, M., et al. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. Nature *528*, 499–503.

Mathieson, I., Alpaslan-Roodenberg, S., Posth, C., Szécsényi-Nagy, A., Rohland, N., Mallick, S., Olalde, I., Broomandkhoshbacht, N., Candilio, F., Cheronet, O., et al. (2018). The genomic history of southeastern Europe. Nature *555*, 197–203.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

Nedergaard, J., and Cannon, B. (2018). Brown adipose tissue as a heat-producing thermoeffector. Handbook of Clinical Neurology 137–152.

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M.J., Clark, A.G., and Bustamante, C. (2005). Genomic scans for selective sweeps using SNP data. Genome Res. *15*, 1566–1575.

O'Connell, J.F., Allen, J., Williams, M.A.J., Williams, A.N., Turney, C.S.M., Spooner, N.A., Kamminga, J., Brown, G., and Cooper, A. (2018). When did Homo sapiens first reach Southeast Asia and Sahul? Proc. Natl. Acad. Sci. U. S. A. *115*, 8482–8490.

Olalde, I., Brace, S., Allentoft, M.E., Armit, I., Kristiansen, K., Booth, T., Rohland, N., Mallick, S., Szécsényi-Nagy, A., Mittnik, A., et al. (2018). The Beaker phenomenon and the genomic transformation of northwest Europe. Nature *555*, 190–196.

Paradis, E., and Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics *35*, 526–528.

Pardo-Seco, J., Gómez-Carballa, A., Amigo, J., Martinón-Torres, F., and Salas, A. (2014). A genome-wide study of modern-day Tuscans: revisiting Herodotus's theory on the origin of the Etruscans. PLoS One *9*, e105920.

Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet. *2*, e190.

Pickrell, J.K., Coop, G., Novembre, J., Kudaravalli, S., Li, J.Z., Absher, D., Srinivasan, B.S., Barsh, G.S., Myers, R.M., Feldman, M.W., et al. (2009). Signals of recent positive selection in a worldwide sample of human populations. Genome Res. *19*, 826–837.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.

Pritchard, J.K., Pickrell, J.K., and Coop, G. (2010). The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr. Biol. *20*, R208–R215.

Przeworski, M., Coop, G., and Wall, J.D. (2005). The signature of positive selection on standing genetic variation. Evolution *59*, 2312–2323.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

Racimo, F. (2016). Testing for Ancient Selection Using Cross-population Allele Frequency Differentiation. Genetics *202*, 733–750.

Rani, V., Deep, G., Singh, R.K., Palle, K., and Yadav, U.C.S. (2016). Oxidative stress and metabolic disorders: Pathogenesis and therapeutic strategies. Life Sciences *148*, 183–193.

Romero, P., Wagg, J., Green, M.L., Kaiser, D., Krummenacker, M., and Karp, P.D. (2005). Computational prediction of human metabolic pathways from the complete human genome. Genome Biol. *6*, R2.

Sabeti, P.C. (2006). Positive Natural Selection in the Human Lineage. Science *312*, 1614–1620.

Schaefer, C.F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., and Buetow, K.H. (2009). PID: the Pathway Interaction Database. Nucleic Acids Research *37*, D674–D679.

Schrider, D.R., and Kern, A.D. (2017). Soft Sweeps Are the Dominant Mode of Adaptation in the Human Genome. Mol. Biol. Evol. *34*, 1863–1877.

Seguin-Orlando, A., Korneliussen, T.S., Sikora, M., -S. Malaspinas, A., Manica, A., Moltke, I., Albrechtsen, A., Ko, A., Margaryan, A., Moiseyev, V., et al. (2014). Genomic structure in Europeans dating back at least 36,200 years. Science *346*, 1113–1118.

Şeker, N. (2013). Forced Population Movements in the Ottoman Empire and the Early Turkish Republic: An Attempt at Reassessment through Demographic Engineering. European Journal of Turkish Studies, 16 | 2013 Demographic Engineering - Part III.

Silvert, M., Quintana-Murci, L., and Rotival, M. (2019). Impact and evolutionary determinants of Neanderthal introgression on transcriptional and post-transcriptional regulation. bioRxiv 532366.

Skoglund, P., and Mathieson, I. (2018). Ancient Genomics of Modern Humans: The First Decade. Annu. Rev. Genomics Hum. Genet. *19*, 381–404.

Smith, J.M., and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. Genetical Research *23*, 23–35.

Sohail, M., Maier, R.M., Ganna, A., Bloemendal, A., Martin, A.R., Turchin, M.C., Chiang, C.W., Hirschhorn, J., Daly, M.J., Patterson, N., et al. (2019). Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. Elife *8*.

Spencer, K.A., Belgacem, Y.H., Visina, O., Shim, S., Genus, H., and Borodinsky, L.N. (2019). Growth at Cold Temperature Increases the Number of Motor Neurons to Optimize Locomotor Function. Curr. Biol. *29*, 1787–1799.e5.

Storey, J.D. (2003). The positive false discovery rate: a Bayesian interpretation and the q - value. The Annals of Statistics *31*, 2013–2035.

Szpak, M., Xue, Y., Ayub, Q., and Tyler-Smith, C. (2019). How well do we understand the basis of classic selective sweeps in humans? FEBS Lett.

Team, R.C. (2012). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. ISBN: 3-900051-07-0.

Teng, C.-C., Yang, Y.-T., Chen, Y.-C., Kuo, Y.-M., and Sze, C.-I. (2013). Role of WWOX/WOX1 in Alzheimer's disease pathology and in cell death signaling. Front. Biosci. *5*, 72–85.

Thul, P.J., Åkesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Ait Blal, H., Alm, T., Asplund, A., Björk, L., Breckels, L.M., et al. (2017). A subcellular map of the human proteome. Science *356*.

Timmermann, A., and Friedrich, T. (2016). Late Pleistocene climate drivers of early human migration. Nature *538*, 92–95.

Van Meerbeeck, C.J., Renssen, H., Roche, D.M., Wohlfarth, B., Bohncke, S.J.P., Bos, J.A.A., Engels, S., Helmens, K.F., Sánchez-Goñi, M.F., Svensson, A., et al. (2011). The nature of MIS 3 stadial–interstadial transitions in Europe: New insights from model–data comparisons. Quaternary Science Reviews *30*, 3618–3637.

Voelker, A.H.L. (2002). Global distribution of centennial-scale records for Marine Isotope Stage (MIS) 3: a database. Quaternary Science Reviews *21*, 1185–1212.

Vy, H.M.T., and Kim, Y. (2015). A Composite-Likelihood Method for Detecting Incomplete Selective Sweep from Population Genomic Data. Genetics *200*, 633–649.

Wang, C.-C., Reinhold, S., Kalmykov, A., Wissgott, A., Brandt, G., Jeong, C., Cheronet, O., Ferry, M., Harney, E., Keating, D., et al. (2019). Ancient human genome-wide data from a 3000-year interval in the Caucasus corresponds with eco-geographic regions. Nat. Commun. *10*, 590.

Weir, B.S., and Cockerham, C.C. (1984). ESTIMATING F-STATISTICS FOR THE ANALYSIS OF POPULATION STRUCTURE. Evolution *38*, 1358–1370.

Whitlock, M.C., and Lotterhos, K.E. (2015). Reliable Detection of Loci Responsible for Local Adaptation: Inference of a Null Model through Trimming the Distribution of F(ST). Am. Nat. *186 Suppl 1*, S24–S36.

Yu, H., Shin, S.M., Wang, F., Xu, H., Xiang, H., Cai, Y., Itson-Zoske, B., and Hogan, Q.H. (2019). Transmembrane protein 100 is expressed in neurons and glia of dorsal root ganglia and is reduced after painful nerve injury. Pain Rep *4*, e703.

Yudin, N.S., Larkin, D.M., and Ignatieva, E.V. (2017). A compendium and functional characterization of mammalian genes involved in adaptation to Arctic or Antarctic environments. BMC Genetics *18*.

ʿÔdēd, B. (1979). Mass deportations and deportees in the Neo-Assyrian Empire (Reichert).