Population Genomics of Stone Age Eurasia:

Supplementary Information (Part 1)

Key to Supplementary Data Tables

No.	Title
Supplementary Table I.	Basic overview of samples and genetic data
Supplementary Table II.	Samples, dates and isotopic data
Supplementary Table III.	Reservoir correction calculations for radiocarbon dates
Supplementary Table IV.	Full sample metadata (dates and isotopic data non-cleaned and non-combined)
Supplementary Table V.	DNA contamination estimation
Supplementary Table VI.	Relatedness Estimates
Supplementary Table VII.	Ancient Genomes dataset (incl. all imputed)
Supplementary Table VIII.	IBD mixture model sets
Supplementary Table IX.	IBD Ancestry proportions for set "deep"
Supplementary Table X.	IBD Ancestry proportions for set "postNeol"
Supplementary Table XI.	IBD Ancestry proportions for set "postBA"
Supplementary Table XII.	IBD Ancestry proportions for set "hgEur"
Supplementary Table XIII.	IBD Ancestry proportions for set "fEur"
Supplementary Table XIV.	IBD Ancestry proportions for set "postNeolScand"
Supplementary Table XV.	CLUES modern
Supplementary Table XVI.	CLUES ancient

1) Data Generation and Authentication

Morten E. Allentoft[1,2], Simon Rasmussen[3], Gabriel Renaud[4], Abigail D. Ramsøe[2], Thorfinn Korneliussen[2], Martin Sikora[2]

[1]Trace and Environmental DNA (TrEnD) Lab, Curtin University, Perth, Australia
[2]Lundbeck Foundation GeoGenetics Centre, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark
[3]Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Copenhagen, Denmark
[4]Department of Health Technology, Section of Bioinformatics, Technical University of Denmark, Kongens Lyngby, Denmark

Sampling, lab work and sequencing

The lab work component of this project followed the same procedures outlined in Allentoft et al. [1] and Damgaard et al. [2], also including sampling of the petrous part of the temporal bone following the discovery of exceptional DNA preservation in these bones [3,4]. While new ancient DNA (aDNA) extraction and library methods are continually optimised and presented in the literature, we prioritised method consistency throughout the project period to avoid the risk of introducing batch effects in the data.

A total of 962 Stone Age and early Bronze Age human skeletons from across Eurasia were sampled for this project. An initial molecular 'screening' to assess the endogenous DNA content (proportion of DNA sequences identified as human) was performed by shallow shotgun sequencing resulting in 317 samples (Supplementary Table I) being selected for deeper sequencing. We applied a threshold of <1% endogenous DNA for rejecting samples in the project, except for a few Danish skeletons that were prioritised despite displaying even lower contents. Of the 317 samples, 211 were teeth, 91 were petrous bones, and 15 were pieces of other types of bones (long bones, ribs, cranial bones).

Following some AMS-dates which were conducted late in the project, two of the 317 samples (NEO901 and NEO902, Supplementary Table I) proved too young to be relevant for this project. These two samples have only been included for imputation purposes together with the data that are released as part of this project, but they have not been included in downstream analyses. Hence the final number of new samples that were sequenced, analysed and discussed as part of this project is 315, and this is the number we refer to throughout the study.

All the pre-PCR-amplification lab work was conducted in dedicated clean laboratories at the Lundbeck Foundation GeoGenetics Centre (GLOBE Institute, University of Copenhagen), according to strict aDNA guidelines [5–7]. To reduce the amount of non-target DNA in the extracts, the outermost surfaces of the samples were first removed using a sterile cutting disc. Teeth were processed by separating the crown from the root by a cutting disc and the inner dentine was then removed from the root with a pointy drilling bit. By this procedure each root sample was proportionally enriched for the outer cementum layer which is known for its high endogenous DNA content [2,8,9]. Petrous bones were sampled by cutting off slices (with a cutting disc) until reaching the dense otic capsule which was used for DNA extraction. The samples were crushed into smaller pieces before the lysis step.

To further increase the endogenous DNA yield, we performed a brief 'pre-digestion' step prior to the extraction protocol following Damgaard et al. [2]. After this pre-digestion, we added 3.5mL of fresh digestion buffer to each sample and incubated them for 24h before the DNA was purified with silica-in-solution similar to Rohland & Hofreiter [10] but using the optimised binding buffer from Allentoft et al. [1]. Double-stranded blunt-end libraries were constructed from the extracted DNA using NEBNext DNA Prep Master Mix Set E6070 (New England Biolabs Inc.) with protocol modifications [11] and then amplified with indexed Illumina-specific adapters prepared as in [12]. The DNA concentration of each amplified library was quantified on an Agilent 2200 Tapestation and sequencing (80bp and 100bp single read) was performed on Illumina HiSeq 2500 and Illumina HiSeq 4000 platforms at the Danish National High-throughput DNA Sequencing Centre.

Basic bioinformatics

The Illumina data was base-called using Illumina software *CASAVA (v.1.8.2)*[13] and sequences were de-multiplexed with the requirement of full matching of the six nucleotide index which was used for library preparation. Adapter sequences and leading/trailing stretches of Ns were trimmed from the reads and bases with quality 2 or less were removed using *AdapterRemoval (v.2.1.3)*. Trimmed reads of at least 30bp were mapped using *bwa (v.0.7.10)*[14] with the seed disabled to allow for higher sensitivity [15]. Reads were mapped to the human reference genome build 37 including mitochondrial DNA (rCRS) and to mitochondrial DNA alone. Mapped reads were filtered for mapping quality 30 and sorted using *Picard (v.1.127)* (http://picard.sourceforge.net) and *samtools* [16]. Data was merged to library level and duplicates removed using *Picard MarkDuplicates (v.1.127)* and hereafter merged to sample level. Sample level BAMs were re-aligned using *GATK (v.3.3.0)* and

hereafter had the md-tag updated and extended BAQs calculated using *samtools calmd (v.1.10)*[16]. Read depth and coverage were determined using *pysam* (https://github.com/pysam-developers/pysam) and *BEDtools (v.2.23.0)* [17].

DNA authentication

To investigate authenticity of the ancient DNA molecules, post-mortem DNA damage patterns were determined using *mapDamage2.0* [18]. Cytosine deamination was recorded for each sample as the fraction of C-to-T transitions at the first 5'position of the DNA reads when compared to the reference genome. For the 317 samples included here, we observed C-to-T deamination fractions ranging from 10.4% to 67.8%, with an average of 38.3% across all 315 samples (Supplementary Table I). These numbers generally reflect highly damaged molecules as expected for DNA that is thousands of years old.

Next, we applied three different methods to estimate levels of DNA contamination; two based on mitochondrial genome data and one method investigating X-chromosomal data in males. All contamination estimates are reported in Supplementary Table V with summary values provided in Supplementary Table I.

First, estimates of present-day human contamination for the mitochondrial genome were performed using the iterative Bayesian framework implemented in *Schmutzi* [19]. Briefly, ancient DNA sequences were realigned to the mitochondrial genome of the revised Cambridge Reference Sequence (rCRS, NCBI Reference Sequence: NC_012920.1) using *BWA* [14] with parameters for increased sensitivity (*-n 0.01 -o 2 -l 16500*). The mapping was performed exclusively to the mitochondrial genome to mitigate the impacts of nuclear NUMTs on the mitochondrial alignments. The resulting BAM file was used as input for *Schmutzi* using five iterations and by subsampling samples with coverage to 500X, should they exceed that [19]. The iterative approach was run using a database of Eurasian mitogenomes. The point estimate for the final contamination rate with the maximum *a posteriori* probability is reported in the sample summary Supplementary Table I, whereas the 95% confidence interval (lower and upper bound as well as the point estimate) are reported in Supplementary Table V.

Next, we applied *ContamMix* in order to estimate the fraction of non-endogenous reads in the mitochondrial genome by comparing the reconstructed mtDNA consensus sequence to

311 possible contaminate genomes [20]. For each sample, an in-house perl script was used to construct two different versions of the endogenous mitochondrial genome. The first approach (CONTAMIX_APPROX_1Xdif05) used sites with at least 1x coverage, and at each position a base was only called if it was observed in at least 50% of reads covering the site. The second approach (CONTAMIX_PRECISE_5Xdif07) only considered sites with at least 5x coverage and 70% of reads agreeing. Both approaches used reads with a base quality of ≥20 and mapping quality of ≥30.

Lastly, we applied *ANGSD* [21] on X-chromosomal data in males. This approach quantifies heterozygosity on the X chromosome. As males only have one copy of the X chromosome, any heterozygosity is expected to arise from either contamination or sequencing error. As heterozygosity due to contamination is expected to be restricted to mainly known diagnostic polymorphic sites, *ANGSD* quantifies the heterozygosity in these sites in. It then compares it to adjacent sites in order to ascertain the level of background sequencing error, and thus estimates the extent of contamination. For each sample, we removed the pseudoautosomal regions on the X chromosome and filtered out reads with a base quality <20 and mapping quality <30.

DNA contamination - results and implications

Supplementary Table V lists all the contamination estimation results for the 317 samples across the three applied methods. The vast majority of the samples show very low levels of contamination (≤ 5%) across all methods. A total of 33 samples, however, display contamination estimates >5% in one or more of the methods, but there are considerable inconsistencies between the methods (Supplementary Table V). It is well known that contamination estimates are not reliable for very low coverage genomic data (refs) and this is further complicated by DNA damage in the sequences. Indeed, we observe that the 33 potentially problematic samples have an average coverage of 0.11X and a median of 0.06X which is considerably lower than the full dataset with its average of coverage of 0.75X and median of 0.26X. So, instead of simply excluding data from these precious samples based on estimates that are likely imprecise, we “flagged” samples as potentially contaminated in downstream analyses and took a more analytical approach in the evaluation. Flagging was applied as follows:

Samples with nuclear coverage <0.1X and MT coverage <10X: Not flagged, since both estimates are likely unreliable

Samples with nuclear coverage <0.1X and MT coverage ≥10X: Flagged as contaminated if any MT estimate is >5%; ignore nuclear estimate as likely unreliable

Male samples with nuclear coverage ≥0.1X: Flagged as contaminated if any nuclear (X-chromosomal) estimate is > 5%; ignore MT as only nuclear data are relevant for genome-wide analyses

Female samples with nuclear coverage ≥0.1X: Flagged as contaminated if any MT estimate is >5% as no nuclear estimate is available

Based on this approach we have a total of 15 samples (NEO1, NEO3, NEO76, NEO77, NEO158, NEO162, NEO168, NEO221, NEO226, NEO537, NEO657, NEO671, NEO677, NEO746, NEO815) that we have flagged as “possibly contaminated” (See Table S2) in our downstream analyses. The further analytical evaluation of these samples is described in Supplement Note S3d.

References

1. Allentoft, M. E. *et al.* Population genomics of Bronze Age Eurasia. *Nature* **522**, 167–172 (2015).

2. Damgaard, P. B. *et al.* Improving access to endogenous DNA in ancient bones and teeth. *Sci. Rep.* **5**, 11184 (2015).

3. Edson, S. M. *et al.* Sampling of the cranium for mitochondrial DNA analysis of human skeletal remains. *Forensic Science International: Genetics Supplement Series* **2**, 269–270 (2009).

4. Gamba, C. *et al.* Genome flux and stasis in a five millennium transect of European prehistory. *Nat. Commun.* **5**, 5257 (2014).

5. Knapp, M., Clarke, A. C., Horsburgh, K. A. & Matisoo-Smith, E. A. Setting the stage--Building and working in an ancient DNA laboratory. *Annals of Anatomy-Anatomischer Anzeiger* **194**, 3–6 (2012).

6. Fulton, T. L. & Shapiro, B. Setting Up an Ancient DNA Laboratory. *Methods Mol. Biol.*

1963, 1–13 (2019).

7. Orlando, L. *et al.* Ancient DNA analysis. *Nature Reviews Methods Primers* **1**, 1–26 (2021).

8. Higgins, D., Kaidonis, J., Townsend, G., Hughes, T. & Austin, J. J. Targeted sampling of cementum for recovery of nuclear DNA from human teeth and the impact of common decontamination measures. *Investig. Genet.* **4**, 18 (2013).

9. Hansen, H. B. *et al.* Comparing Ancient DNA Preservation in Petrous Bone and Tooth Cementum. *PLoS One* **12**, e0170940 (2017).

10. Rohland, N. & Hofreiter, M. Comparison and optimization of ancient DNA extraction. *Biotechniques* **42**, 343–352 (2007).

11. Margaryan, A. *et al.* Population genomics of the Viking world. *Nature* **585**, 390–396 (2020).

12. Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* **2010**, db.prot5448 (2010).

13. Hosseini, P., Tremblay, A., Matthews, B. F. & Alkharouf, N. W. An efficient annotation and gene-expression derivation tool for Illumina Solexa datasets. *BMC Res. Notes* **3**, 183 (2010).

14. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

15. Schubert, M. *et al.* Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics* **13**, 178 (2012).

16. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

17. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

18. Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F. & Orlando, L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage

parameters. *Bioinformatics* **29**, 1682–1684 (2013).

19. Renaud, G., Slon, V., Duggan, A. T. & Kelso, J. Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. *Genome Biol.* **16**, 224 (2015).

20. Fu, Q. *et al.* A revised timescale for human evolution based on ancient mitochondrial genomes. *Curr. Biol.* **23**, 553–559 (2013).

21. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15**, 356 (2014).

2) Imputation of ancient DNA

Bárbara Sousa da Mota[1,2], Andrew Vaughn[3] & Olivier Delaneau[1,2]

[1] Department of Computational Biology, University of Lausanne, Switzerland
[2] Swiss Institute of Bioinformatics, University of Lausanne, Switzerland
[3] Center for Computational Biology, University of California, Berkeley, USA

Introduction

GLIMPSE is a statistical method developed to impute low-coverage human genomes. It has been shown that GLIMPSE efficiently produces accurate results when employed to impute low-coverage present-day genomes. Here we seek to demonstrate that GLIMPSE is also a suitable imputation tool of low-coverage ancient genomes. To benchmark it, we used a subset of 42 previously published ancient genomes with a mean depth of coverage above 10x (sample list is given in Table S2.1) that we downsampled to lower coverages in order to match the coverage we observed for the 317 genomes sequenced in the present study. Then, we imputed the resulting genomes and assessed the accuracy of the imputed calls by comparing with the original high-coverage genomes. Specifically, we examined the imputation performance regarding (i) depth of coverage, (ii) minor allele frequency, (iii) ancestry and living period of target samples. We also imputed the samples with the present 'gold standard' method, Beagle4.1 [1], to show how it compares with GLIMPSE v1.0.1 [2] (https://odelaneau.github.io/GLIMPSE/).

Methods

We first prepared all necessary files for the imputation step. We used samtools 1.10 to downsample the 42 high-coverage genomes to coverages 0.1x, 0.2x, 0.4x, 0.8x,1.0x, 2.0x and 4.0x. Then, we prepared a list of candidate variant sites at which imputation was performed by retaining all sites in 1000 Genomes version 3 that were (i) bi-allelic SNPs and (ii) non-singleton (i.e. informative for imputation). For each of the seven tested coverages, we computed genotype likelihoods (VCF/PL field) at all candidate variant sites across all target samples using bcftools 1.10. To minimise computation time, we restricted this data generation procedure to chromosome 20.

Then, we performed imputation of all the resulting VCF files. We first divided chromosome 20 into 35 chunks with size between 1Mb and 2Mb. To prevent edge effects, we also added

additional buffer regions of 200kb on each side. Splitting chromosome 20 according to these parameters was done using GLIMPSE_chunk v1.0.1. We then performed the imputation with GLIMPSE_phase using the reference panel 1000 Genomes version 3, a cosmopolitan collection of whole genome sequenced modern samples that we feel well adapted to the various ancestries included in our data set. GLIMPSE_phase was run using the following parameters: 10 burn-in iterations (*--burn 10*), 15 main iterations (*--main 15*) and a depth of 2 for the conditional state selection based on Positional Burrows-Wheeler transform (*--pbwt-depth 2*). Finally, we ligated all imputed chunks back together into chromosome-wide VCF files using GLIMPSE_ligate v1.0.1.

In addition to this, we also performed imputation with Beagle 4.1 with exactly the same input data and reference panel and set its parameters *modelscale* to 2 and *niterations* to 0. These parameter settings allow Beagle v4.1 to run with tractable running times on the data while retaining good accuracy. The chunks of data imputed with Beagle v4.1 were ligated together with bcftools concat 1.10.

To evaluate imputation performance, we employed GLIMPSE_concordance v1.0.1 using as validation set all genotypes in high-coverage data that were covered by at least eight sequencing reads and at which the most likely genotype was at least 1,000 times more likely than the second best given the genotype likelihoods reported in the VCF/PL fields. Specifically, we computed (i) the squared correlation and (ii) the concordance between imputed and validation genotypes. For (i), we compared minor allele dosages (VCF/DS field) within multiple minor allele frequency (MAF) bins. For (ii), we compared best guess genotypes (VCF/GT field) and stratified the results depending on the type of validation genotype: homozygous reference allele, heterozygous and homozygous alternative allele. As further validation, we increased genomic coverage to 27.5X, 18.9X and 5.4X by deep sequencing a previously published family trio (mother, father, son) from the Late Neolithic mass burial at Koszyce in Poland [3]. This presented an opportunity to validate imputed genotypes and haplotypes on the basis of Mendel's rules of inheritance [4].

Results

In Fig. S2.1 and S2.2, we present the imputation accuracy per downsampled genome to 1.0x for chromosome 20. We divided samples into eight classes based on expected genetic proximity and plotted each class separately. In Fig. S2.1, the minor allele frequency (MAF) for each of these groups was estimated from the 1000G reference panel, using European, African, South East Asian, East Asian or American allele frequencies according to the place

of origin of samples. As expected, imputation accuracy decreases as minor allele frequency decreases too. For common variants (MAF≥5%), imputation accuracy is remarkably high (>0.9) and closely matches what is usually obtained for modern samples. An exception to this are African samples which exhibit lower accuracy in some cases (especially for baa01). This likely results from the reference panel we used that does not represent well the underlying ancestries of these samples. In Fig. S2.2, we present imputation accuracy per genome as genotype discordance: the fraction of validation genotypes incorrectly imputed stratified by homozygous and heterozygous genotypes. As expected, homozygous reference alleles exhibit lower error rates than heterozygous and homozygous alternative alleles. Error rates are remarkably low: less than 1% overall and less than 5% for the most challenging genotypes to impute (RA and AA). Again, African samples exhibit much higher error rates.

In Fig. S2.3, we present how imputation accuracy varies for all 42 samples depending on multiple factors expected to affect accuracy. First, we look at coverage and find that 0.4x coverage is enough to get 0.9 imputation accuracy at common variants (MAF≥10%). Of note, even 0.1x allows reaching 0.8 at common variants. Second, we considered whether imputing the 42 jointly with the remaining 1,622 low coverage samples could decrease imputation accuracy and did not find evidence of this happening: we get very similar accuracy results. Finally, we check how GLIMPSE imputation does compare to Beagle 4.1 in case of ancient low coverage samples and find that GLIMPSE brings substantial accuracy boost across the entire frequency range. In Fig. S2.4, we show the phasing and imputation performance we obtained across the entire genome for multiple coverages (0.1x to 4x). We obtained Mendelian error rates from 0.1% at 4X to 0.55% at 0.1X (Extended Figure 1E, Supplementary figure S2.4A). When looking only at sites at which at least one sample is heterozygote (i.e. excluding triple homozygotes), we find that Mendel error rate ranges from less than 2% at 4x and up to 8-10% for 0.1x (Supplementary figure S2.4B). Similarly, we obtained switch error rates between 2% and 6%. Altogether, our validation analysis showed that ancient European genomes can be imputed confidently from coverages above 0.4x and highly valuable data can still be obtained with coverages as low as 0.1X when using specific QC on the imputed data.

Imputation of the full dataset

Given the outcome of the benchmarking described above, we then proceeded with the imputation of the full dataset. In total, we retained 1,664 samples with at least 0.1x mean coverage. Similarly as before, we extracted all variable positions in 1000 Genomes version 3 that correspond to non-singleton bi-allelic SNPs and call genotype likelihoods at all these

variants for all samples using bcftools v1.10, thereby resulting in a VCF containing data for 1,664 samples across 43,285,119 SNPs. Of note, the reference genome used in this analysis was hg19, b37. We then used GLIMPSE_chunk to split all the data into 1,841 chunks of 1Mb to 2Mb with overlapping 200kb buffers on each side. All these chunks of data were imputed using GLIMPSE_phase v1.0.1 with 1000 Genomes version 3 as a reference panel of haplotypes. Imputed chunks were ligated back together using GLIMPSE_ligate v1.0.1, resulting in chromosome-wide VCF files containing the following information: (i) best genotype guesses (VCF/GT field), (ii) expected non-reference allele dosage (VCF/DS field), (iii) genotype posterior probabilities /VCF/GP field) and (iv) haplotype pairs sampled during imputation (VCF/HS field). Finally, we used GLIMPSE_sample v1.0.1 to produce consensus haplotype calls at all variants for all samples from the VCF/HS field.

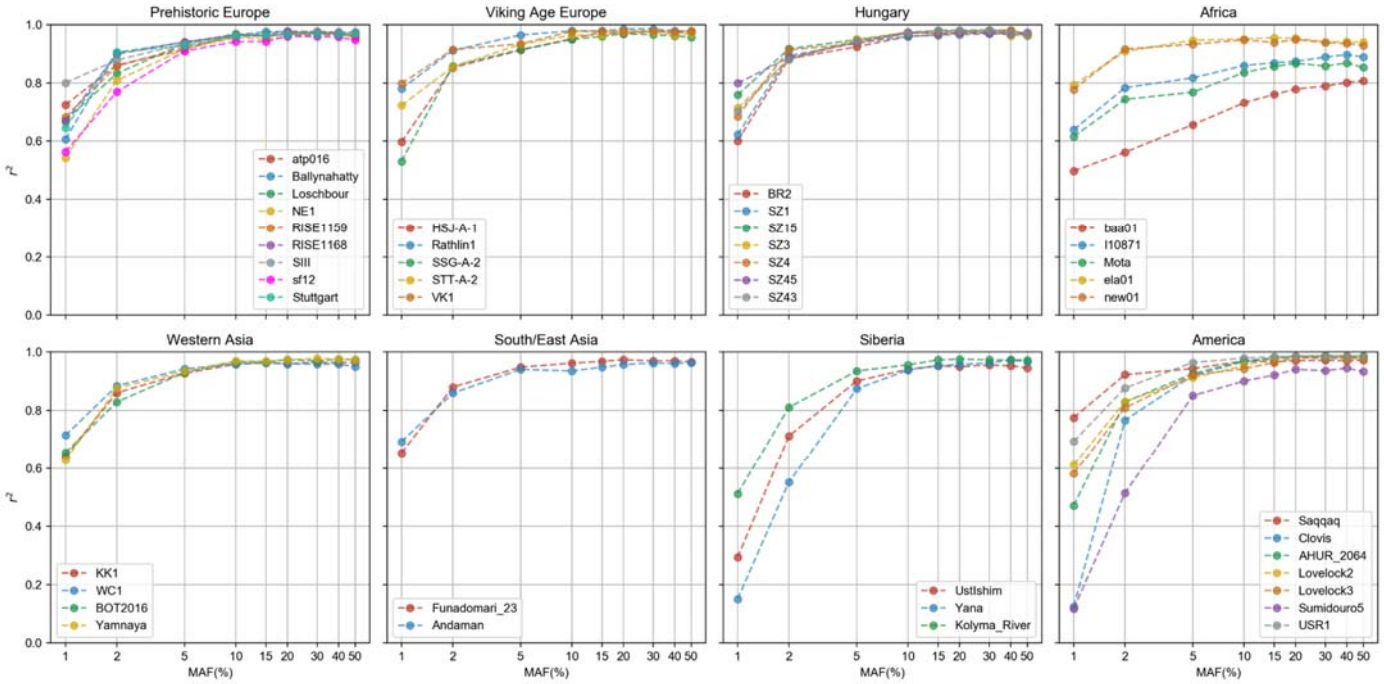
Effects of Low coverage

It should be noted, however, that certain issues may arise when using this imputation on very low coverage data. Specifically, the imputation errors GLIMPSE makes with low coverage data tend to predominantly occur by incorrectly filling in the major allele at a given SNP. This is not an issue specific to GLIMPSE itself but is instead inherent to any kind of Bayesian approach to imputation. In the absence of informative data about the allele at a particular SNP, imputation methods will fall back on the reference panel, or a set of confidently imputed genomes, for imputation, which will tend to fill in missing data with the major allele at each SNP.

To illustrate this phenomenon, we took the 1,492 samples that passed all filters (described in detail in Supplementary Note 3d) and retained only SNPs passing the 1000 Genomes strict mask. For computational considerations, we considered only chromosome 8, which gave 1,139,150 total SNPs. In Figure S2.5, we plotted, for each of the 2,984 haplotypes, the total number of allele differences between that genome and the reference sequence against the coverage of that sample. Computing Spearman's ρ showed a substantial correlation, which persisted after filtering for SNPs on INFO > 0.5, leaving 584,280 SNPs, and INFO > 0.8, leaving 336,842 SNPs (Figures S2.6 and S2.7 respectively). Visually, it appeared that this correlation was driven by a reduced number of differences to the reference sequence among very low coverage genomes, and we confirmed this by considering SNPs of all INFO scores and noticing that the correlation decreases sharply when samples below 0.3x are dropped (Figure S2.9) and can be decreased even further by dropping samples below 2x coverage (Figure S2.10). We also confirmed that the correlation is very small when only retaining very

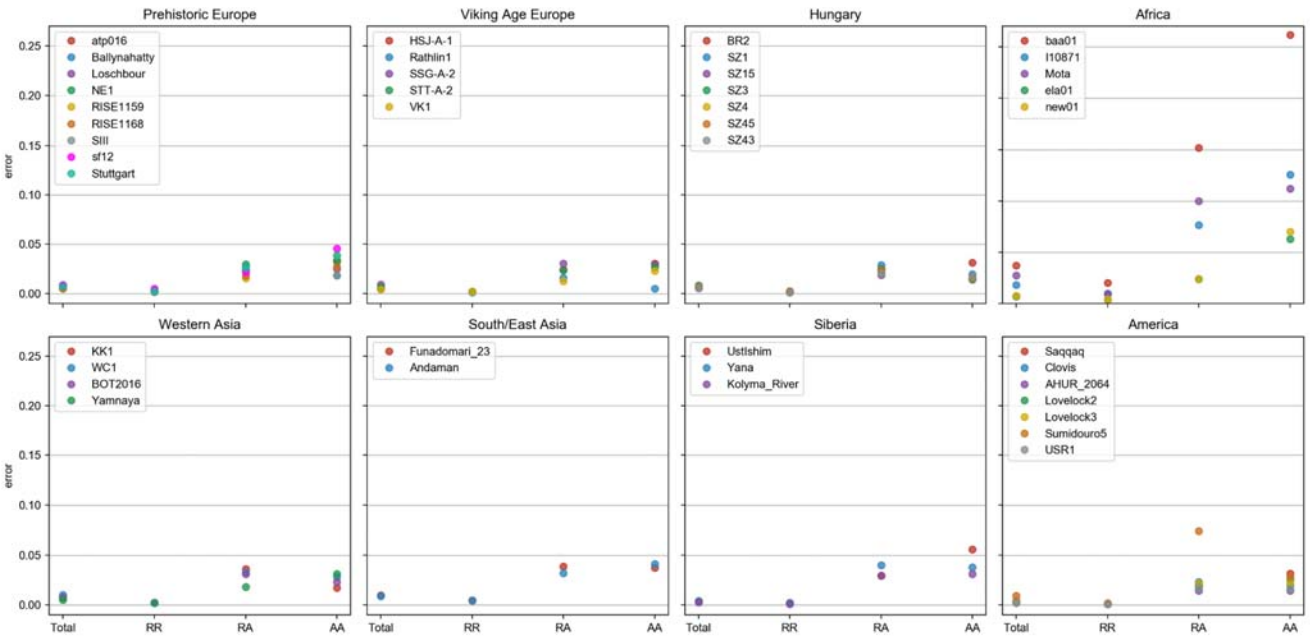
high INFO score SNPs, specifically INFO > 0.97, which retained 56,925 SNPs (Figure S2.8). This is to be expected, as the imputation for high-confidence SNPs should show no significant biases.

This phenomenon, where coverage can be predictive of sequence, is important to keep in mind when running certain types of analyses on these data. Analyses such as PCA and admixture modelling, which mainly rely on common SNPs that are shared among many individuals, are not expected to be significantly affected, as imputation is quite accurate for SNPs with high MAF (see Supplementary Note 3d for a thorough analysis of how coverage affects PCA). However, this observation has important implications for genealogy reconstruction and other analyses that rely on overall sequence similarity or otherwise utilise rare SNPs. We recommend that researchers using these imputed data carefully consider what effect the inclusion of low coverage samples might have on their analyses and then utilise appropriate MAF/INFO filters on SNPs and/or coverage filters on samples as necessary.

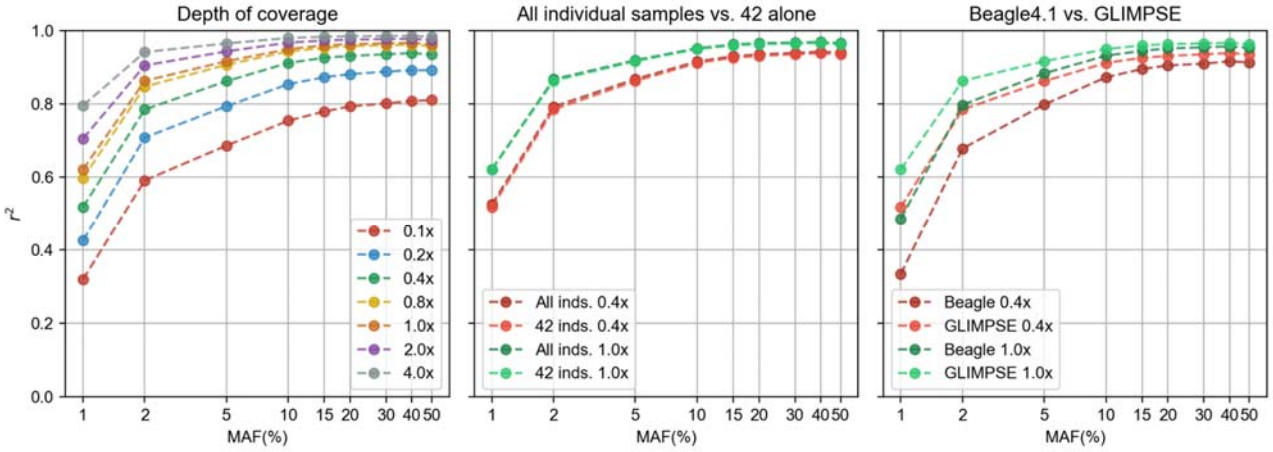


564

565 **Figure S2.1: Per-sample imputation accuracy (1).** Imputation accuracy as squared
566 correlation between imputed and validation genotypes (y-axis). Samples were grouped into
567 eight broad categories based on genetic proximity. Each of the plots corresponds to one of
568 such categories. Minor allele frequencies (MAF; x-axis) were estimated from 1000 Genomes
569 version 3 for matched continental groups (see Table S2.1 for details).
570



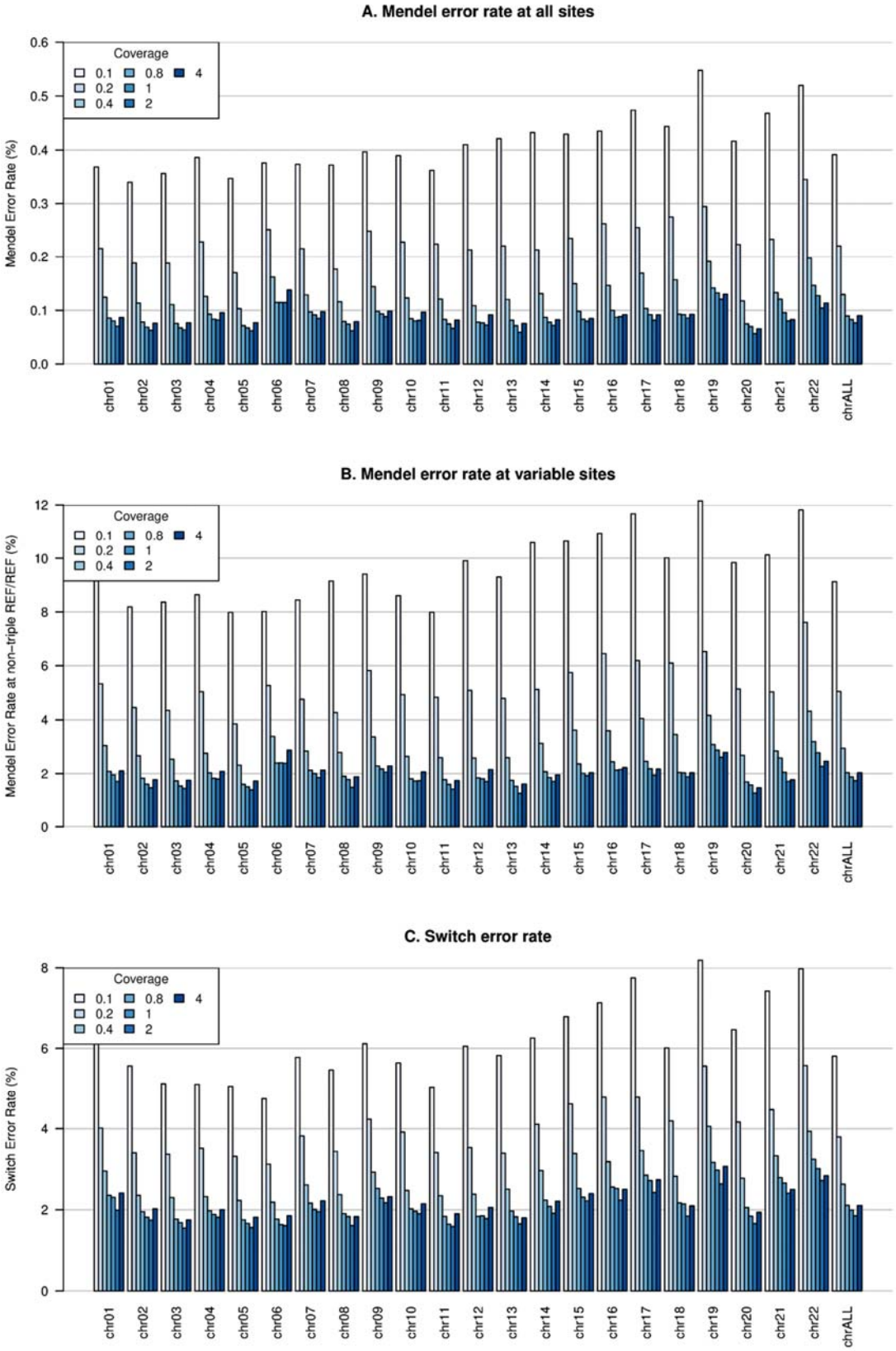
571 **Figure S2.2: Per-sample imputation accuracy (2).** Imputation accuracy as discordance
572 between imputed and validation genotypes (y-axis). Samples were grouped into eight broad
573 categories based on genetic proximity. Each of the plots corresponds to one of such
574 categories. We report results across four types of genotypes: (i) Total; all genotypes
575 together, (ii) RR; validation genotypes with two copies of the reference allele, (iii) RA;
576 heterozygous genotypes (iv) AA; validation genotypes with two copies of the alternative
577 allele.
578



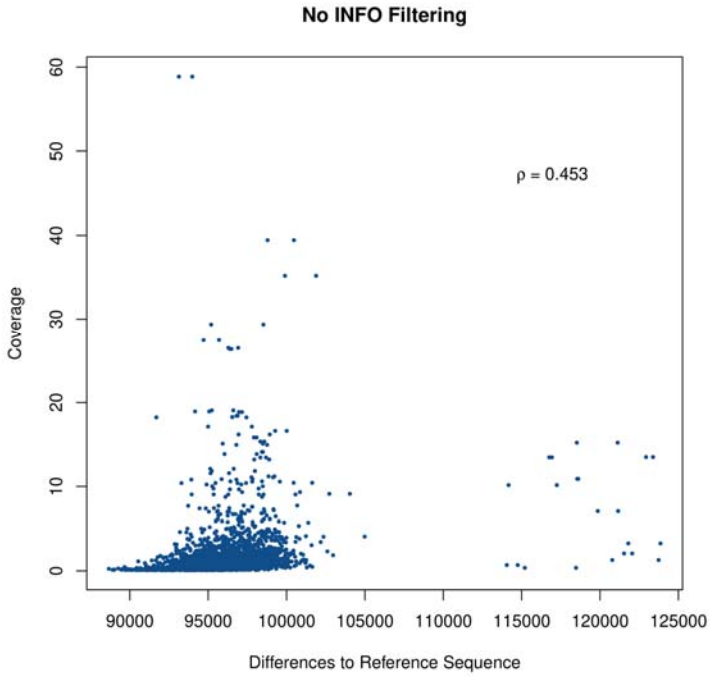
579 **Figure S2.3: Main parameters affecting imputation accuracy.** Imputation accuracy of
580 GLIMPSE across the 42 samples regarding, from left to right: (i) sequencing coverage; (ii)
581 effect of jointly imputing all 1.6K individual samples compared to imputing only the 42
582 downsampled genomes, (iii) imputation done with Beagle4.1.
583
584

585

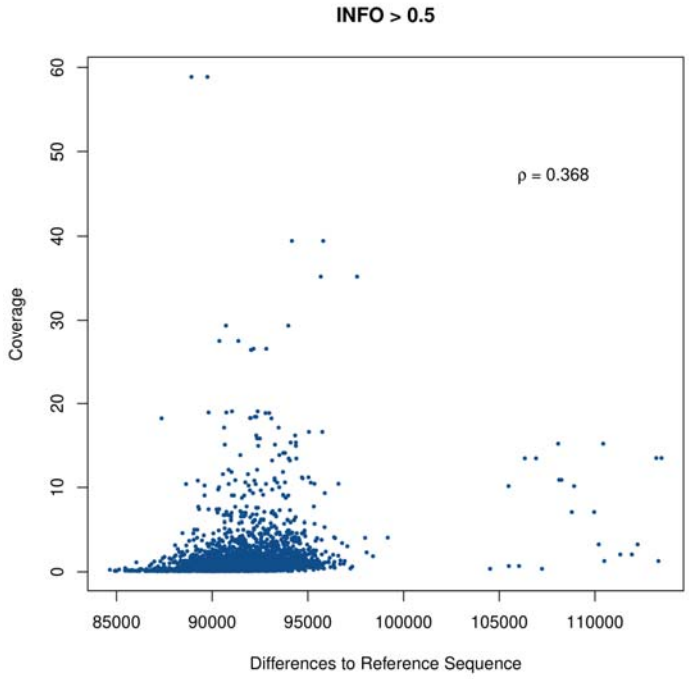
Figure S2.4: Imputation and phasing accuracy for the Koszyce trio. (A) Mendel error rate across the 22 autosomes. A Mendel error is counted when the parental and offspring genotypes violate Mendel transmission rules. (B) Same as before, excluding sites at which all three samples are REF/REF in the high coverage data. (C) Switch error rates averaged over the 3 samples. A switch error is counted between two consecutive heterozygous



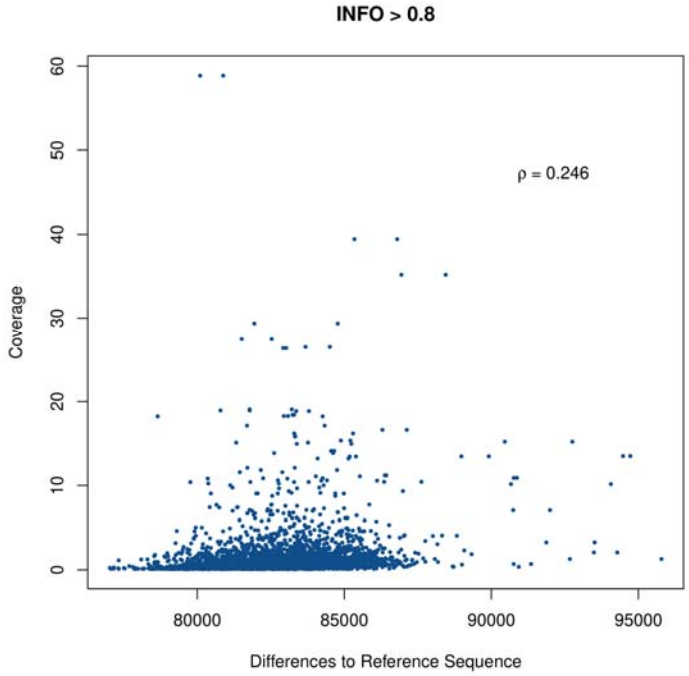
591 genotypes when the reported haplotypes are not consistent with those derived from the trio.



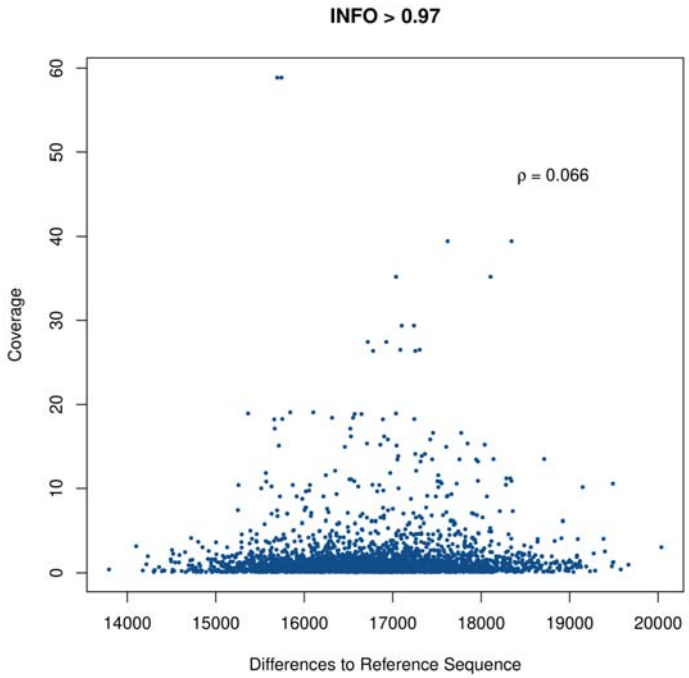
592 **Figure S2.5.** Total number of allele differences to the reference sequence for each of the
593 2,984 haplotypes against coverage, without INFO filtering.



594 **Figure S2.6.** Total number of allele differences to the reference sequence for each of the
595 2,984 haplotypes against coverage, with filtering for SNPs on INFO >0.5.



596 **Figure S2.7.** Total number of allele differences to the reference sequence for each of the
597 2,984 haplotypes against coverage, with filtering for SNPs on INFO >0.8.



598 **Figure S2.8.** Total number of allele differences to the reference sequence for each of the
599 2,984 haplotypes against coverage, with filtering for SNPs on INFO >0.97.

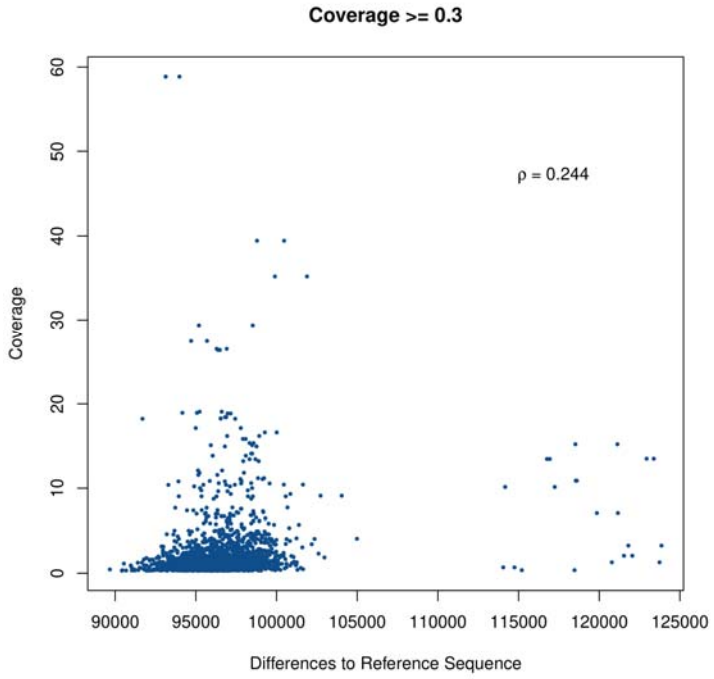


Figure S2.9. Total number of allele differences to the reference sequence for each of the 2,984 haplotypes against coverage, with filtering for samples on coverage ≥0.3.

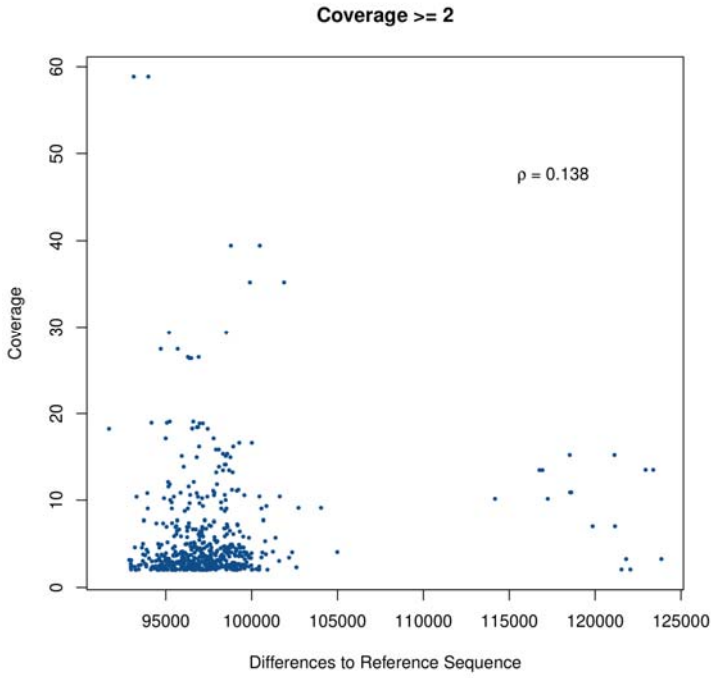


Figure S2.10. Total number of allele differences to the reference sequence for each of the 2,984 haplotypes against coverage, with filtering for samples on coverage ≥2.

605
606
607
608

SampleID	Country	Age (yBP)	MAF 1000G group	Coverage	Reference
atp16	Spain	4867 – 5212	EUR	13X	Günther et al. 2015[5]
Stuttgart	Germany	7020 – 7260	EUR	16X	Lazaridis et al. 2014[6]
Loschbour	Luxembourg	7940 – 8160	EUR	18X	Lazaridis et al. 2014[6]
Ballynahatty	Ireland	4970 – 5293	EUR	10X	Cassidy et al. 2016[7]
sf12	Sweden	8757 – 9033	EUR	59X	Günther et al. 2018[8]
NE1	Hungary	7021 – 7256	EUR	18X	Gamba et al. 2014[9]
Sunghir III	Russia	33031 – 35154	EUR	11X	Sikora et al. 2017[10]
Rathlin1	Ireland	3835 – 3976	EUR	11X	Cassidy et al. 2016[7]
SSG-A2	Iceland	950 – 1100	EUR	10X	Ebenesersdóttir et al. 2018[11]
HSJ-A1	Iceland	950 – 1080	EUR	29X	Ebenesersdóttir et al. 2018[11]
STT-A2	Iceland	950 – 1050	EUR	14X	Ebenesersdóttir et al. 2018[11]
VK1	Greenland	750 – 950	EUR	12X	Margaryan et al. 2020[12]
BR2	Hungary	3060 – 3220	EUR	18X	Gamba et al. 2014[9]
SZ15	Hungary	1346 – 1538	EUR	11X	Amorim et al. 2018[13]
SZ3	Hungary	1346 – 1538	EUR	11X	Amorim et al. 2018[13]
SZ4	Hungary	1347 – 1538	EUR	10X	Amorim et al. 2018[13]
SZ45	Hungary	1348 – 1538	EUR	10X	Amorim et al. 2018[13]
SZ43	Hungary	1349 – 1538	EUR	12X	Amorim et al. 2018[13]
SZ1	Hungary	1150 – 1350	EUR	11X	Amorim et al. 2018[13]
baa01	South Africa	1831 – 1986	AFR	14X	Schlebusch et al. 2017[14]
ela01	South Africa	453 – 533	AFR	13X	Schlebusch et al. 2017[14]
new01	South Africa	327 – 508	AFR	11X	Schlebusch et al. 2017[14]
I10871	Cameroon	7800 – 7970	AFR	15X	Lipson et al. 2020[15]
Mota	Ethiopia	4419 – 4525	AFR	10X	Gallego Llorente et al. 2015[16]

KK1	Georgia	9550 – 9890	EUR	12X	Broushaki et al. 2016[17]
WC1	Iran	9032 – 9405	EUR	10X	Broushaki et al. 2016[17]
BOT2016	Kazakhstan	5318 – 5582	EUR	14X	de Barros Damgaard et al. 2018[18]
Yamnaya Karagash	Kazakhstan	4837 – 4968	EUR	26X	de Barros Damgaard et al. 2018[18]
Andaman	India	30 – 150	SAS	17X	Moreno-Mayar et al. 2018[19]
Funadomari 23	Japan	3550 – 3960	EAS	39X	Kanzawa-Kiriyama et al. 2019[20]
Ust'-Ishim	Russia	42560 – 47480	ALL	35X	Fu et al. 2014[21]
Yana 1	Russia	30950 – 32950	ALL	27X	Sikora et al. 2019[22]
Kolyma 1	Russia	9665 – 9906	ALL	15X	Sikora et al. 2019[22]
USR1	USA	11270 – 11600	ALL	17X	Moreno-Mayar et al. 2018[23]
AHUR_2064	USA	10770 – 11170	AMR	19X	Moreno-Mayar et al. 2018[19]
Lovelock2	USA	1818 – 1942	AMR	15X	Moreno-Mayar et al. 2018[19]
Lovelock3	USA	567 – 687	AMR	19X	Moreno-Mayar et al. 2018[19]
Saqqaq	Greenland	3600 – 4170	AMR	13X	Rasmussen et al. 2010[24]
Clovis	USA	12572 – 12726	AMR	15X	Rasmussen et al. 2014[25]
Sumidouro5	Brazil	10258 – 10552	AMR	16X	Moreno-Mayar et al. 2018[19]
RISE1159 *	Poland	4840 – 4709	EUR	27X	Schroeder et al. 2019[3]
RISE1168 *	Poland	4840 – 4709	EUR	19X	Schroeder et al. 2019[3]
RISE1160 *	Poland	4840 – 4709	EUR	5X	Schroeder et al. 2019[3]

Table S2.1: Detailed list of high coverage ancient validation genomes. From left to right: (1) original sample ID, (2) country of origin, (3) estimated age, (4) best matching continental group in 1000 Genomes used to stratify imputation accuracy results and (5) original coverage from which down-sampling has been performed. *indicates the Neolithic Koszyce trio that was first published in Schroeder et al. [3] but now sequenced to higher depth in this study and used for imputation validation purposes. RISE1160 is not counted among the 42 high coverage genomes.

References

1. Browning, B. L. & Browning, S. R. Genotype Imputation with Millions of Reference Samples. *Am. J. Hum. Genet.* **98**, 116–126 (2016).

2. Rubinacci, S., Ribeiro, D. M., Hofmeister, R. J. & Delaneau, O. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nat. Genet.* **53**, 412 (2021).

3. Schroeder, H. *et al.* Unraveling ancestry, kinship, and violence in a Late Neolithic mass grave. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 10705–10710 (2019).

4. Bateson, W. *Mendel's Principles of Heredity*. (Cambridge University Press, 1902).

5. Günther, T. *et al.* Ancient genomes link early farmers from Atapuerca in Spain to modern-day Basques. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 11917–11922 (2015).

6. Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409–413 (2014).

7. Cassidy, L. M. *et al.* Neolithic and Bronze Age migration to Ireland and establishment of the insular Atlantic genome. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 368–373 (2016).

8. Günther, T. *et al.* Population genomics of Mesolithic Scandinavia: Investigating early postglacial migration routes and high-latitude adaptation. *PLoS Biol.* **16**, e2003703 (2018).

9. Gamba, C. *et al.* Genome flux and stasis in a five millennium transect of European prehistory. *Nat. Commun.* **5**, 5257 (2014).

10. Sikora, M. *et al.* Ancient genomes show social and reproductive behavior of early Upper Paleolithic foragers. *Science* **358**, 659–662 (2017).

11. Ebenesersdóttir, S. S. *et al.* Ancient genomes from Iceland reveal the making of a human population. *Science* **360**, 1028–1032 (2018).

12. Margaryan, A. *et al.* Population genomics of the Viking world. *Nature* **585**, 390–396 (2020).

13. Amorim, C. E. G. *et al.* Understanding 6th-century barbarian social

organization and migration through paleogenomics. *Nat. Commun.* **9**, 3547 (2018).

14. Schlebusch, C. M. *et al.* Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science* **358**, 652–655 (2017).

15. Lipson, M. *et al.* Ancient West African foragers in the context of African population history. *Nature* **577**, 665–670 (2020).

16. Gallego Llorente, M. *et al.* Ancient Ethiopian genome reveals extensive Eurasian admixture in Eastern Africa. *Science* **350**, 820–822 (2015).

17. Broushaki, F. *et al.* Early Neolithic genomes from the eastern Fertile Crescent. *Science* **353**, 499–503 (2016).

18. de Barros Damgaard, P. *et al.* The first horse herders and the impact of early Bronze Age steppe expansions into Asia. *Science* **360**, (2018).

19. Moreno-Mayar, J. V. *et al.* Early human dispersals within the Americas. *Science* **362**, (2018).

20. Kanzawa-Kiriyama, H. *et al.* Late Jomon male and female genome sequences from the Funadomari site in Hokkaido, Japan. *Anthropol. Sci.* **127**, (2019).

21. Fu, Q. *et al.* Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**, 445–449 (2014).

22. Sikora, M. *et al.* The population history of northeastern Siberia since the Pleistocene. *Nature* **570**, 182–188 (2019).

23. Moreno-Mayar, J. V. *et al.* Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans. *Nature* **553**, 203–207 (2018).

24. Rasmussen, M. *et al.* Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463**, 757–762 (2010).

25. Rasmussen, M., Anzick, S. L., Waters, M. R. & Skoglund, P. The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature* (2014).

3) Demographic inference

3a) Phylogenetic analysis of mtDNA sequences

Tharsika Vimala[1], Martin Sikora[1]

[1]Lundbeck Foundation GeoGenetics Centre, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark

Methods

We carried out a phylogenetic tree analysis using the reconstructed mitochondrial genomes from the human remains presented in this study. Only sequences with less than 2000 missing sites were included in the analysis to avoid biases caused by missing data. Sequences were aligned using *mafft* [1] and subsequently inputted to the maximum likelihood (ML) based phylogenetic tree inference tool *RAxML-ng* [2]. The analysis was carried out by using the 'all-in-one'- option performing both an ML search and bootstrapping (--all –bs-trees 100) along with the substitution model GTR+I+G4.

Results

From the resulting phylogenetic tree, we obtain an overview of how the remains are distributed across the haplogroups in our dataset. We find haplogroup U to be the most common haplogroup in the analysed set of individuals. Especially, subclade U5 is commonly observed among European hunter-gatherers. Focusing on the subclade U5a, we find the remains distributed into two main sub-haplogroups classified as U5a1 and U5a2. U5a1 is mainly influenced by Eastern European hunter-gatherers from Russia and Ukraine, while we also find the Scandinavian remains, NEO752 and NEO18, represented in this clade. In particular, the 9.8 kya old remains of NEO18 are interesting as the genetic structure analysis of the autosomes of NEO18 show evidence of Ukrainian hunter-gatherer ancestry (Figure 2, main). Haplogroup U5a2 shows a higher representation of Danish hunter-gatherers, specifically in subclade U5a2a, in which we also identify two Mesolithic Iberian individuals, NEO648 and NEO938. We likewise observe a Mesolithic Latvian individual (NEO307) within the U5a2d subclade, which is primarily dominated by Ukrainian and Russian hunter-gatherers (Figure S3a.1). These observations are congruent with the autosomal structure analysis displaying Ukrainian hunter-gatherer ancestry in these individuals. The clade

representing haplogroup U5b is mainly influenced by Danish hunter-gatherers along with a few Western European hunter-gatherers from Britain, France, and Iberia. We do, however, also observe a few Ukrainian hunter-gatherers clustering closely to the Danish hunter-gatherer individuals, which could indicate a continuous level of gene flow from Eastern Europe into Scandinavia. Additionally, we identify a single farmer individual, NEO597, carrying U5b, which is a rare example of genetic continuity of a hunter-gatherer associated haplogroup (Figure S3a.1). This contrasts the genetic transition otherwise observed with the arrival of the early farmers. We find a similar overall pattern within the genetic variation of U4, which is mainly influenced by Eastern European hunter-gatherers. Furthermore, we identify individuals with steppe-ancestry as well as two Danish hunter-gatherer remains clustering within the same clade. In the clade of haplogroup U2 we find a single Mesolithic Iberian carrying haplogroup U2'3'4'7'8'9, while the rest of the remains in the U2-clade belong to the sub-haplogroup U2e. U2e is carried by Eastern European hunter-gatherers, although we also identify a significant number of remains from the forest steppe clustering in U2e as well. The Danish Neolithic remains of NEO792 are interesting as this individual carried the highest proportion of steppe-ancestry among the Danish individuals. Haplogroup K, a descending haplogroup of U8, includes a combination of farmers and hunter-gatherers. Specifically, we find the haplogroup K1e to be carried by Danish hunter-gatherers, while K1a and K1b are mainly influenced by Neolithic individuals from Scandinavia (Figure S3a. 2). The highest frequency of early farmers is found within the genetic variation of haplogroup H, a descending haplogroup of HV. Both HV and JT are mainly influenced by Western European farmers (Figure S3a.3).

Conclusion

In overall we find most of the Scandinavian hunter-gatherers clustering within the variation of haplogroup U. Given the high number of human remains represented, we were able to obtain a phylogeny of a relatively high resolution of this particular haplogroup. Our results show evidence of a continuous migration of especially Eastern hunter-gatherers into Scandinavia. Most of the early farmers carried haplogroup H or fell within haplogroup JT, while a few farmers carried haplogroup K. We find individuals with steppe ancestry mainly clustering together under macro haplogroup M, although we also identify a few steppe individuals in U2 closely related to a Danish Neolithic individual, NEO792, who also carried a high proportion of steppe-ancestry in the nuclear genome.

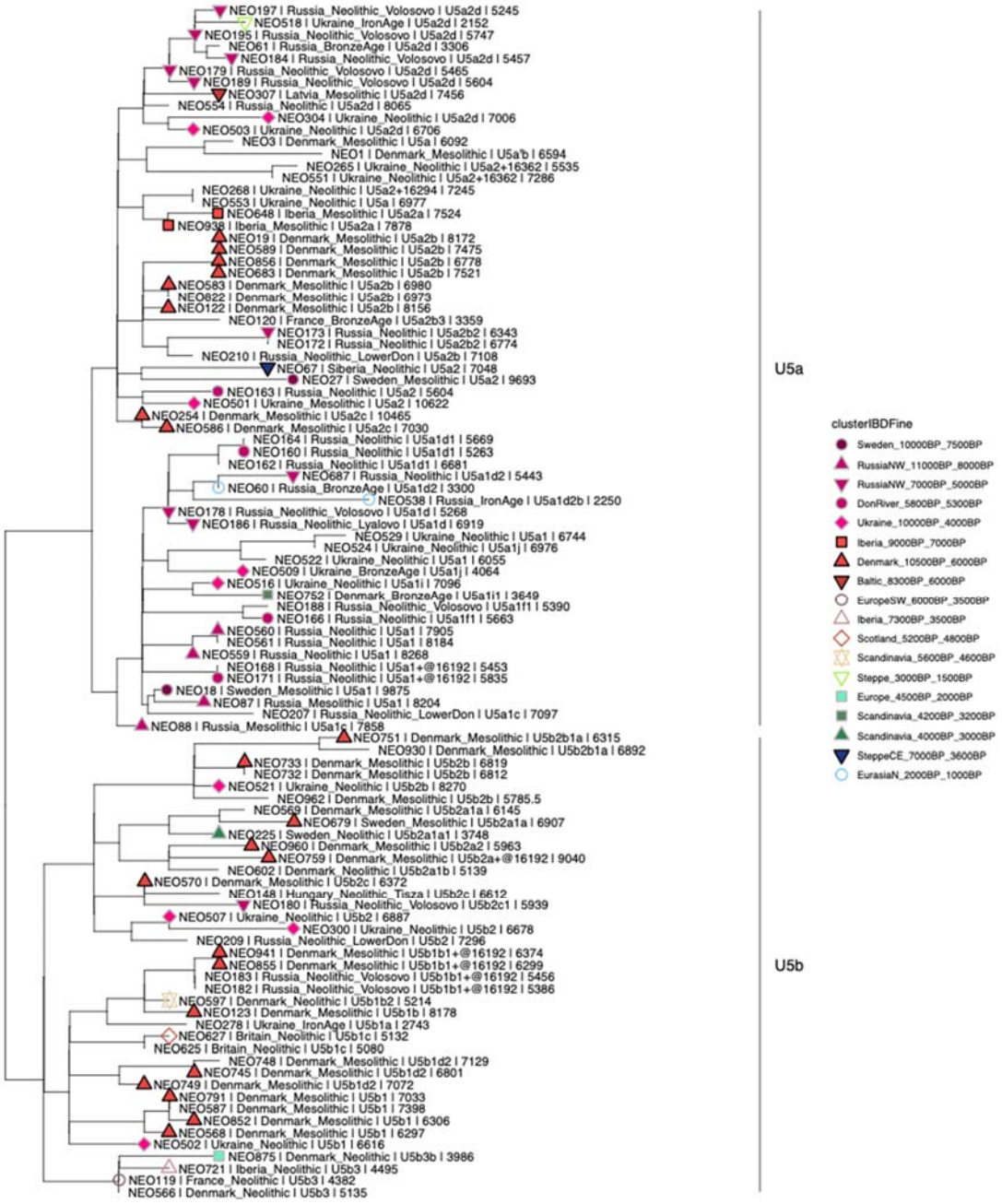


Figure S3a.1. Maximum likelihood tree of haplogroup U5

Maximum likelihood tree displaying the phylogenetic relationship between the human remains carrying haplogroup U5 and more specifically the two subclades U5a and U5b. Labels include information on sample ID, group ID, haplogroup, and age of the respective remains. Symbols indicate the specific fine scale IBD cluster listed in the legend.

0

Figure S3a.2. Maximum likelihood tree of haplogroup U excluding U5

Maximum likelihood tree displaying the phylogenetic relationship between the human remains carrying a descending haplogroup of U with the exception of U5. Labels include information on sample ID, group ID, haplogroup, and age of the respective remains. Symbols indicate the specific fine scale IBD cluster listed in the legend.

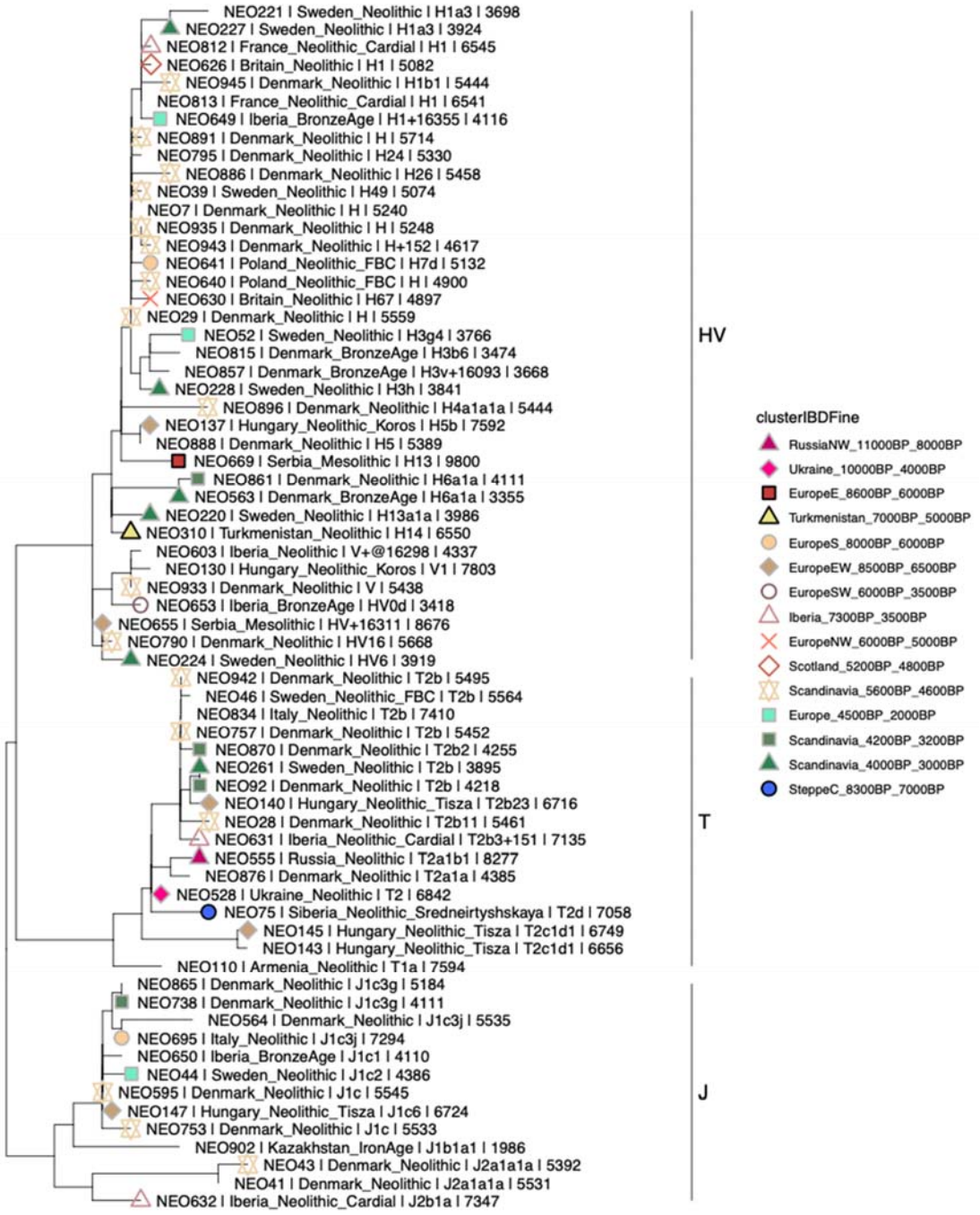


Figure S3a.3. Maximum likelihood tree of haplogroup farmer-associated haplogroups

Maximum likelihood tree displaying the phylogenetic relationship between the human remains carrying an HV or JT descending haplogroup. Labels include information on sample ID, group ID, haplogroup, and age of the respective remains. Symbols indicate the specific fine scale IBD cluster listed in the legend.

References

1. Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, Takashi Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, *Nucleic Acids Research*, Volume 30, Issue 14, 15 July 2002, Pages 3059–3066, https://doi.org/10.1093/nar/gkf436
2. Alexey M Kozlov, Diego Darriba, Tomáš Flouri, Benoit Morel, Alexandros Stamatakis, RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference, *Bioinformatics*, Volume 35, Issue 21, 1 November 2019, Pages 4453–4455, https://doi.org/10.1093/bioinformatics/btz305

3b) Y chromosome / Sex determination

Martin Sikora[1]

[1]Lundbeck Foundation GeoGenetics Centre, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark

Methods

We calculated the ratio of reads aligning to either of the sex chromosomes (R_Y statistic)[1] to determine genetic sex of the study individuals. Y chromosomes of inferred male individuals were further analysed using phylogenetic placement[2]. We built a reference phylogenetic tree of 1,244 male individuals from the 1000 Genomes project with *RAxML-NG*[3], using the general time-reversible model including among-site rate heterogeneity and ascertainment correction (model GTR+G+ASC_LEWIS). For each ancient sample, haploid genotypes given the positions and alleles in the reference panel were called using '*bcftools call'* (options *'-C alleles –ploidy 1 -i*'). The resulting genotypes were converted to fasta format and placed onto

the reference tree using *EPA-ng*[2]. Phylogenetic placements were processed and visualised using *gappa*[4].

To convert phylogenetic placements into haplogroup calls, we assigned each branch of the reference phylogeny to its representing haplogroup, using SNP annotations from ISOGG (version 15.73). For each ancient sample, haplogroups were then called using the most basal branch accumulating 99% of the placement weights, obtained using the '*accumulate'* command in *gappa.*

For in-depth phylogenetic analyses of haplogroups I, R1, and Q, we compiled extended reference panels of high-coverage modern individuals belonging to those haplogroups from publicly available sources[5–8]. To increase resolution for the placement of ancient samples, we also included ancient individuals with Y chromosome coverage ≥1.5X in the reference panels. For each haplogroup panel, we called haploid genotypes individually per sample using '*bcftools call'*, setting genotypes with read depth < 2 or quality score <30 to missing. Individual VCF files were then merged and filtered to retain only biallelic SNPs polymorphic in the reference panel. For each haplogroup reference panel, we built phylogenetic trees using *RAxML-NG* and performed phylogenetic placement as described above, restricting to target samples with >0.1X coverage.

Results

Sex determination

We unambiguously determined genetic sex for all 317 study individuals (118 female, 199 males; Supplementary Table VII). In a plot of normalised sequencing depth across the X and Y chromosomes, the final dataset individuals form two clearly separated clusters corresponding to XX and XY karyotypes (Fig. S3b.1). The exception is individual YGS-B-2, an Icelandic Viking Age individual previously found to carry an XXY karyotype[9].

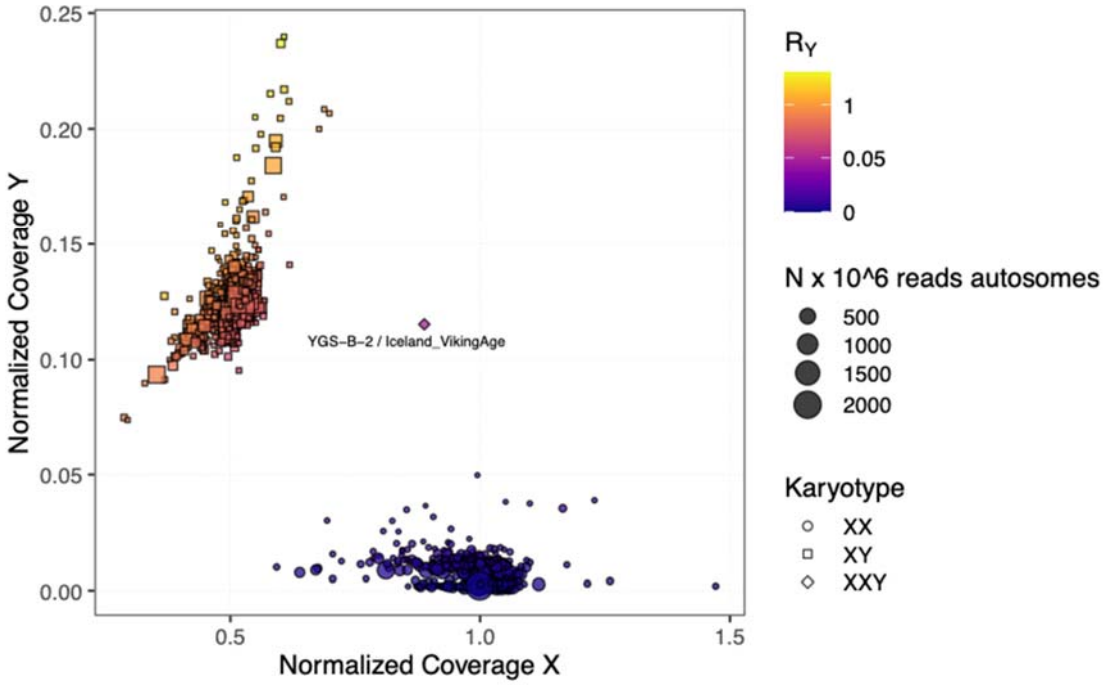


Fig. S3b.1. Sex determination. Plot shows coverage on X and Y chromosomes normalised by autosomal coverage, for each individual. Symbol color indicates R_Y ration values, and shape the inferred sex chromosome karyotype. Total number of autosomal reads are indicated by symbol size.

Phylogenetic placement

We used phylogenetic placement to analyse Y chromosome diversity in our dataset. For each ancient sample we obtain a distribution of placement weights across the reference phylogeny, hereby incorporating uncertainty in the placement due to low coverage. The placements can be subjected to analyses such as grafting as a pendant edge to the most likely placement, or clustering of multiple samples. As Y chromosome haplogroups are labels for clades of the phylogeny descending from specific ancestral branches, we can convert the placements into haplogroup calls for a specific sample by assigning haplogroup labels to each branch in the reference phylogeny, and finding the most basal branch that accumulates placement weights up to a specified threshold for the sample. We chose a conservative threshold of 0.99 for the weight accumulation; lower thresholds result in more derived haplogroup calls but with potentially higher uncertainty. Fig. S3b.2 gives an example of this approach for NEO962, a Mesolithic individual from Denmark with low coverage of 0.036X.

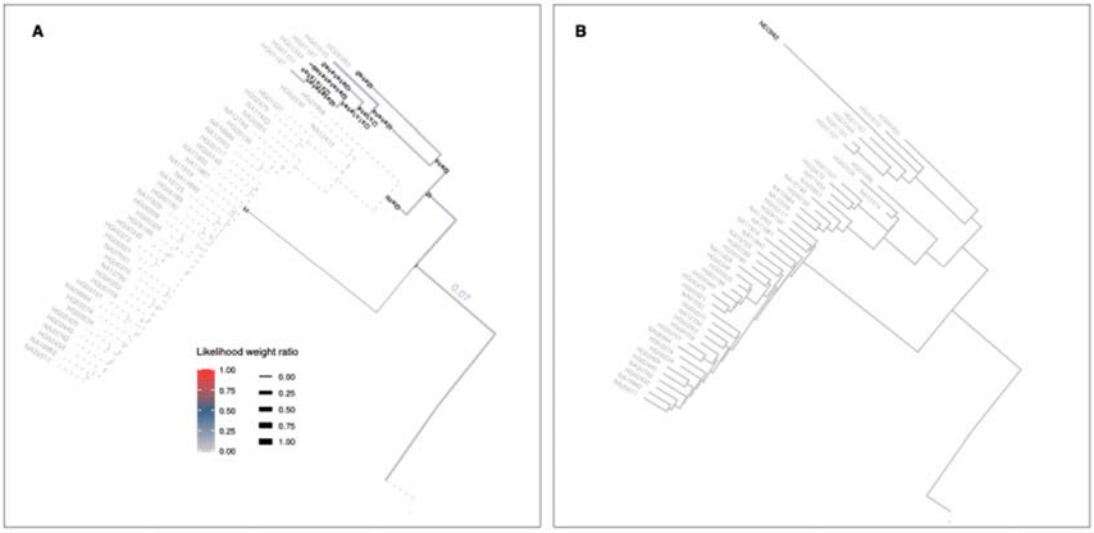


Fig. S3b.2. Phylogenetic placement. Plot showing phylogenetic placement weights (A) and graft tree with most likely placement (B) for individual NEO962 on a subtree representing reference individuals with haplogroup I. (A) Weights for individual branches are indicated with edge colour and width, edges without placements are indicated with dashed line. While the majority of the placement weight mass is distributed among branches of haplogroup I2, non-zero weights are also found on branches ancestral to I1 (0.02) and I (0.07). The individual is hence conservatively assigned to haplogroup I. (B) The most likely branching of NEO962 was found within the subclade I2a1a1a, albeit this placement is associated with considerable uncertainty.

Sub-haplogroup analyses of newly reported samples

Haplogroup I

Haplogroup I2 was common among newly reported samples from western Eurasian hunter-gatherer contexts, as well as later Neolithic farmers. In particular, the 25 Danish Mesolithic male individuals were exclusively carriers of haplogroup I2, albeit with considerable diversity across different sub-haplogroups (Fig. S3b.3). Neolithic farmer individuals from Scandinavia were predominantly placed within an ancient-only subclade of haplogroup I2a1a2, containing other individuals from Neolithic farmer contexts across Europe.

The earliest presence of haplogroup I1, which is the most common haplogroup among present-day Scandinavians, was found ~4,000BP among late Neolithic and early Bronze Age Scandinavians newly reported in this study (Fig. S3b.3). A single Swedish Mesolithic

individual (sf11) was placed at the base of the I1 clade; however, its low coverage (0.1X) precludes to conclude with certainty whether early I1-related lineages were indeed present among Scandinavian hunter-gatherers.

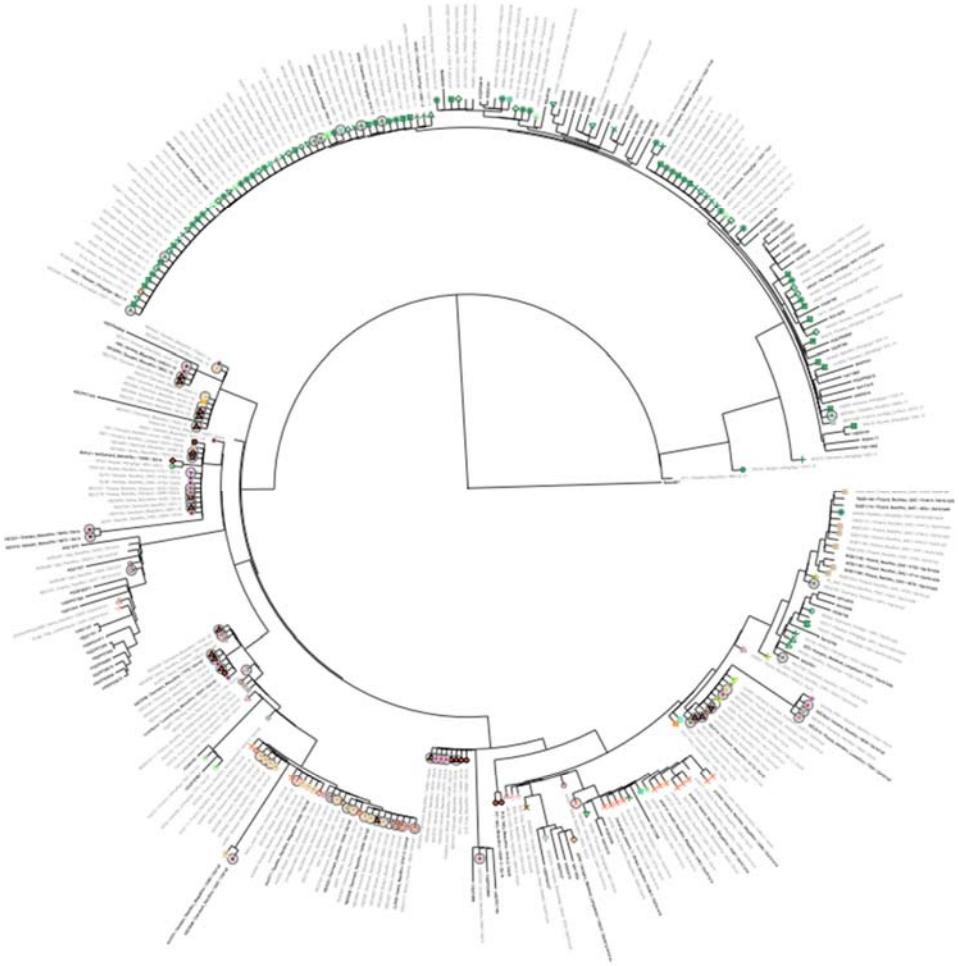


Fig. S3b.3. Phylogeny of haplogroup I. Phylogenetic tree with most likely placements of ancient samples. Samples labelled with black colour were used to infer the reference tree, whereas samples with grey labels were grafted from phylogenetic placement. Terminal branches for ancient samples were shortened to aid visualisation. Symbol colours and shapes indicate genetic clusters from IBD-based clustering. Newly reported individuals are highlighted with circled symbols.

Haplogroup Q

Haplogroup Q1 was common among newly reported Neolithic hunter-gatherer individuals from the Siberian Forest steppe and the Lake Baikal region (Fig. S3b. 4). We observed haplogroup Q1b2, rare among ancient West Eurasians, in two Ukrainian hunter-gatherers (NEO501, NEO516) as well as two Danish Neolithic farmer individuals (NEO599, NEO744).

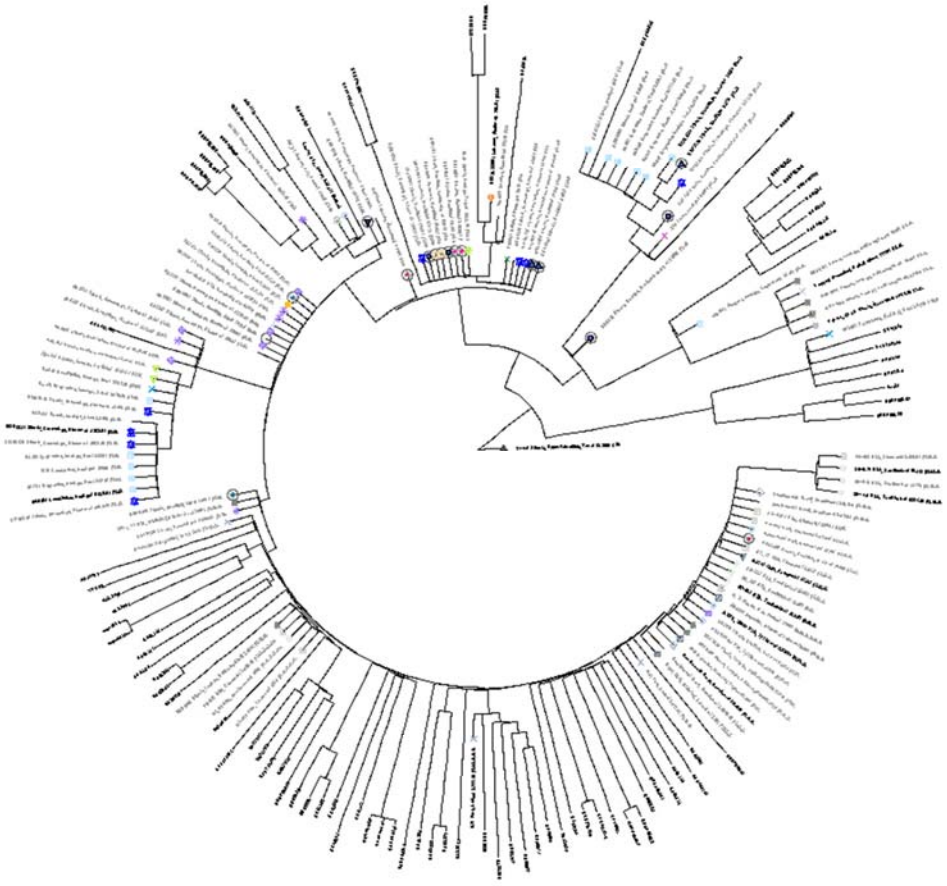


Fig. S3b.4. Phylogeny of haplogroup Q. Phylogenetic tree with most likely placements of ancient samples. Samples labelled with black colour were used to infer the reference tree, whereas samples with grey labels were grafted from phylogenetic placement. Terminal branches for ancient samples were shortened to aid visualisation. Symbol colours and shapes indicate genetic clusters from IBD-based clustering. Newly reported individuals are highlighted with circled symbols.

Haplogroup R1a

Haplogroup R1a was found in the newly reported samples mainly among Eastern European hunter-gatherer individuals. Phylogenetic placement suggests that the oldest individuals from Mesolithic and Neolithic Russia represent early diverging lineages (Fig. S3b.5). Notably, a ~7,300-year-old Neolithic individual from the Middle Don region (NEO113) was placed in a basal R1a clade together with early individuals associated with the Corded Ware complex (poz81, RISE446), which would make it the earliest observation of this lineage reported to date.

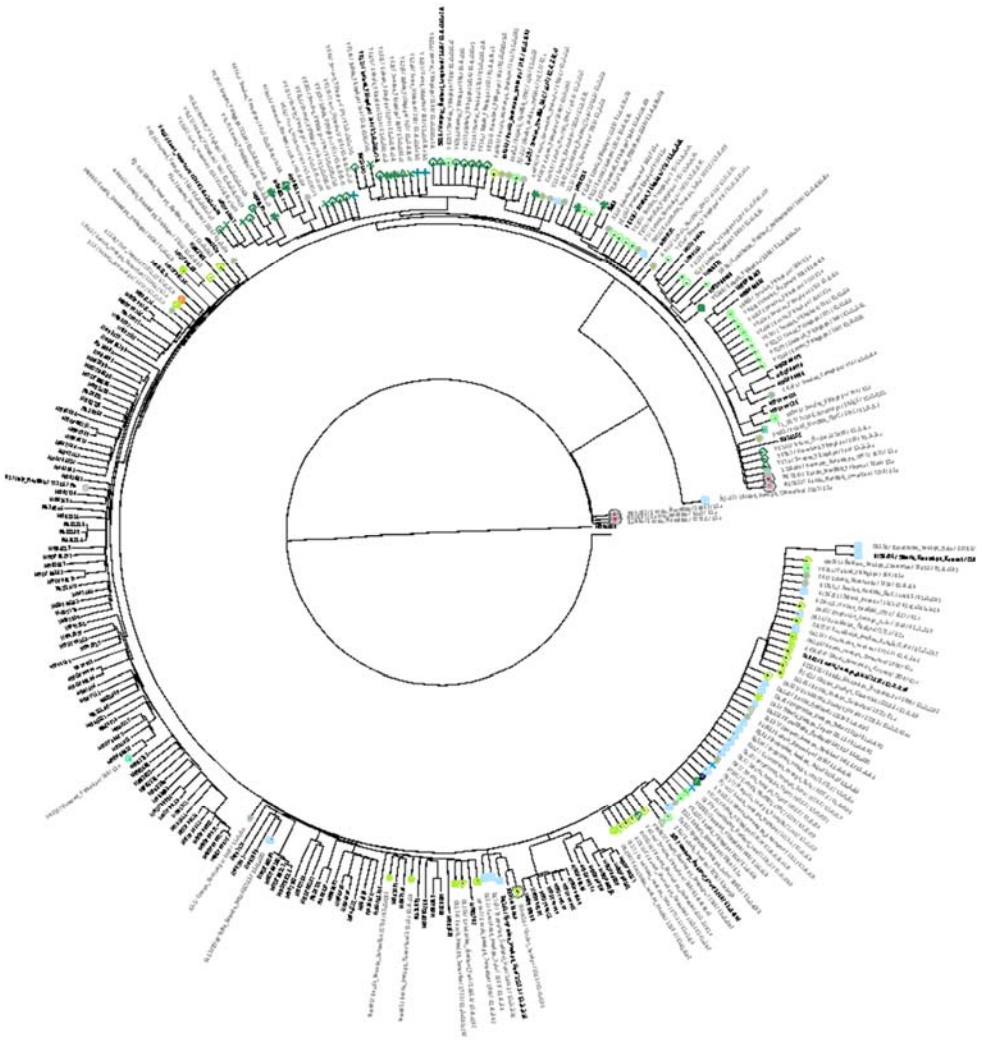


Fig. S3b.5. Phylogeny of haplogroup R1a. Phylogenetic tree with most likely placements of ancient samples. Samples labelled with black colour were used to infer the reference tree, whereas samples with grey labels were grafted from phylogenetic placement. Terminal branches for ancient samples were shortened to aid visualisation. Symbol colours and shapes indicate genetic clusters from IBD-based clustering. Newly reported individuals are highlighted with circled symbols.

Haplogroup R1b

Newly reported samples belonging to haplogroup R1b were distributed between two distinct groups depending on whether they formed part of the major European subclade R1b1a1b (R1b-M269). Individuals placed outside this subclade were predominantly from Eastern European Mesolithic and Neolithic contexts, and formed part of rare early diverging R1b lineages (Fig. S3b.6). Two Ukrainian individuals belonged to a subclade of R1b1b (R1b-V88) found among present-day Central and North Africans, lending further support[5,10] to an ancient Eastern European origin for this clade. Haplogroup R1b1a1a (R1b-M73) was frequent among Russian Neolithic individuals.

Individuals placed within the R1b-M269 clade on the other hand were from Scandinavian Late Neolithic and early Bronze Age contexts (Fig. S3b.6). Interestingly, more fine-scale sub-haplogroup placements of those individuals revealed that Y chromosome lineages distinguished samples from distinct genetic clusters inferred from autosomal IBD sharing (Fig. S3b.6, S3b.7). In particular, individuals associated with the Scandinavian cluster *Scandinavia_4200BP_3200BP* were all placed within the sub-haplogroup R1b1a1b1a1a1 (R1b-U106), whereas the two Scandinavian males associated with the Western European cluster *Europe_4500BP_2000BP* were placed within R1b1a1b1a1a2 (R1b-P312) (Fig. S3b.7).

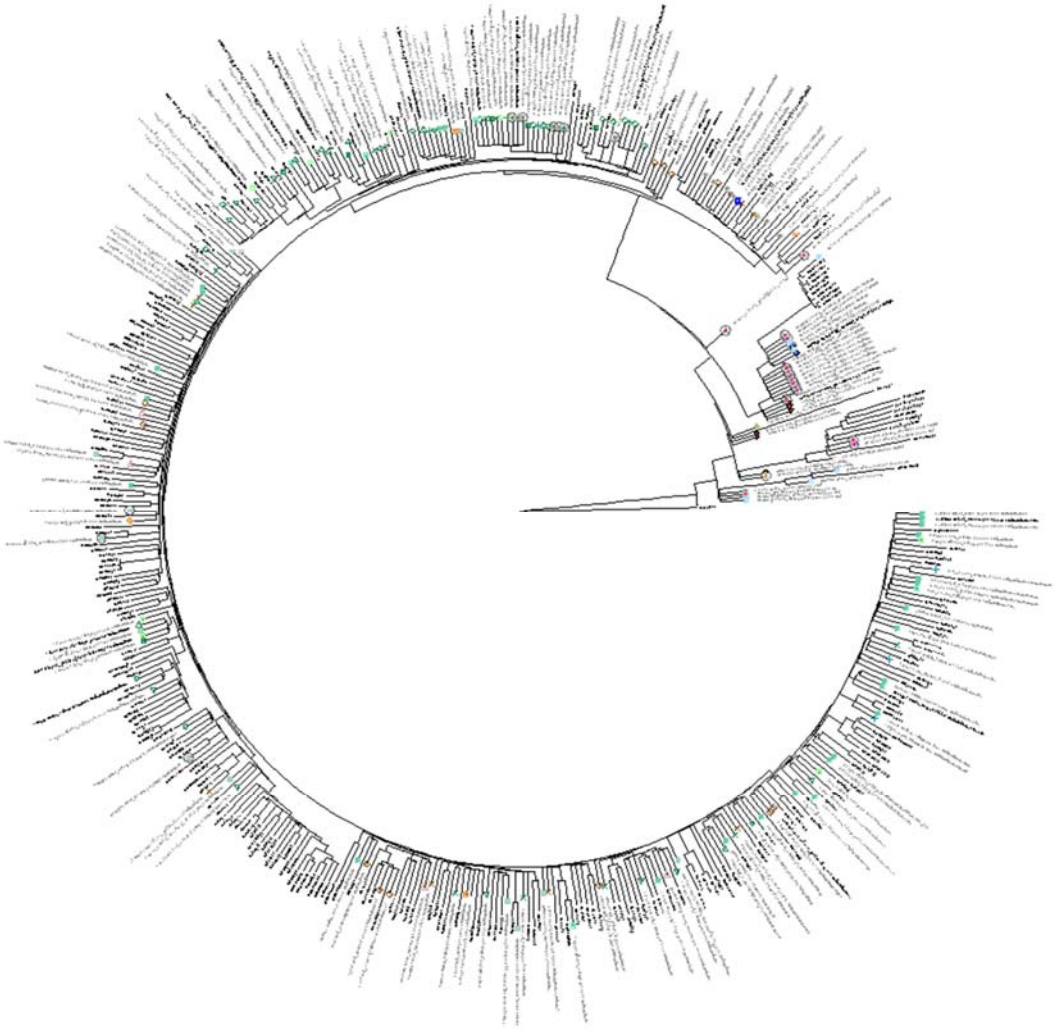


Fig. S3b.6. Phylogeny of haplogroup R1b. Phylogenetic tree with most likely placements of ancient samples. Samples labelled with black colour were used to infer the reference tree, whereas samples with grey labels were grafted from phylogenetic placement. Terminal branches for ancient samples were shortened to aid visualisation. Symbol colours and shapes indicate genetic clusters from IBD-based clustering. Newly reported individuals are highlighted with circled symbols.

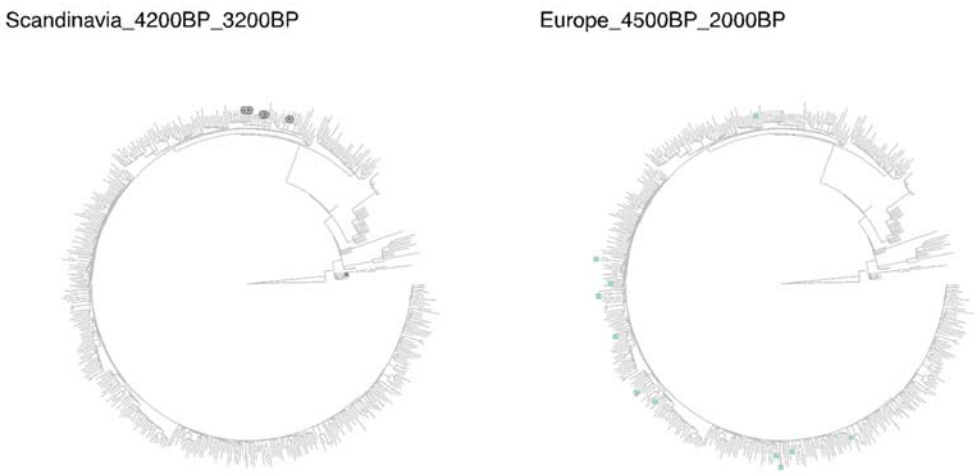


Fig. S3b.7. Phylogeny of haplogroup R1b for genetic clusters. Phylogenetic trees showing most likely placements of ancient samples from Danish Late Neolithic and Bronze Age genetic clusters Scandinavia_4200BP_3200BP (left) and Europe_4500BP_2000BP (right). Terminal branches for ancient samples were shortened to aid visualisation. Symbol colours and shapes indicate genetic clusters from IBD-based clustering. Newly reported individuals are highlighted with circled symbols.

References

1. Skoglund, P., Storå, J., Götherström, A. & Jakobsson, M. Accurate sex identification of ancient human remains using DNA shotgun sequencing. *J. Archaeol. Sci.* **40**, 4477–4482 (2013).
2. Barbera, P. *et al.* EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences. *Syst. Biol.* **68**, 365–369 (2019).
3. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455 (2019).
4. Czech, L., Barbera, P. & Stamatakis, A. Genesis and Gappa: processing, analyzing and visualizing phylogenetic (placement) data. *Bioinformatics* **36**, 3263–3265 (2020).
5. Haber, M. *et al.* Chad Genetic Diversity Reveals an African History Marked by Multiple Holocene Eurasian Migrations. *Am. J. Hum. Genet.* **0**, (2016).
6. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
7. Bergström, A. *et al.* Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**, (2020).
8. Byrska-Bishop, M. *et al. High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios*. 2021.02.06.430068 https://www.biorxiv.org/content/10.1101/2021.02.06.430068v1 (2021) doi:10.1101/2021.02.06.430068.
9. Ebenesersdóttir, S. S. *et al.* Ancient genomes from Iceland reveal the making of a human population. *Science* **360**, 1028–1032 (2018).
10. Marcus, J. H. *et al.* Genetic history from the Middle Neolithic to present on the Mediterranean island of Sardinia. *Nat. Commun.* **11**, 1–14 (2020).

3c) Relatedness

Martin Sikora[1]

[1]Lundbeck Foundation GeoGenetics Centre, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark

Methods

To infer genetic relatedness between the study individuals we used the allele-frequency free inference method introduced by Waples et al.[1]. For each pair of individuals, we calculated the three relatedness estimators R0, R1 and KING-robust[2] using the site-frequency-spectrum (SFS)-based approach. We used the *realSFS*[3] method implemented in the *ANGSD*[4] package to infer the 2D-SFS, selecting the SFS with the highest likelihood across ten replicates. We used a set of 1,191,529 autosomal transversion SNPs with minor allele frequency ≥ 0.05 from the 1000 Genomes Project[5] for the analysis.

We used previously established cut-offs[2] for the KING-robust estimator to assign individual pairs to first-, second- or third-degree relationships. Parent-offspring relationships were distinguished from sibling relationships using R0 and R1 ratios, by requiring that R0 ≤ 0.02 and 0.4 ≤ R1 ≤ 0.6 to infer a parent-offspring relative pair. We excluded individual pairs with less than 20,000 sites contributing to the estimators.

Results

We detected a total of 92 close relative pairs among the 1,664 dataset individuals, including 24 parent-offspring pairs, 36 siblings and 30 2nd degree pairs (Fig S3c.1, Supplementary Table VI). We further found evidence of two duplicate / monozygotic twin relationships. Sample NEO70 presented in this study was inferred to be from the same individual as RISE554 previously reported in Allentoft *et al*[6]. Additionally, two male individuals MJ-15 and MJ-32 reported in Järve *et al*[7] were also inferred as duplicate/twin pairs. In both cases genetic sex, mitochondrial as well as Y chromosome haplogroups were all consistent with their inferred relatedness status.

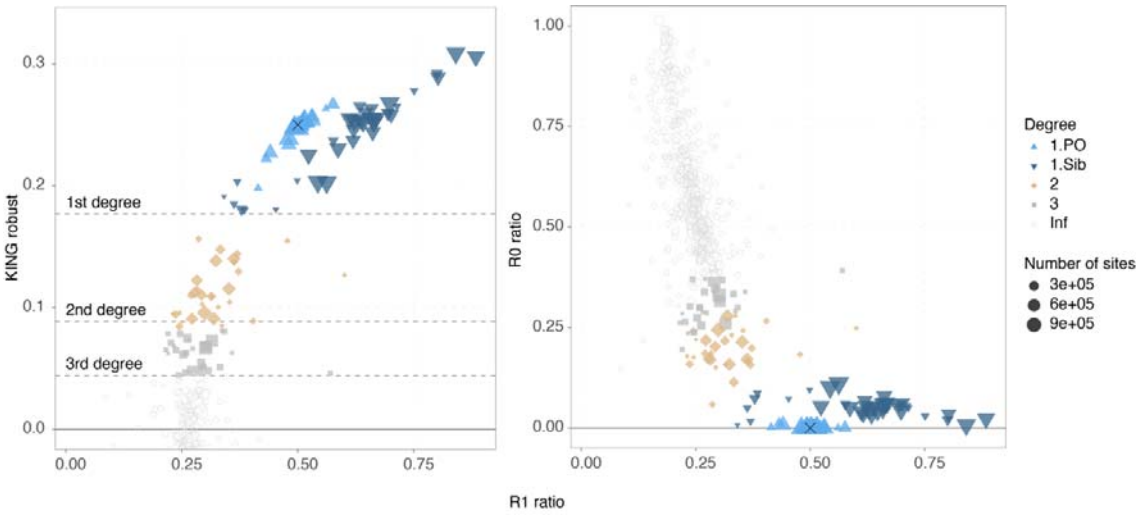


Fig. S3c.1. Relatedness inference. Plots showing relatedness estimators R1/KING-robust (left) and R1/R2 (right) for pairs of individuals. Inferred relatedness status for each pair is indicated by plot symbol and colour, with symbol size scaled by the total number of informative sites. Black crosses indicate expected values for parent-offspring relationships.

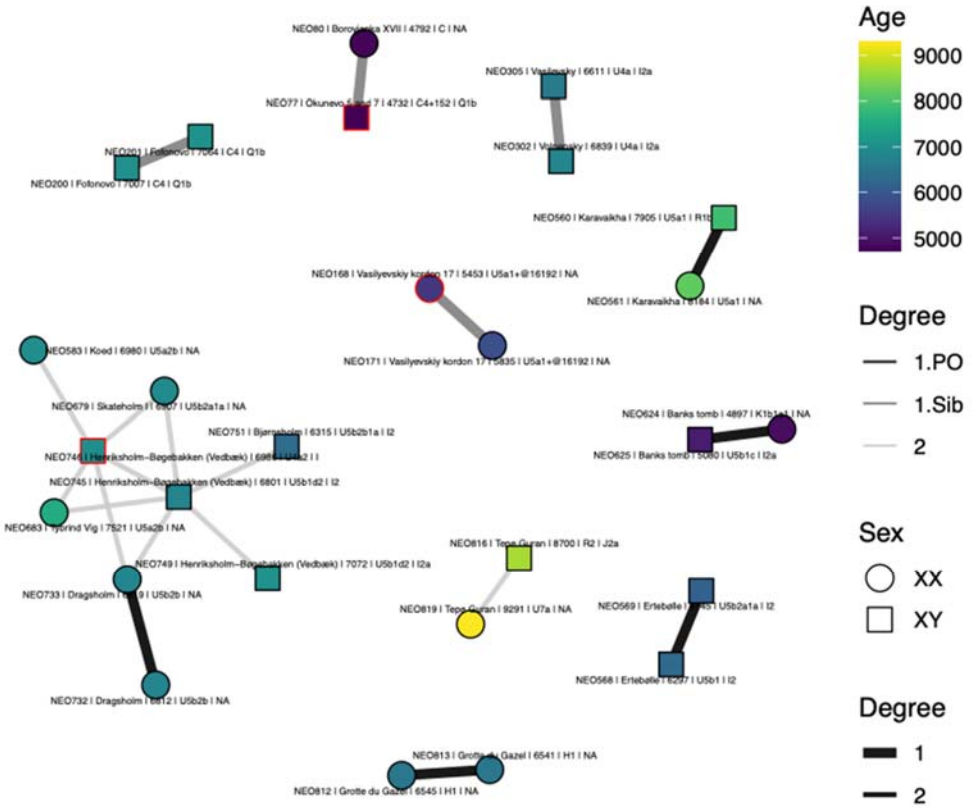


Fig. S3c.2. Relatedness among newly reported individuals. Network showing first and second-degree relationships, indicated by edge width and colour. Age of individuals is indicated by fill colour, and individuals are labelled with site name, age, mitochondrial and Y chromosome haplogroups. Individuals flagged for possible contamination are indicated in red.

We identified a total of 10 first- and 12 second-degree relative pairs among the newly reported individuals. However, inspection of the relatedness network revealed that the majority of 2nd degree connections are between Danish Mesolithic individuals from different sites and two individuals from Henriksholm (NEO745, NEO746), one of which was flagged as contaminated (Fig. S3c.2). As excess heterozygosity due to contamination can lead to artificially increased relatedness estimates, we excluded any pair involving those two individuals, as well as two other pairs involving contaminated individuals from the final list of close relatives (Table S3c.1). Finally, three individuals reported here were inferred to be either the same individual (2) or close relatives (1) of samples previously published using targeted SNP capture (Table S3c.2).

Individual 1	Individual 2	Site	Country	Degree	Notes
NEO568	NEO569	Ertebølle	Denmark	1.PO	NEO568 ("Ertebølle man") is the father of infant NEO569
NEO732	NEO733	Dragsholm	Denmark	1.PO	Mother-daughter relationship, direction unknown
NEO624	NEO625	Banks tomb	UK	1.PO	NEO625 is the father of juvenile NEO624
NEO813	NEO812	Grotte du Gazel	France	1.PO	NEO813 is the mother of infant NEO812
NEO560	NEO561	Karavaikha	Russia	1.PO	NEO561 likely the mother of NEO560 (age and MT haplogroup)
NEO201	NEO200	Fofonovo	Russia	1.Sib	
NEO302	NEO305	Volnensky / Vasilevsky	Ukraine	1.Sib	
NEO816	NEO819	Tepe Guran	Iran	2	

Table S3c.1. Close relatives among newly reported individuals.

Study Individual	Related individual	Publication	Relationship

NEO73	I1960	Narasimhan et al 2019 Science	same individual
NEO669	I5407	Mathieson et al 2018 Nature	same individual
NEO60	BOO005	Lamnidis et al 2018 Nature Communications	first degree, infant NEO60 likely daughter of BOO005

Table S3c.2. Study individuals with related published individuals.

References

1. Waples, R. K., Albrechtsen, A. & Moltke, I. Allele frequency-free inference of close familial relationships from genotypes or low-depth sequencing data. *Mol. Ecol.* **28**, 35–48 (2019).
2. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
3. Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y. & Wang, J. SNP Calling, Genotype Calling, and Sample Allele Frequency Estimation from New-Generation Sequencing Data. *PLoS ONE* **7**, e37558 (2012).
4. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15**, 356 (2014).
5. The 1000 Genomes Project Consortium,. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
6. Allentoft, M. E. *et al.* Population genomics of Bronze Age Eurasia. *Nature* **522**, 167-+ (2015).
7. Järve, M. *et al.* Shifts in the Genetic Landscape of the Western Eurasian Steppe Associated with the Beginning and End of the Scythian Dominance. *Curr. Biol.* **29**, 2430-2441.e10 (2019).

3d) Pop structure general, PCA/Admixture (Martin)

Martin Sikora[1]

[1]Lundbeck Foundation GeoGenetics Centre, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark

Methods

We generated a dataset for population genetic analysis by combining the 317 newly sequenced individuals with 1,347 previously published ancient genomes with genomic

coverage >0.1X generated using shotgun sequencing (Supplementary Table VII). Imputed genotype data (Supplementary note S2) for this set of 1,664 ancient genomes was merged with genotypes of 2,504 modern individuals from the 1,000 Genomes project[1] used as a reference panel in the imputation. We retained only SNPs passing the 1000 Genomes strict mask, resulting in a final dataset of 4,168 individuals genotyped at 7,321,965 autosomal SNPs ("1000G" dataset). In addition to imputed genotypes, we also generated pseudo-haploid genotypes for each ancient individual by randomly sampling an allele from sequencing reads covering those SNPs. For population structure analyses in the context of global genetic diversity, we generated a second dataset by intersecting the ancient genotype data with SNP array data of 2,180 modern individuals from 213 world-wide populations[2–5] ("HO" dataset).

To facilitate filtering for downstream analyses, we flagged individuals to potentially exclude based on the following criteria:

- Contamination estimate >5% ("*contMT5pct*", "*contNuc5pct*"; Supplementary note S1)
- Autosomal coverage < 0.1X ("*lowcov*")
- Genome-wide average imputation genotype probability < 0.98 (*"lowGpAvg")*
- Individual is the lower quality sample in a close relative pair (*"1d_rel"*, *"2d_rel";* Supplementary note S3c)

A total of 1,492 individuals (213 newly reported) passed all filters, which were used in the majority of downstream analyses unless otherwise noted.

We investigated overall population structure among the dataset individuals using principal component analyses (PCA) and model-based clustering (ADMIXTURE[6]). We carried out PCA using different subsets of individuals in the "HO" dataset. For the PCA including only imputed diploid samples, we used GCTA[7], excluding SNPs with minor allele frequency (MAF) < 0.05 in the respective panel. For PCA projecting low coverage or flagged individuals, we used *smartpca*[8] with options '*lsqproject: YES*' and '*autoshrink: YES*' on a fixed set of 400,186 SNPs with MAF ≥ 0.05 in non-African individuals passing all filters.

We ran *ADMIXTURE* on a set of 1,593 ancient individuals from the "1000G" dataset, excluding individuals flagged as close relatives or a contamination estimate >5%. For the 1,492 individuals passing all filters we used imputed genotypes, the remaining 101 lower coverage samples were represented by pseudo-haploid genotypes. We restricted the analysis to transversion SNPs with imputation INFO score ≥ 0.8 and MAF ≥ 0.05. We further

performed linkage disequilibrium (LD) pruning and filtering for missingness using *plink*[9] (options '*--indep-pairwise 500 50 0.4 –geno 0.8'*), for a final analysis set of 142,550 SNPs.

Results

Comparison of imputed and pseudo-haploid genotypes in PCA space

To determine the consistency of imputed and pseudo-haploid genotypes when used in PCA, we followed the approach of Antonio *et al.*[10] comparing the coordinates of both sets of genotypes for each individual when projected onto principal components inferred from modern individuals (Fig. S3d.1, S3d.3). We did not use the "*autoshrink*" option of *smartpca* for this analysis to avoid possible systematic differences in the projection correction between the two sets of genotypes. Projected PCA positions for samples passing all filters were consistent between imputed and pseudo-haploid genotypes, with no evidence for systematic shifts (Fig. S3d.2, S3d.4) and only a very subtle relationship of PCA distance between genotypes with genomic coverage (Fig. S3d.3, S3d.6). More substantial shifts were only observed with low coverage (<0.1X) flagged samples.

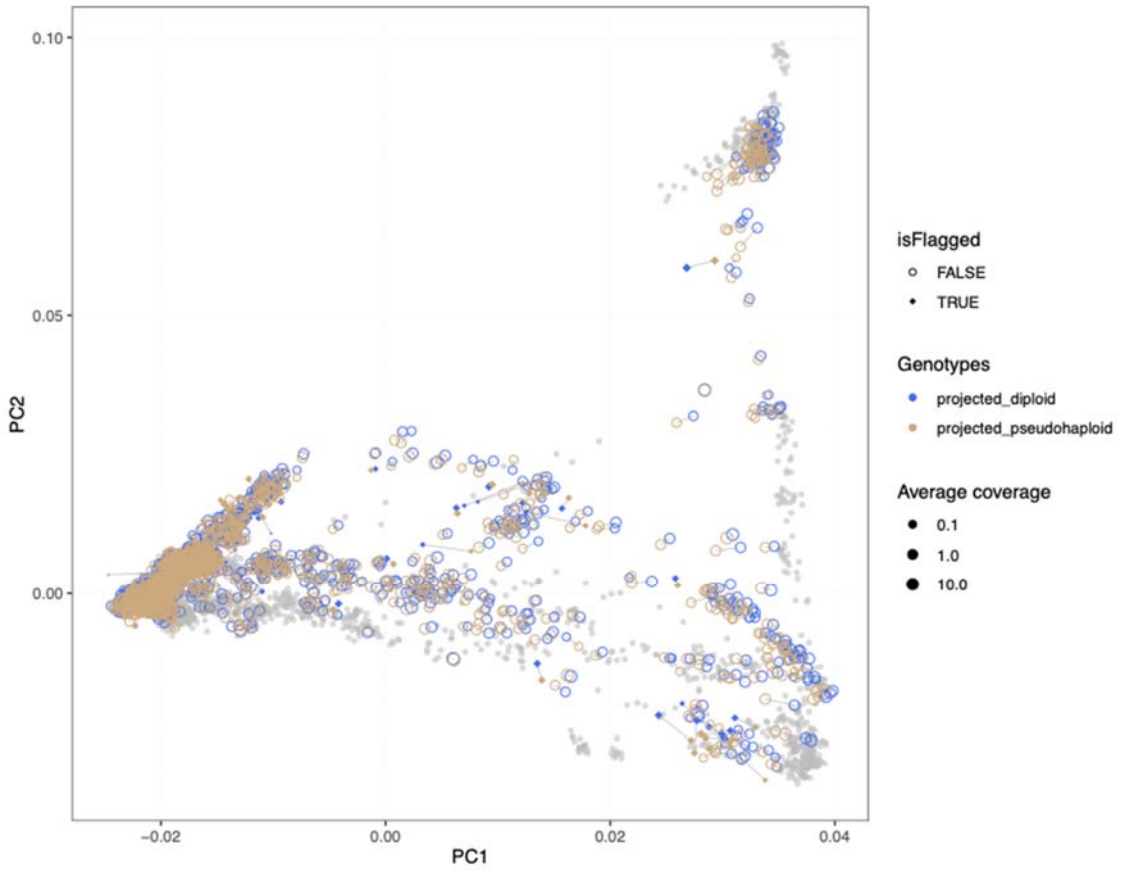


Fig. S3d.1. PCA projection of imputed and pseudo-haploid genotypes. Colored symbols show the position of imputed diploid (blue) and pseudo-haploid (beige) genotypes for each

ancient individual, projected onto principal components inferred from modern individuals from Eurasia, Oceania, and the Americas. Genotype pairs from the same individual are connected by grey lines.

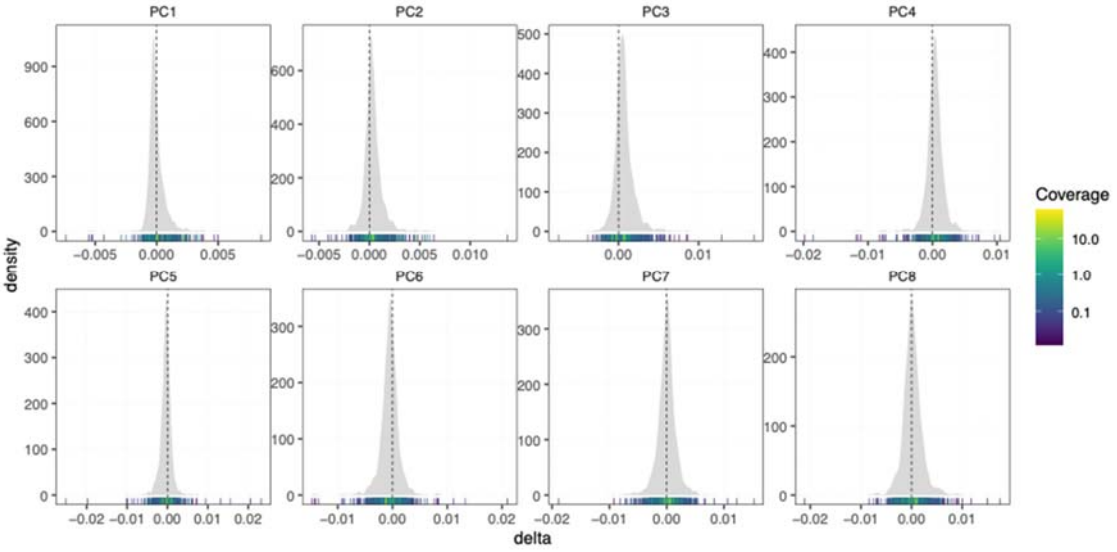


Fig. S3d.2. Distribution of differences between genotypes in PC space. Density plots show differences along individual PCs between imputed and pseudo-haploid genotypes for each individual in Fig. S3d.1, as a function of their average read depth. Marginal rug plots show individual observations, colored by the average read depth of the respective individual.

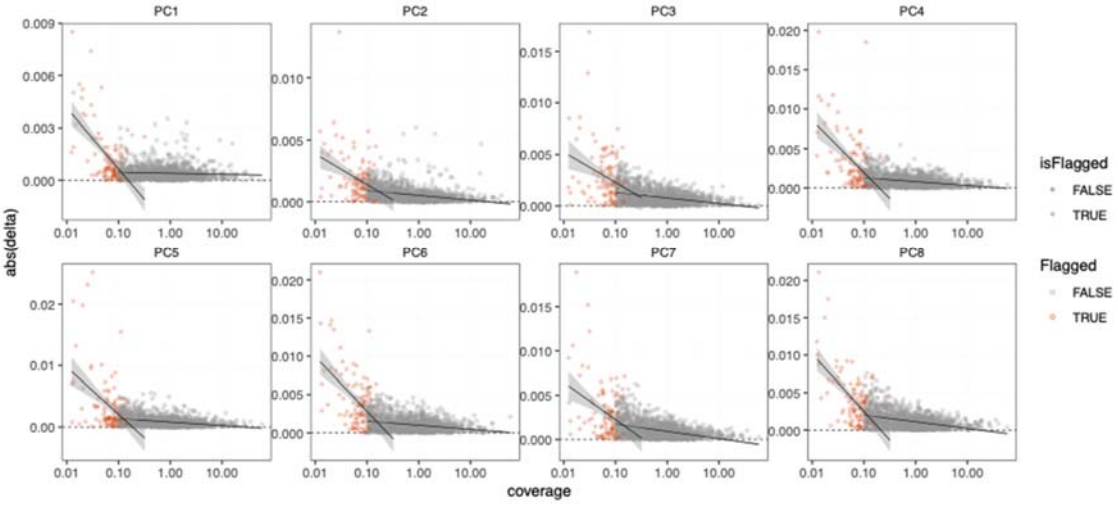


Fig. S3d.3. Relationship of read depth and PCA position. Plot shows absolute value of differences along individual PCs between imputed and pseudo-haploid genotypes for each individual in Fig. S3d.1, as a function of their average read depth. Individuals flagged for low

coverage or low GP average are indicated with red symbols. Linear regression lines for flagged and unflagged individuals are shown with black lines.

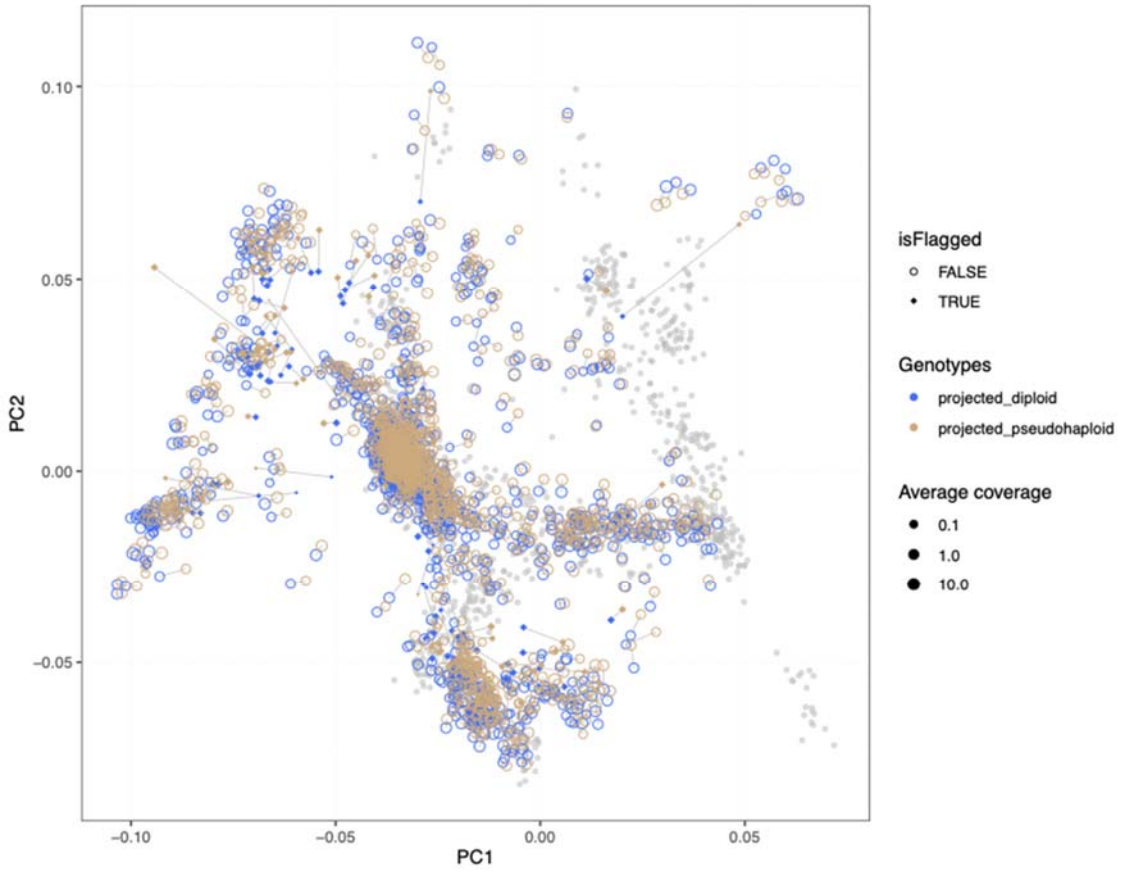


Fig. S3d.4. PCA projection of imputed and pseudo-haploid genotypes. Colored symbols show the position of imputed diploid (blue) and pseudo-haploid (beige) genotypes for each ancient individual, projected onto principal components inferred from modern individuals from Western Eurasia. Genotype pairs from the same individual are connected by grey lines.

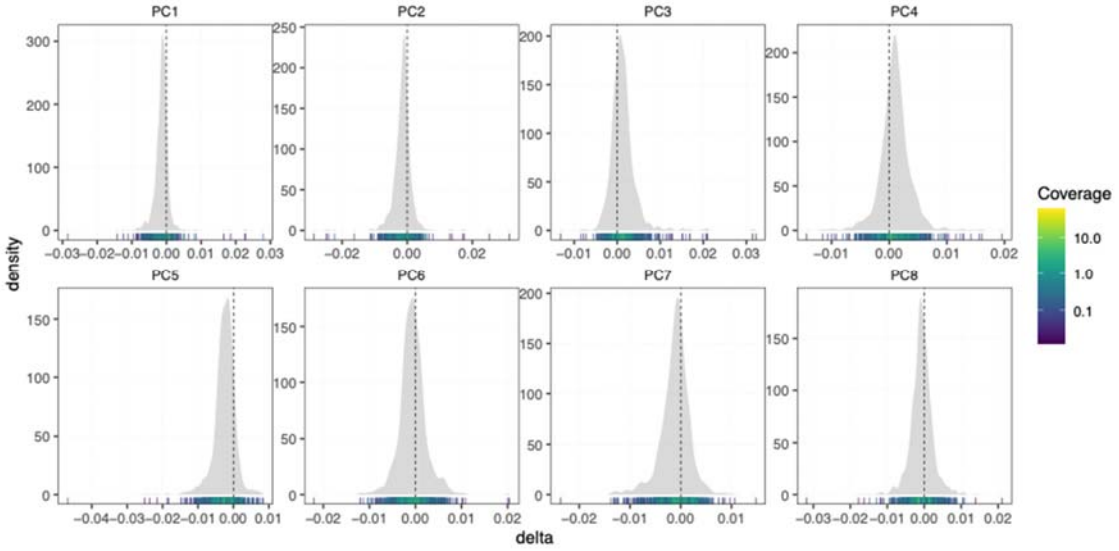


Fig. S3d.5. Distribution of differences between genotypes in PC space. Density plots show differences along individual PCs between imputed and pseudo-haploid genotypes for each individual in Fig. S3d.4, as a function of their average read depth. Marginal rug plots show individual observations, coloured by average read depth of the respective individual.

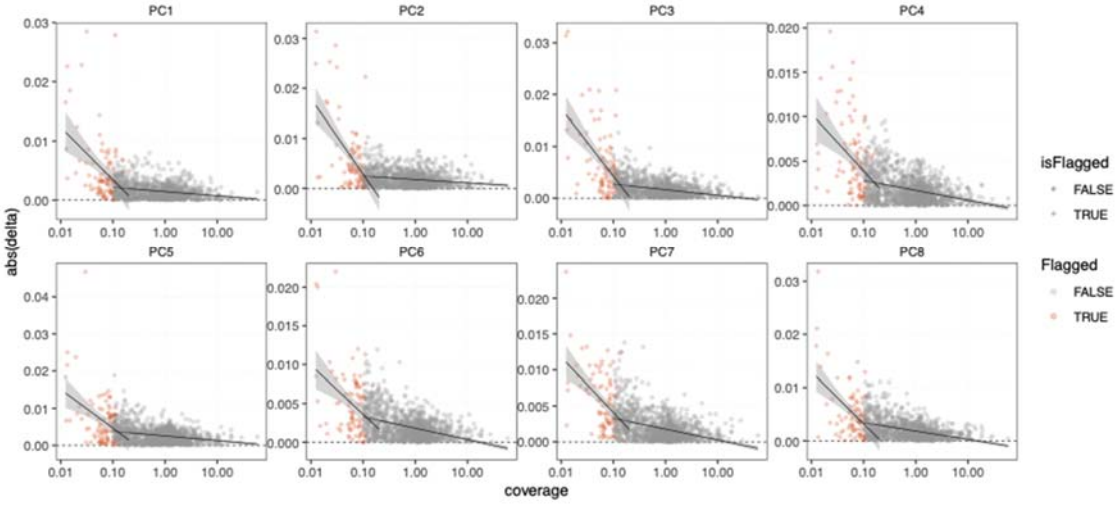


Fig. S3d.6. Relationship of read depth and PCA position. Plot shows absolute difference along individual PCs between imputed and pseudo-haploid genotypes for each individual in Fig. S3d.4, as a function of their average read depth. Individuals flagged for low coverage or low GP average are indicated with red symbols. Linear regression lines for flagged and unflagged individuals are shown with black lines.

PCA position of samples flagged as contaminated

To investigate the effect of elevated contamination estimates on the position of individuals flagged as possibly contaminated, we projected them onto the principal components inferred from modern and ancient individuals passing all filters. We found that the majority of those individuals projected consistently with ancient samples of related age and regional contexts (Fig. S3d.7). An exception to this is seen in the Mesolithic Danish individual NEO1, which shows a clear shift towards present-day Europeans along PC1 and PC2. Overall, our results suggest that inferences about broad patterns of deep Eurasian population structure are likely not affected in the majority of the flagged individuals. We nevertheless opted for a conservative approach and excluded those individuals from in-depth analyses further downstream.

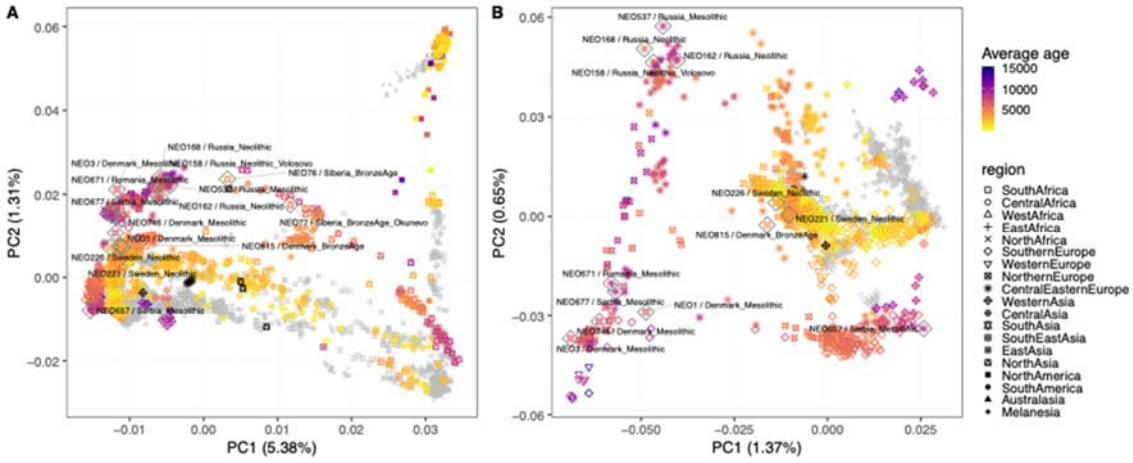


Fig. S3d.7. PCA positions of individuals flagged as contaminated. Flagged individuals are labelled and outlined with grey diamonds. Principal components were inferred using ancient and modern individuals from (A) Eurasia, Oceania, and the Americas or (B) Western Eurasia. Plot symbols indicate geographic region, coloured by age of the respective individual. Present-day individuals are indicated in grey.

Genetic ancestry of newly reported samples

Overview

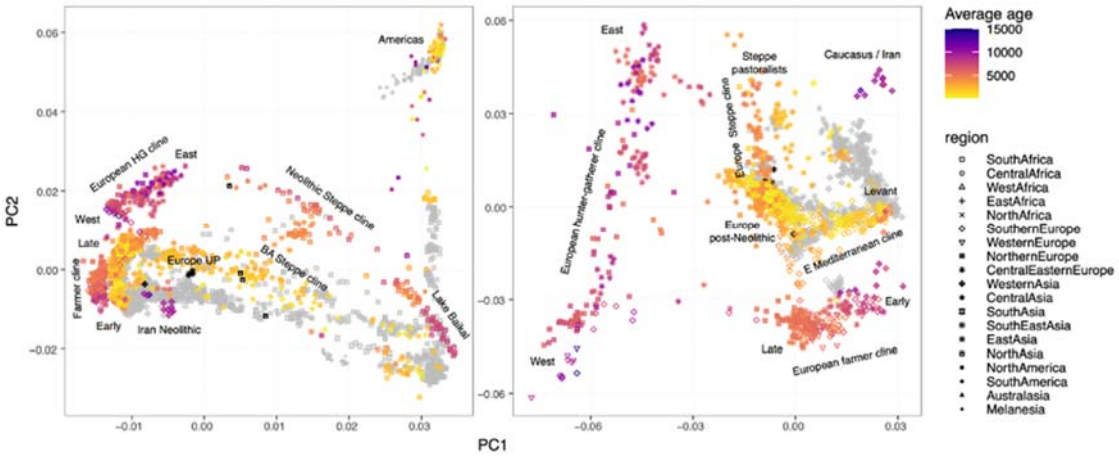


Fig. S3d.8. Overview of genetic structure. PCA of ancient and modern individuals from Eurasia, Oceania, and the Americas (left), or Western Eurasia (right). Plot symbols indicate geographic regions, colored by the age of the respective individual. Present-day individuals are indicated in grey. Terms for spatiotemporal ancestry clusters and clines are indicated.

Genetic structure in a PCA using 3,316 individuals from regions outside Africa is dominated by continental-scale differentiation among western Eurasia (defined here as west of the Urals), east Asia and the Americas (Fig S3d.8). Two west-east clines of ancient individuals connect western and eastern Eurasia: A “Neolithic Steppe cline” between hunter-gatherers of the West Siberian Forest Steppe and Lake Baikal; as well as a later “BA Steppe cline” linking Western Steppe pastoralists with the Altai mountain region (Fig S3d.8).

Focusing the PCA on 2,126 modern and ancient individuals from Western Eurasia, the extremes of the PCA space are defined by clines and clusters related to previously described “deep” ancestry sources, including: A “European hunter-gatherer cline” between western and eastern European Mesolithic individuals; A “European farmer cline” ranging from early Neolithic individuals from Anatolia and Southern Europe to mid- and late Neolithic European individuals; and hunter-gatherers and early farmers from Iran and the Caucasus. European individuals from the late Neolithic and early Bronze Age onwards form an extended “European post-Neolithic” cluster in the centre of the PCA, differentiated along either a “European Steppe cline” between Steppe pastoralists and late European farmers, or an “Eastern Mediterranean cline” anchored in the east by Anatolian and Levantine Bronze Age individuals (Fig S3d.8). The newly reported genomes from western Eurasia cluster across

the entire range of the PCA, resulting in increased fine-scale resolution along the major ancestry clines, particularly the European hunter-gather and farmer clines. The following sections provide regional descriptions for the patterns of ancestry observed in the newly reported samples.

Southern Europe

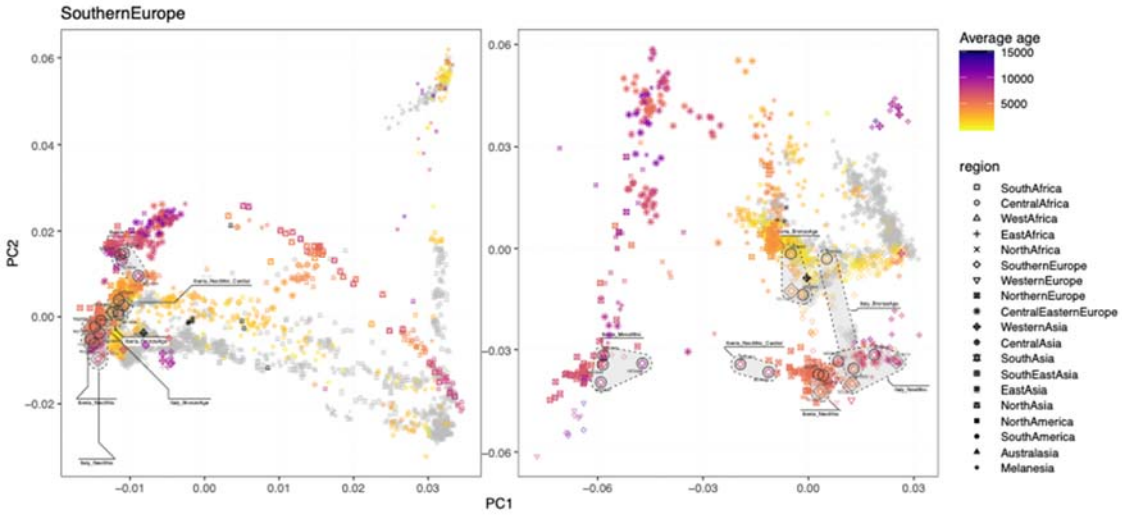


Fig. S3d.9. Newly reported individuals from Southern Europe. PCA positions of newly reported individuals are highlighted with black circles (imputed) or grey diamonds (pseudo-haploid, projected). Individuals from the same spatiotemporal group are connected with shaded hulls.

We report 18 new individuals from Italy (6) and the Iberian Peninsula (12), distributed across European hunter-gatherer (HG), farmer, and post-Neolithic ancestry clusters (Fig S3d.9). Four Iberian Mesolithic individuals cluster with other Southern European Mesolithic individuals at the “western” end of the European HG cline. Among the individuals falling within the European farmer cline, two early Neolithic individuals from Portugal (NEO631, NOE632; ~ 7,300 BP) are shifted towards the European HG cline suggestive of increased HG ancestry. The four most recent individuals (from ~4,100 BP) form part of the extended European post-Neolithic cluster.

Western Europe

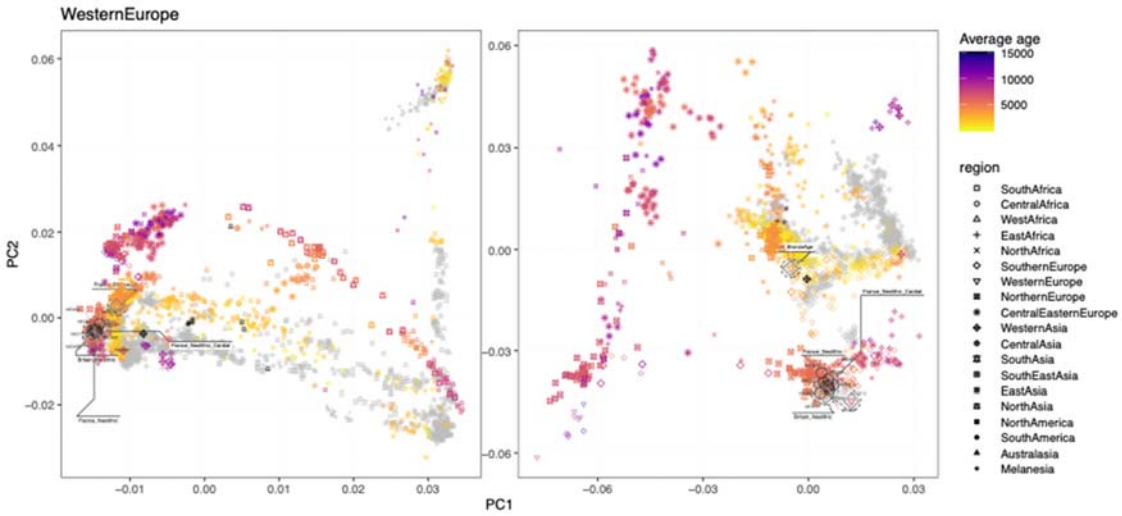


Fig. S3d.10. Newly reported individuals from Western Europe. PCA positions of newly reported individuals are highlighted with black circles (imputed) or grey diamonds (pseudo-haploid, projected). Individuals from the same spatiotemporal group are connected with shaded hulls.

We report 12 new individuals from France (5) and the UK (7), from the early Neolithic to the Bronze Age. All 11 Neolithic individuals fall within the European farmer cline, whereas a single Bronze Age individual from Grotte Mandrin (NEO120, ~3,400BP) clustered with post-Neolithic Europeans (Fig. S3d.10).

Northern Europe

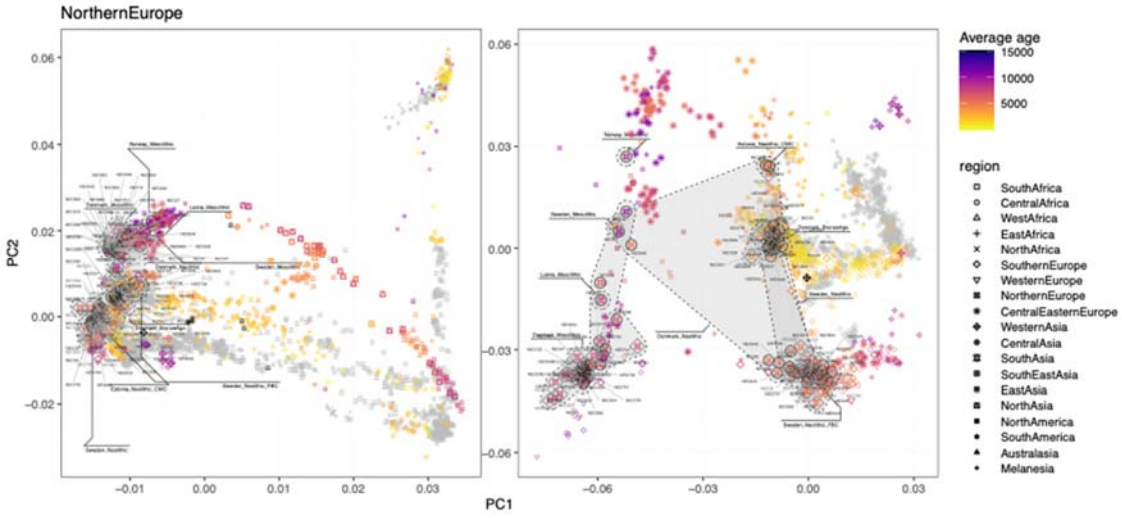


Fig. S3d.11. Newly reported individuals from Northern Europe. PCA positions of newly reported individuals are highlighted with black circles (imputed) or grey diamonds (pseudo-

haploid, projected). Individuals from the same spatiotemporal group are connected with shaded hulls.

We report 124 new individuals from Denmark (100), Sweden (21), Norway (1) and the Baltic (2), spanning a period from ~10,500BP to 3,100BP. This transect includes 46 Mesolithic individuals, all of which cluster within the European HG cline (Fig. S3d.10). The 40 HG individuals from Denmark fall towards the “western” end of the cline, whereas the other Scandinavian and Baltic individuals occupy varied positions shifted towards the “eastern” end of the cline. Neolithic Scandinavian individuals generally fall towards the “late” end of the European farmer cline, with later Neolithic individuals also found among the extended post-Neolithic Europe cluster. Three late Neolithic (~4,500BP) individuals from Denmark (NEO876, NEO792) and Estonia (NEO306) are shifted further up along PC2 towards the Steppe pastoralist cluster, suggesting higher amounts of Steppe-related ancestry (Fig. S3d.11).

Central and Eastern Europe

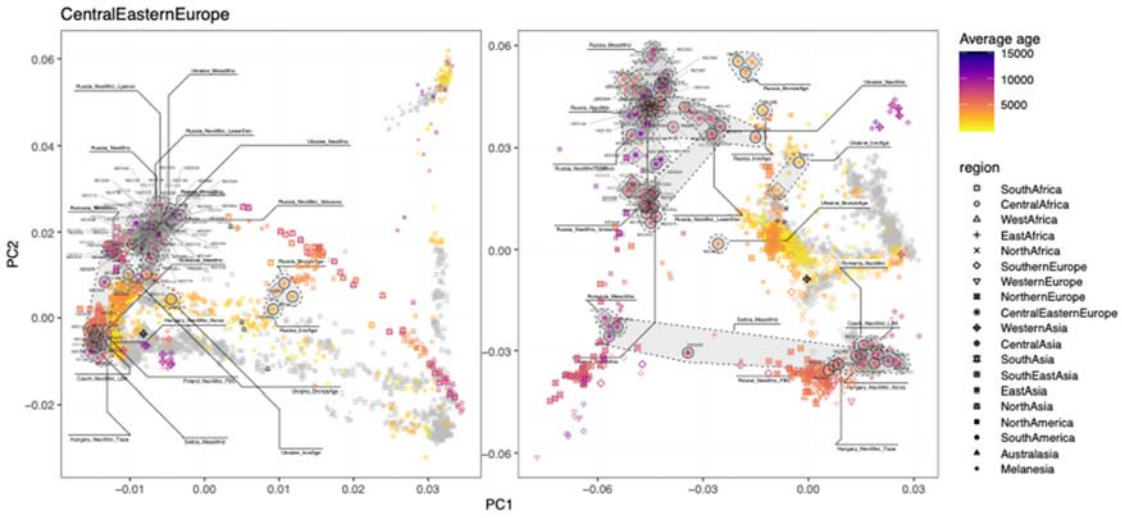


Fig. S3d.12. Newly reported individuals from Southern Europe. PCA positions of newly reported individuals are highlighted with black circles (imputed) or grey diamonds (pseudo-haploid, projected). Individuals from the same spatiotemporal group are connected with shaded hulls.

We report 112 new individuals from Central and Eastern Europe, falling into two distinct groups. The 92 individuals from Russia (57) and Ukraine (35) predominantly occupy a broad area between the centre and "eastern" end of the European HG cline, roughly corresponding to a geographic cline from the south (Ukraine) to the north (Russia) (Fig. S3d.11). Among the southern Russian samples, six individuals from Golubaya Krinitsa in the Middle Don region are shifted on a cline along PC1 towards Iranian and Caucasus Mesolithic and Neolithic at the other extreme, falling close to later Steppe pastoralists from the region. Three Bronze Age individuals from Northwestern Russia (Bol'shoy Oleni Ostrov; NEO60, NEO61, NEO62) are positioned between the Neolithic and BA Steppe clines, centrally between the West and East Eurasian poles in the extended PCA of all non-Africans (Fig. S3d.12).

The remaining 20 samples from Central and Southeastern Europe include Mesolithic individuals at the "western" end of the European HG cline, as well as early Neolithic individuals on the farmer cline. An early Neolithic individual from Iron Gates, Serbia (NEO658) is found intermediate between the HG and farmer clines, suggestive of recent farmer/HG admixture (Fig. S3d.12).

Western Asia

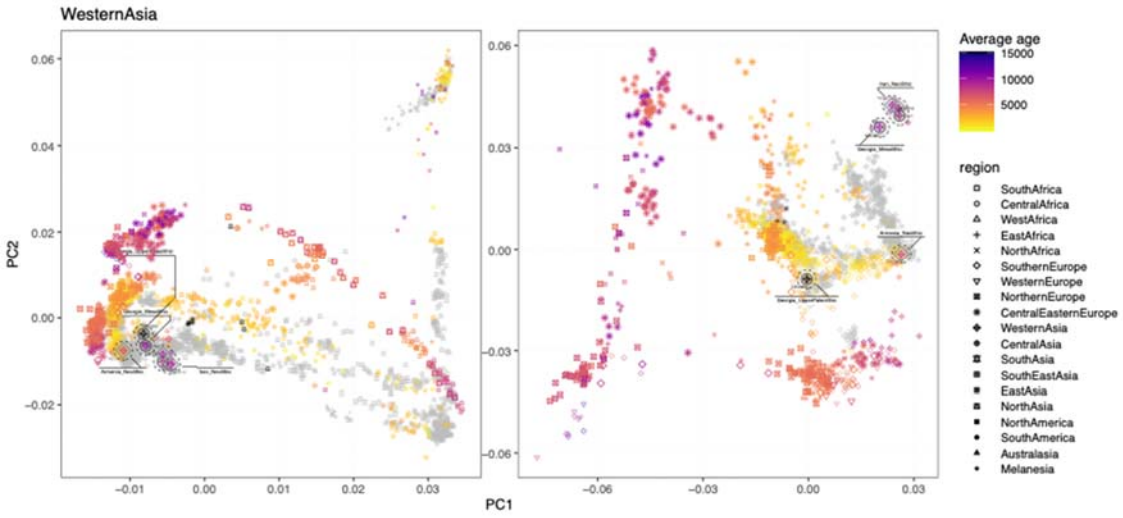


Fig. S3d.13. Newly reported individuals from Western Asia. PCA positions of newly reported individuals are highlighted with black circles (imputed) or grey diamonds (pseudo-haploid, projected). Individuals from the same spatiotemporal group are connected with shaded hulls.

We report 6 new individuals from Iran (3) and the South Caucasus region (3). The oldest sample in the dataset, a ~25,000-year-old individual from Georgia is positioned intermediate between Upper Paleolithic Europeans and early Neolithic farmers in both the west Eurasian and extended non-African PCA (Fig. S3d.13). Three Iranian Neolithic individuals (~9,200 BP) as well as one Mesolithic Georgian individual (NEO281; ~9,700 BP) fall with other previously published samples of similar provenance, defining one of the extremes of PC1/PC2 space. One Neolithic individual from Armenia (NEO110, ~7,600 BP) is found at the “eastern” extreme of an eastern Mediterranean cline between ancient Levantine individuals and Southern post-Neolithic Europeans.

Central Asia

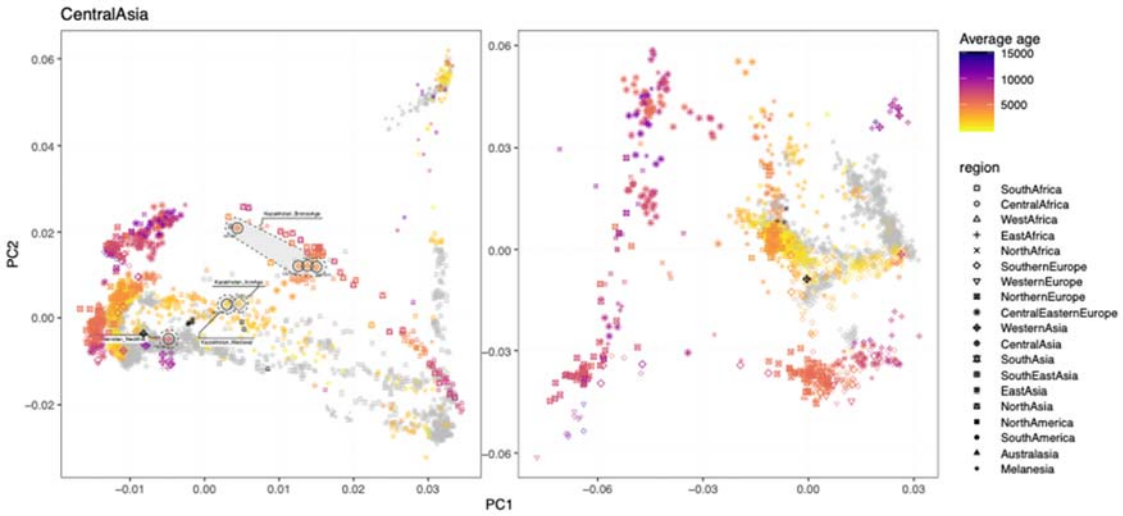


Fig. S3d.14. Newly reported individuals from Central Asia. PCA positions of newly reported individuals are highlighted with black circles (imputed) or grey diamonds (pseudo-haploid, projected). Individuals from the same spatiotemporal group are connected with shaded hulls.

We report 7 new individuals from Kazakhstan (6) and Turkmenistan (1). The Neolithic individual from Turkmenistan (~6,500 BP) clusters close to Neolithic Iranians. The individuals from Kazakhstan are more recent (~4,500BP – 2,000BP), with the older individuals forming part of the Neolithic Steppe cline, and the younger individuals along the BA Steppe cline (Fig. S3d.14).

North Asia

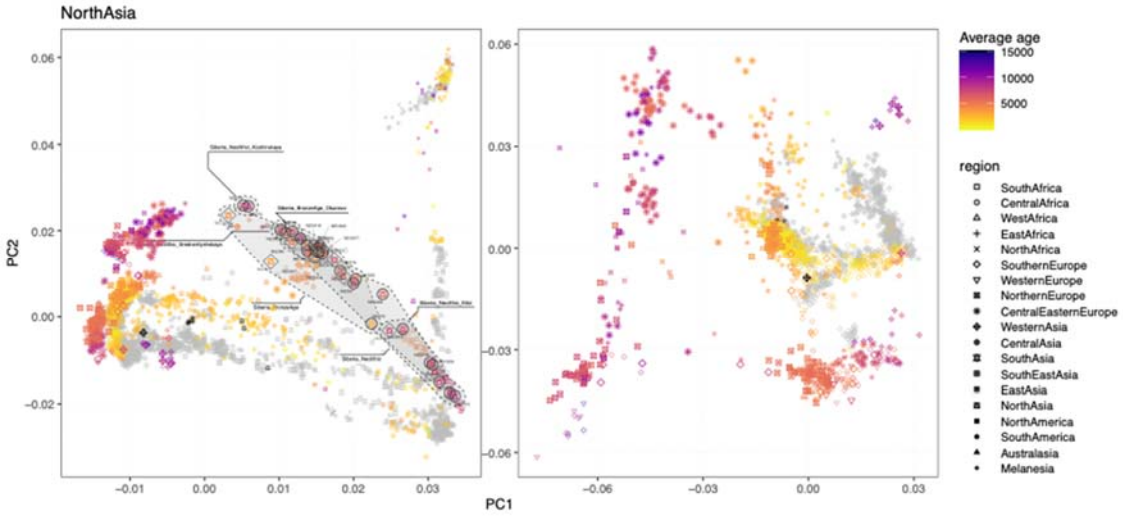


Fig. S3d.15. Newly reported individuals from North Asia. PCA positions of newly reported individuals are highlighted with black circles (imputed) or grey diamonds (pseudo-

haploid, projected). Individuals from the same spatiotemporal group are connected with shaded hulls.

We report 38 new individuals from Western Siberia and Lake Baikal, spanning a period from ~8,300BP to 2,800 BP. The individuals fall along the entire range of the Neolithic Steppe cline, spanning from early Forest Steppe hunter-gatherers at the “western” end (NEO72, NEO73) to Lake Baikal hunter-gatherers at the “eastern” end (Fig. S3d.15)

References

1. The 1000 Genomes Project Consortium,. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
2. Patterson, N. *et al.* Ancient Admixture in Human History. *Genetics* **192**, 1065–1093 (2012).
3. Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409–413 (2014).
4. Lazaridis, I. *et al.* Genomic insights into the origin of farming in the ancient Near East. *Nature* **536**, 419–424 (2016).
5. Pickrell, J. K. *et al.* The genetic prehistory of southern Africa. *Nat. Commun.* **3**, 1143 (2012).
6. Shringarpure, S. S., Bustamante, C. D., Lange, K. & Alexander, D. H. Efficient analysis of large datasets and sex bias with ADMIXTURE. *BMC Bioinformatics* **17**, 218 (2016).
7. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A Tool for Genome-wide Complex Trait Analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
8. Patterson, N., Price, A. L. & Reich, D. Population Structure and Eigenanalysis. *PLoS Genet* **2**, e190 (2006).
9. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).
10. Antonio, M. L. *et al.* Ancient Rome: A genetic crossroads of Europe and the Mediterranean. *Science* **366**, 708–714 (2019).

3e) Inferring the spatiotemporal spread of population movements in the past 13 millennia

Fernando Racimo[1]

[1]Lundbeck Foundation GeoGenetics Centre, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark

Introduction

We aimed to infer the geographic and temporal spread of major population movements in the past 13 millennia of Western Eurasian history. We used a method developed in [1], which uses spatiotemporal ordinary kriging on latent ancestry proportion estimates from ancient and present-day genomes. This way, we obtained detailed spatiotemporal maps reflecting the dynamics of the spread of ancestry during the transition from the Mesolithic to the Neolithic, Bronze Age, Iron Age and more recent periods, finally resulting in the complex ancestry make-up of present-day populations in the region.

Methods

We obtained ancestry proportions estimated using Admixture[2] with K=9 latent ancestry clusters (Supplementary Note S3d) on a sequence dataset including both whole-genome shotgun-sequenced genomes and genomic sequences obtained via SNP capture, after imputation (Supplementary Note S2). We performed spatiotemporal kriging[3] of these proportions over the last 12,900 years, in intervals of 300 years, with a 5,000-point spatial grid spanning Western and Central Eurasia. We used the R package *gstat* to fit a spatiotemporal variogram via a metric covariance model, and perform ordinary kriging[4]. We focused on the ancestry clusters for which we could fit variogram models that were not static over time.

Results

We were able to fit spatiotemporal variogram functions to six of the nine ancestries. We label these as WHG, EHG, IRN, LVN, SIB and EAS. The first four are roughly maximised in Mesolithic western European hunter-gatherers, Mesolithic eastern European hunter-gatherers, Iranian Neolithic populations, Levant Neolithic populations and ancient Siberian populations, respectively (see [5] for a model positing the first four of these populations as the major sources of ancestry in present-day Europeans). We depict the spatiotemporal spread

of the first four of these ancestries in Fig. S3e.1 and Supplementary Animations 1-4. The fifth ancestry (SIB) occurs at much lower rates in western Eurasia, and rises in frequency in northeastern Europe during the Iron Age (Fig. S3e.2, Supplementary Animation 5)[6–8]. A sixth ancestry (EAS) has affinities to East Asians, and expands into the Caucasus in recent times (Fig. S3e.2, Supplementary Animation 6).

These spatiotemporal maps evince interesting patterns of ancestry change across the landscape. For example, the advancement of Neolithic Levant (LVN) ancestry appears staggered: we observe different periods of advancement followed by stasis. In addition to the Bronze Age movement of EHG ancestry, there is a southern incursion of IRN ancestry via South Europe[9–11]. This is particularly obvious in Bronze Age Greek and Iron Age Roman samples, and may be due to contacts with Anatolia and Northern Africa (where this ancestry is also present). Additionally, we observe small incursions of very late SIB ancestry into Eastern Europe (Fig. S3e.2 Supplementary Animation 5). This signal is driven by the presence of SIB ancestry in Iron Age Cimmerian nomads[12] and in a medieval Serbian[13], and could perhaps be linked to the introduction of languages from the Finno-Ugric family into the Hungarian Plain. An incursion of this ancestry into Western Eurasia can also be seen in Medieval Ottoman Anatolians[14].

We can focus on local timelines of kriged ancestry changes in different points of the map (Fig. S3e.3). Here, we observe that the timing and duration of the rise in LVN ancestry was different in different points in Europe (Fig. S3e.3). We also observe that, in certain regions of Europe, the rise in IRN and EHG ancestry are largely decoupled from each other (see e.g. “Rome” in Fig. S3e.2)[9–11,15].

Figures

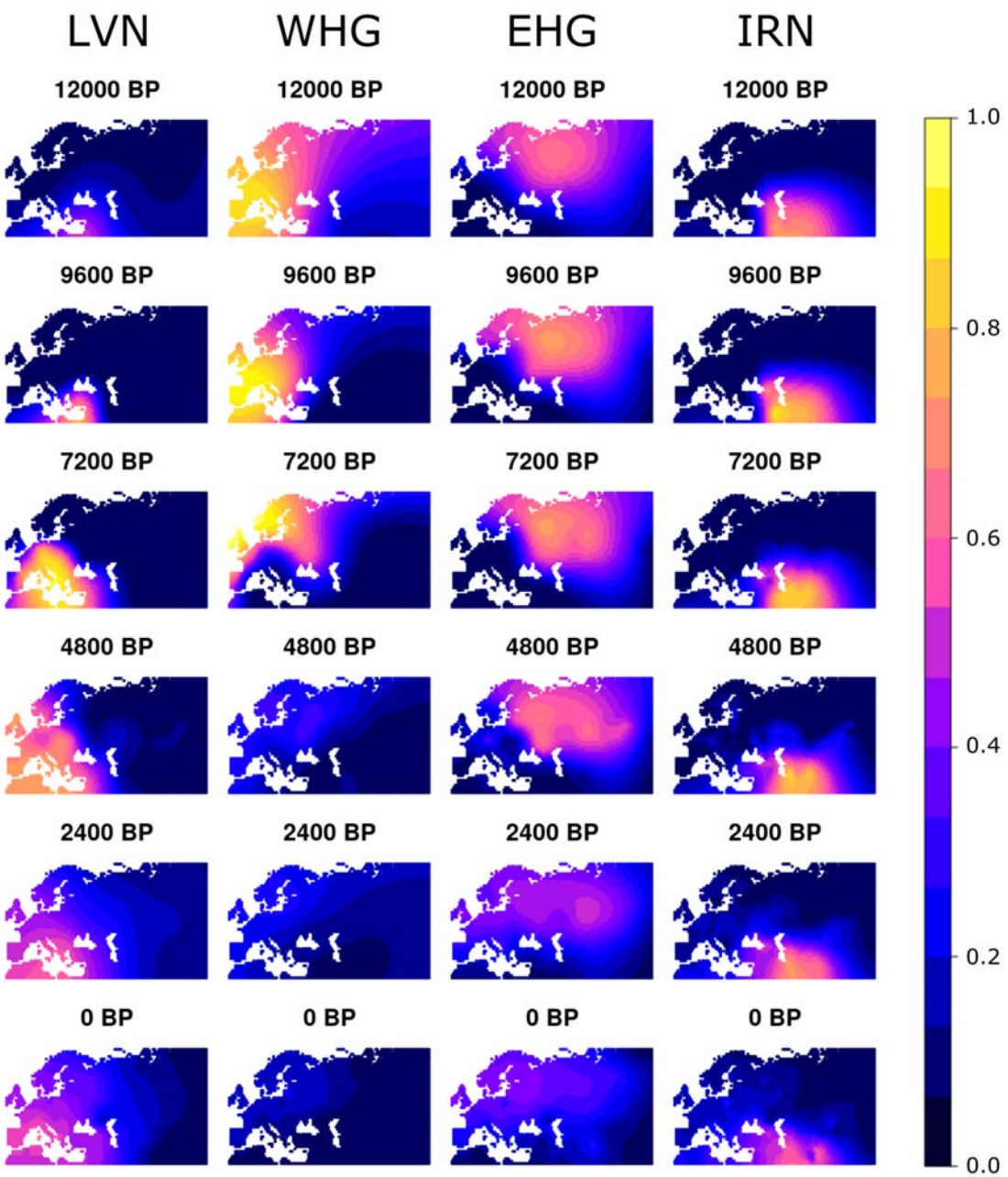
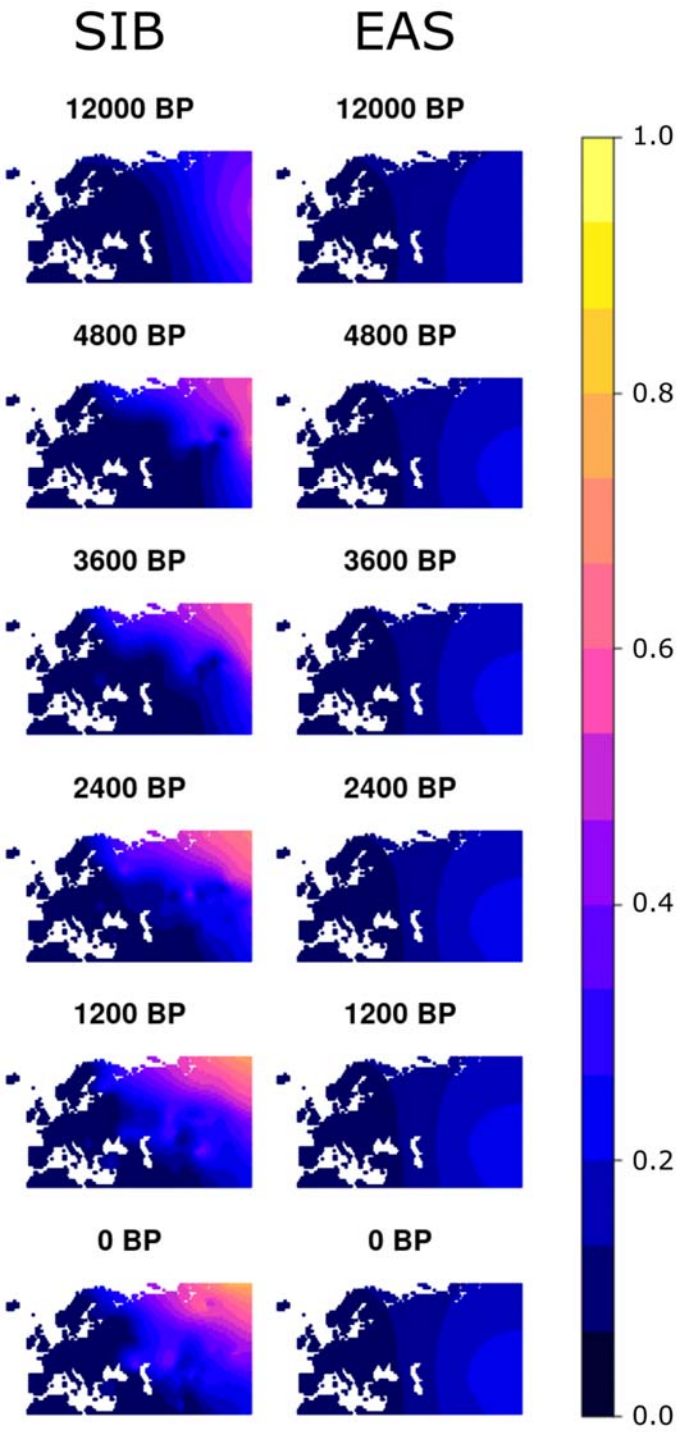


Figure S3e.1. Spatiotemporal kriging of four major ancestry clusters over the last 12,000 years of human history. LVN = ancestry maximised in Anatolian farmer populations. WHG = ancestry maximised in western European hunter-gatherers. EHG = ancestry maximised in eastern European hunter-gatherers. IRN = ancestry maximised in Iranian Neolithic individuals and Caucasus hunter-gatherers.



1421

1422 **Figure S3e.2.** Spatiotemporal kriging of two additional ancestry clusters with later incursions
1423 into Western Eurasia over the last 12,000 years of human history, particularly focusing on
1424 the last 5,000 years. SIB = ancestry maximised in ancient Siberian individuals. EAS =
1425 ancestry maximised in East Asian individuals.

1426

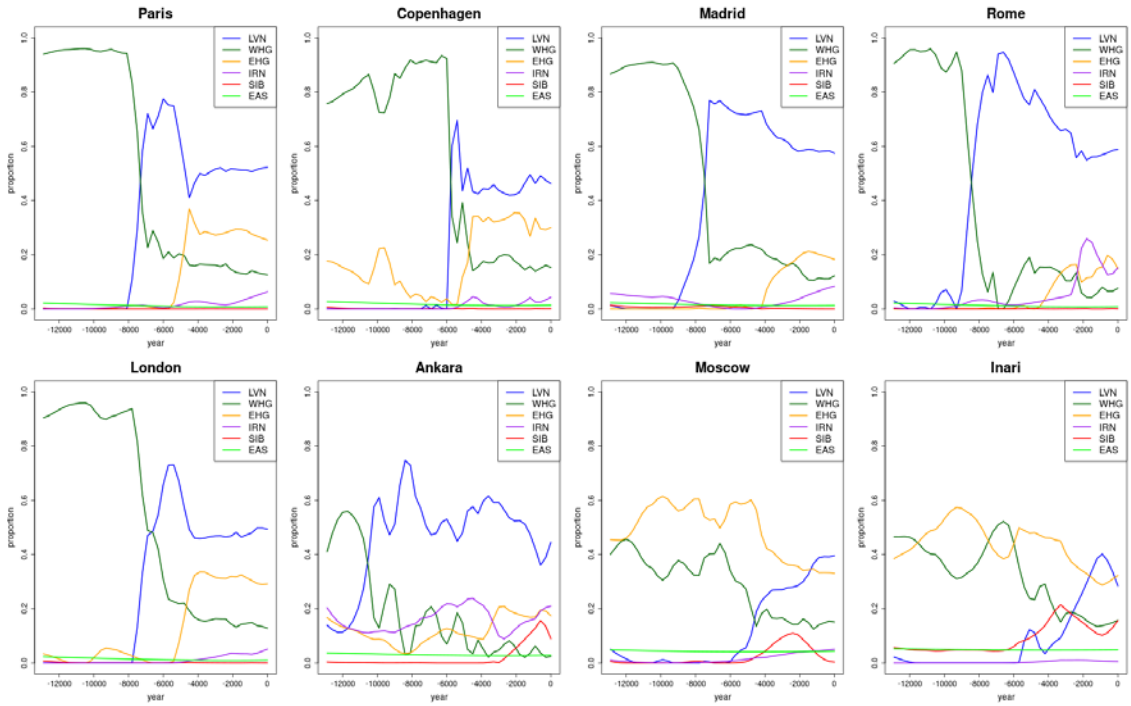
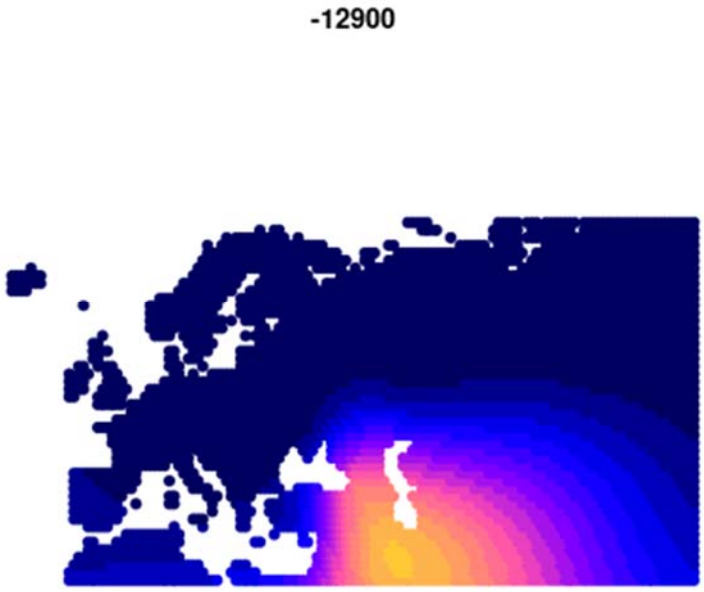
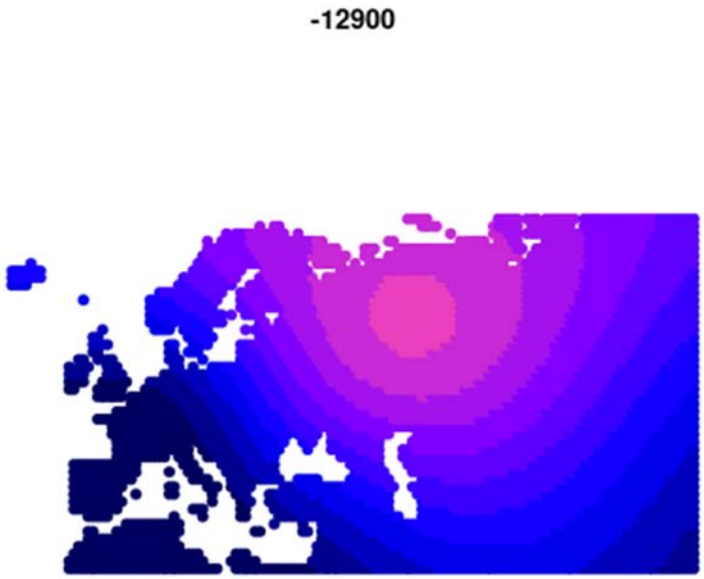


Figure S3e.3. Local timelines in what are now 8 urban centres across western Eurasia, reflecting local differences in the tempo and mode of ancestry changes over time. LVN = ancestry maximised in Anatolian farmer populations. WHG = ancestry maximised in western European hunter-gatherers. EHG = ancestry maximised in eastern European hunter-gatherers. IRN = ancestry maximised in Iranian Neolithic individuals / Caucasus hunter-gatherers. SIB = ancestry maximised in ancient Siberian individuals. EAS = ancestry maximised in East Asian individuals.

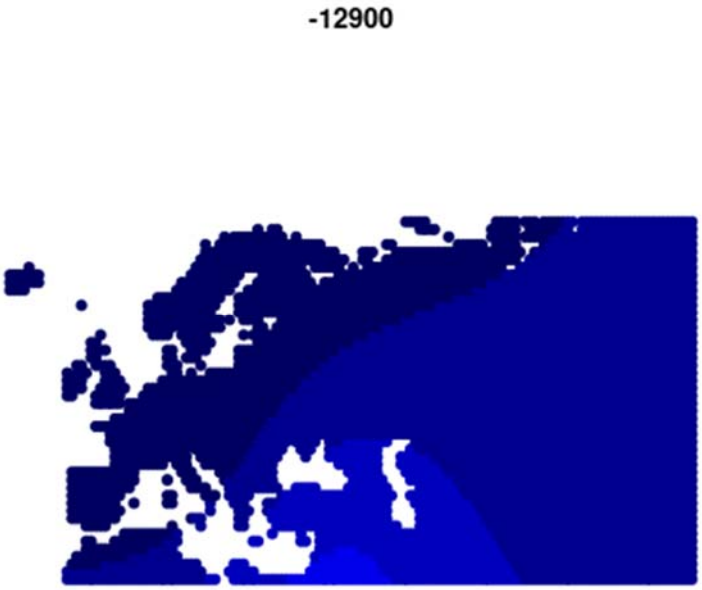
Animation S3e.1. IRN



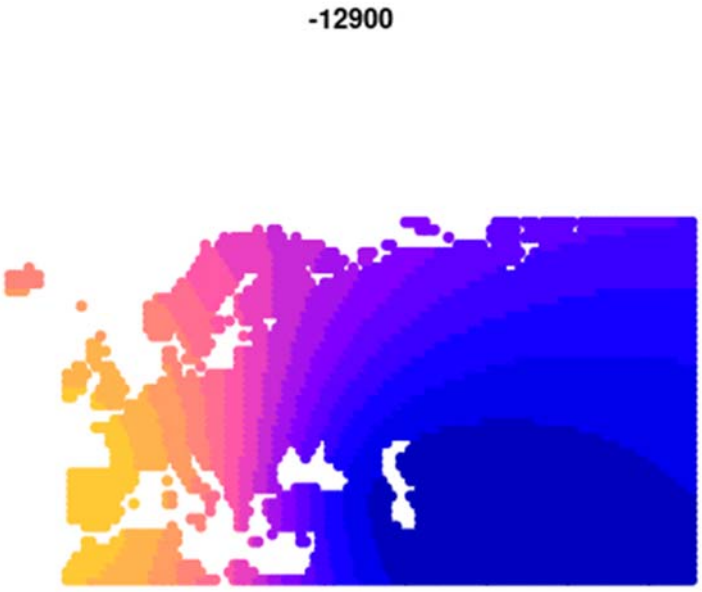
Animation S3e.2. EHG



Animation S3e.3. LVN



Animation S3e.4. WHG



References

1. Racimo, F. *et al.* The spatiotemporal spread of human migrations during the European Holocene. *Proc. Natl. Acad. Sci. U. S. A.* (2020) doi:10.1073/pnas.1920051117.

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

3. Cressie, N. & Wikle, C. K. *Statistics for Spatio-Temporal Data*. (John Wiley & Sons, 2015).

4. Gräler, B., Pebesma, E. & Heuvelink, G. Spatio-Temporal Interpolation using gstat. *The R Journal* vol. 8 204 (2016).

5. Lazaridis, I. *et al.* Genomic insights into the origin of farming in the ancient Near East. *Nature* **536**, 419–424 (2016).

6. Lamnidis, T. C. *et al.* Ancient Fennoscandian genomes reveal origin and spread of Siberian ancestry in Europe. *Nat. Commun.* **9**, 5018 (2018).

7. Tambets, K. *et al.* Genes reveal traces of common recent demographic history for most of the Uralic-speaking populations. *Genome Biology* vol. 19 (2018).

8. Saag, L. *et al.* The Arrival of Siberian Ancestry Connecting the Eastern Baltic to Uralic Speakers further East. *Curr. Biol.* **29**, 1701–1711.e16 (2019).

9. Antonio, M. L. *et al.* Ancient Rome: A genetic crossroads of Europe and the Mediterranean. *Science* **366**, 708–714 (2019).

10. Mathieson, I. *et al.* The genomic history of southeastern Europe. *Nature* **555**, 197–203 (2018).

11. Marcus, J. H. *et al.* Genetic history from the Middle Neolithic to present on the Mediterranean island of Sardinia. *Nat. Commun.* **11**, 939 (2020).

12. Krzewińska, M. *et al.* Ancient genomes suggest the eastern Pontic-Caspian steppe as the source of western Iron Age nomads. *Sci Adv* **4**, 4457 (2018).

13. Veeramah, K. R. *et al.* Population genomic analysis of ellimpse ongated skulls reveals extensive female-biased immigration in Early Medieval Bavaria. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 3494–3499 (2018).

14. Damgaard, P. de B. *et al.* The first horse herders and the impact of early Bronze Age steppe expansions into Asia. *Science* vol. 360 7711 (2018).

15. Fernandes, D. M. *et al.* The spread of steppe and Iranian-related ancestry in

the islands of the western Mediterranean. *Nat Ecol Evol* **4**, 334–345 (2020).

3f) HBD/ IBD sharing/ROH/clustering

Martin Sikora[1]
[1]Lundbeck Foundation GeoGenetics Centre, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark

Methods

We used *IBDseq*[1] to detect genomic segments shared identical-by-descent (IBD) between all individuals in the "1000G" dataset, restricting to transversion SNPs with imputation INFO score ≥ 0.8 and MAF ≥ 0.01. We filtered the resulting IBD segments for LOD score ≥ 3 and a minimum length of 2 centimorgans (cM), and further removed regions of excess IBD following the approach of Browning and Browning[2]. First, we used the *GenomicRanges*[3] package in R to calculate the total number of IBD segments overlapping each position along the genome, and calculated their 3% trimmed mean and standard deviation (SD). We then called regions of excess IBD if they were > 3 trimmed SD from the trimmed mean. We split IBD segments overlapping the excess IBD regions, and removed any segments with length < 2 cM after splitting. For analyses of runs of homozygosity (ROH) we used a shorter length cutoff of 1cM.

We carried out genetic clustering of the ancient individuals using hierarchical community detection on a network of pairwise identity-by-descent (IBD)-sharing similarities[4]. To facilitate detection of clusters at a finer scale, we ran *IBDseq* on a dataset restricting to ancient samples only, and applied more lenient filters of imputation INFO score > 0.5, and minimum IBD segment length of 1 cM. We constructed a weighted network of the individuals using the *igraph*[5] package in R, with the fraction of the genome shared IBD between pairs of individuals as weights. We then performed iterative community detection on this network using the Leiden algorithm[6] implemented in the *leidenAlg* R package (*v1.01,* https://github.com/kharchenkolab/leidenAlg). We used a resolution parameter of r=0.5 as the starting value for each level of community detection. If more than one community was detected, we split the network into the respective communities, and repeated the community detection step. If no communities were detected, we incremented the resolution parameter in steps of 0.5 until a maximum value of r=3. The initial clustering was completed when no more communities were detected at the highest resolution parameter, across all

subcommunities. To convert the resulting hierarchy into a final clustering, we simplified the initial clustering by collapsing nodes into single clusters based on observed spatiotemporal annotations of the samples.

To estimate ancestry proportions from patterns of pairwise IBD sharing, we used an approach akin to "chromosome painting"[7]. We first inferred an IBD-based "painting profile" for each target individual, by summing up the total amount of IBD shared with each "donor" group (using population labels for modern donors or IBD-based genetic clusters for ancient donors), and normalising them to the interval [0,1]. We used a leave-one-out approach as in[8] to account for the fact that recipient individuals cannot be included as donors from their own group. We then used these painting profiles in supervised modelling of target individuals as mixtures from different sets of putative source groups[8,9], using non-negative least squares implemented in the R package *limSolve*[10].

To investigate ancestry compositions across the full set of ancient individuals, we used three sets of source groups reflecting different temporal depths:

- "deep", a set of groups representing highly differentiated deep ancestry sources
- "postNeol", using diverse Neolithic and earlier source groups
- "postBA", using Late Neolithic and Bronze Age source groups

We also performed two analyses of more restricted spatiotemporal scope:

- "hgEur", modelling European hunter-gatherers as mixtures of early European hunter-gatherers and selected outgroups
- "fEur", modelling later European farmers as mixtures of earlier farmers and hunter-gatherers
- "postNeolScand", modelling Scandinavian Late Neolithic and early Bronze Age individuals as mixtures of other European early Bronze Age groups.

Results

IBD-based hierarchical graph clustering

We performed hierarchical graph clustering on the 1,492 ancient individuals passing all filters, which were assigned into a final curated set of 122 genetic clusters (Fig. S3f.1). The obtained clusters captured both broad and finer-scale genetic structure, corresponding to shared ancestry within particular spatiotemporal ranges and/or archaeological contexts (Fig.

S3f.2). We named these cluster using a “geographic-temporal” nomenclature[11] (e.g. “Denmark_10500BP_6000BP”), in concert with more traditional names for groups of multiple clusters with shared archaeological or subsistence contexts (e.g. “Farmer_Europe_early”) where applicable.

At the highest level of the clustering hierarchy, the individuals were partitioned into six global clusters representing broad continent-wide genetic structure.

- Africa_8000BP_400BP
- Europe_15000BP_4000BP
- EuropeWCAsia_25000BP_300BP
- Eurasia_5000BP_200BP
- Asia_45000BP_200BP
- Americas_12000BP_100BP

The following sections provide more detailed descriptions of relevant sub-clusters within the four global clusters from Eurasia.

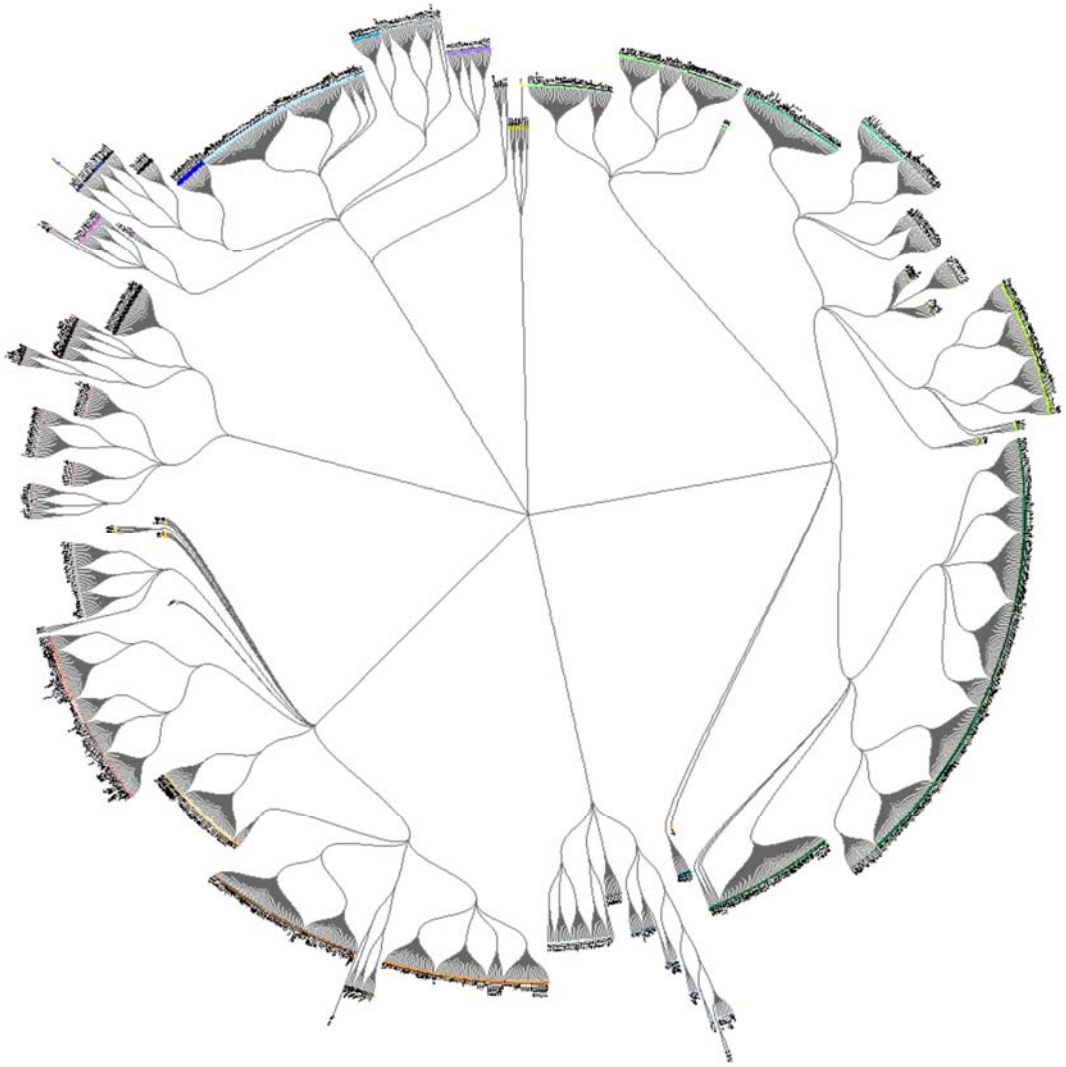


Fig. S3f.1. Hierarchical graph clustering. Tree diagram showing final curated hierarchical clustering relationship among the 1,492 ancient individuals passing all filters. Genetic clusters are differentiated using plot symbol colours and shapes.

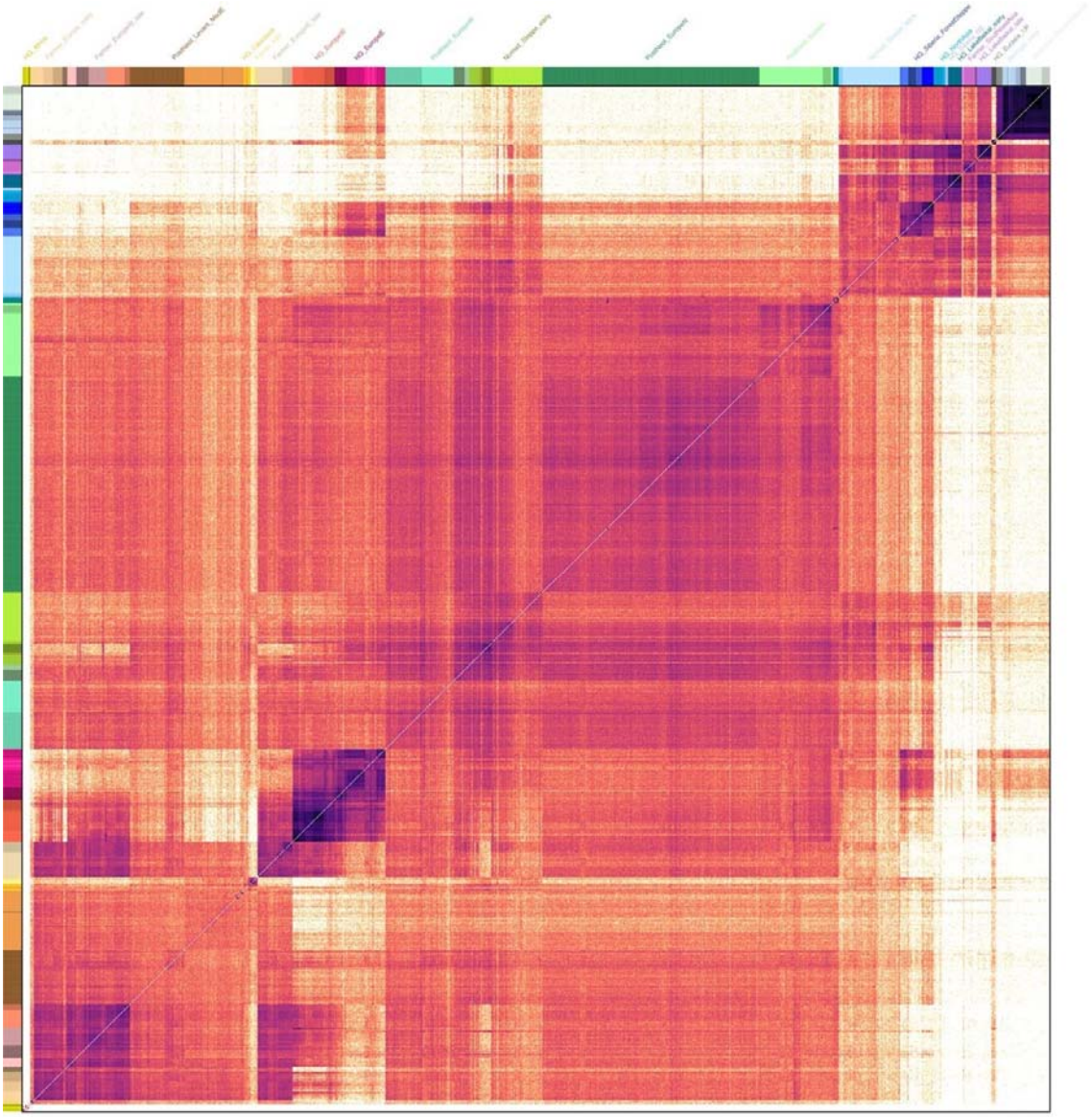


Fig. S3f.2. IBD sharing similarities. Heatmap of pairwise IBD-sharing similarities between the 1,492 ancient individuals passing all filters, sorted according to clustering hierarchy. Colored bars indicate cluster membership of individuals. Selected cluster group labels are shown in the top colour bar.

Europe_15000BP_4000BP

This global cluster includes individuals from western Eurasian Mesolithic and Neolithic contexts with hunter-gatherer ancestry. The individual genetic clusters are partitioned into two cluster groups corresponding to “Eastern hunter-gatherers” and “Western hunter-gatherers” as previously used in the literature (Fig. S3f.3-6):

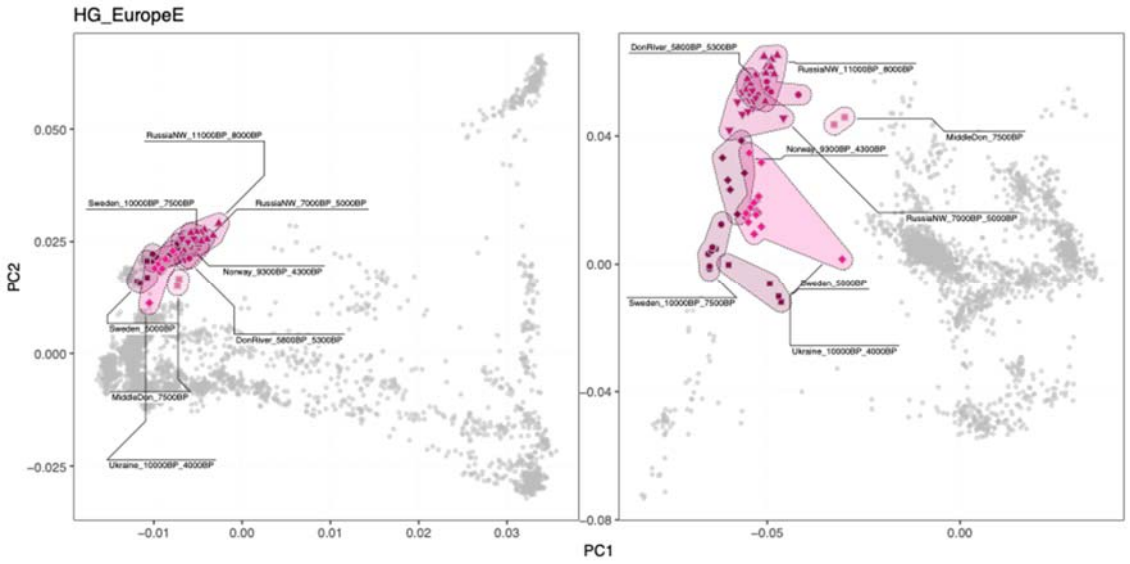


Fig. S3f.3 PCA for cluster group *HG_EuropeE*. PCA positions of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

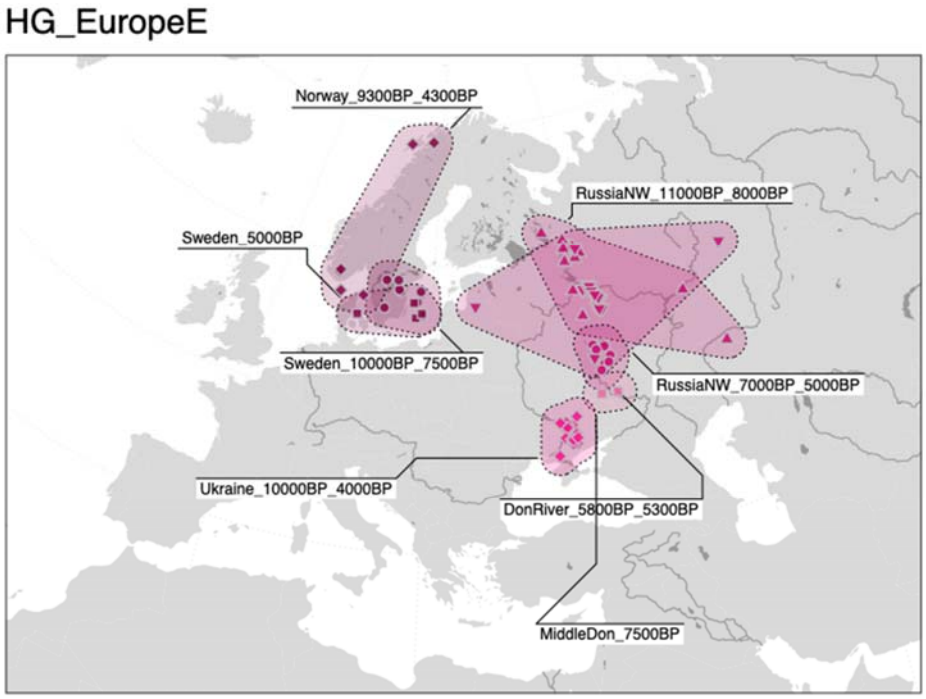


Fig. S3f.4 Geographic distribution of individuals in cluster group *HG_EuropeE*. Geographic locations of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

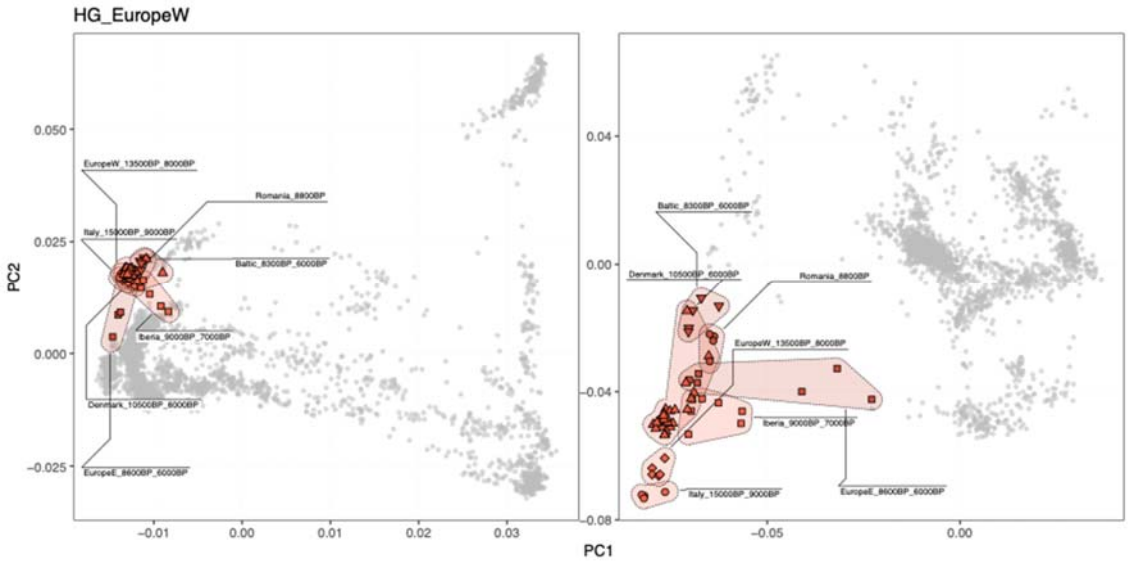


Fig. S3f.5 PCA for cluster group *HG_EuropeW*. PCA positions of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

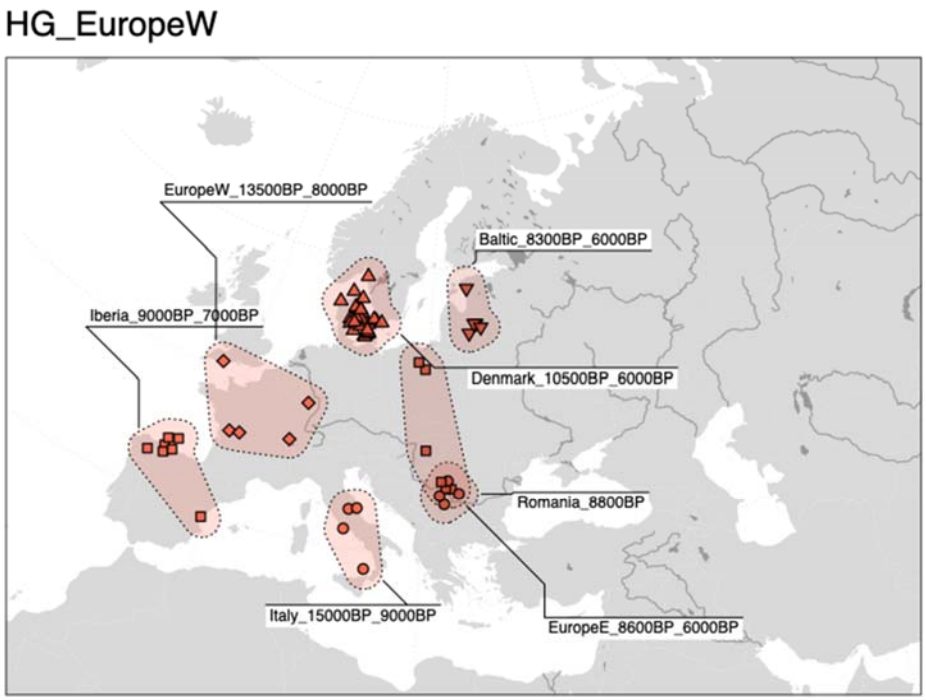


Fig. S3f.6 Geographic distribution of individuals in cluster group *HG_EuropeW*. Geographic locations of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

EuropeWCAsia_25000BP_300BP

This global cluster includes individuals from western Eurasia with ancestry related to Mesolithic and Neolithic Near Eastern groups. The individual genetic clusters are partitioned into a total of eight cluster groups, including all clusters of “Neolithic farmers” and “Caucasus hunter-gatherers” previously used in the literature (Fig. S3f.7-20).

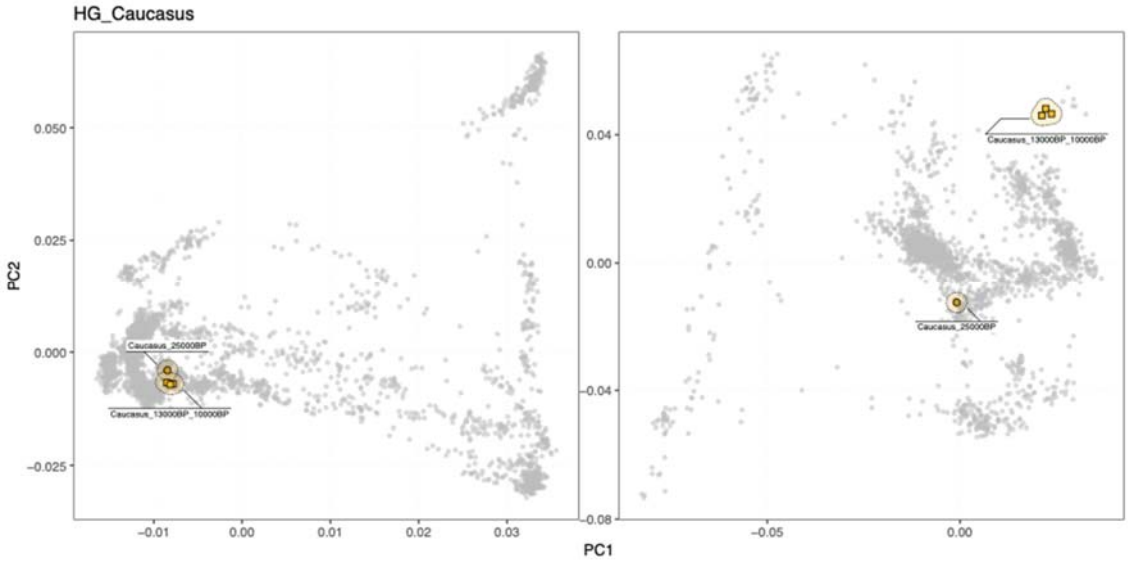


Fig. S3f.7 PCA for cluster group *HG_Caucasus*. PCA positions of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

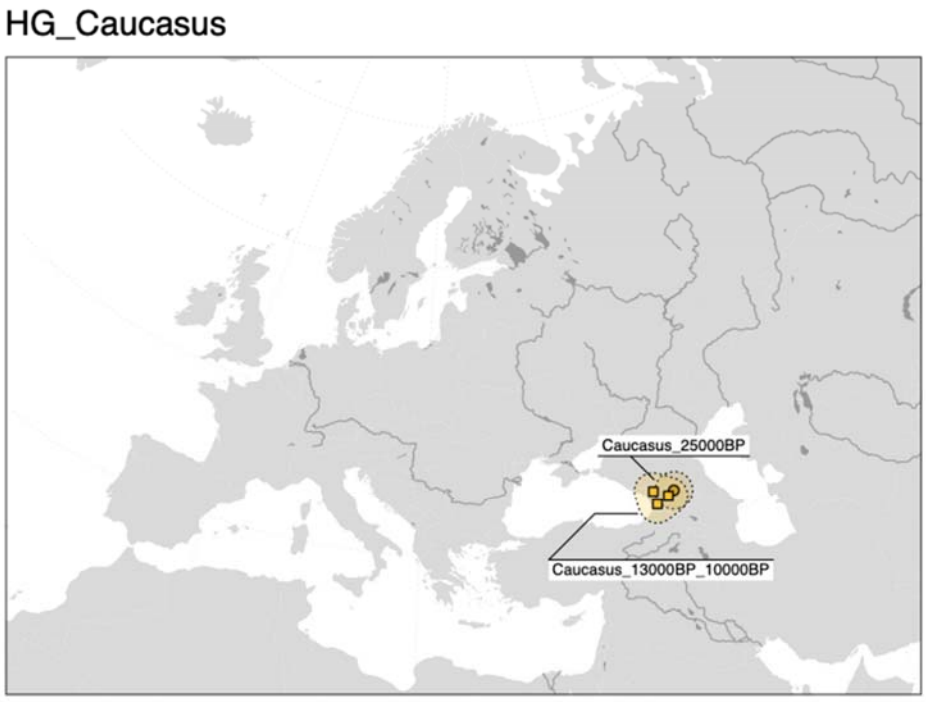


Fig. S3f.8 Geographic distribution of individuals in cluster group *HG_Caucasus.* Geographic locations of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

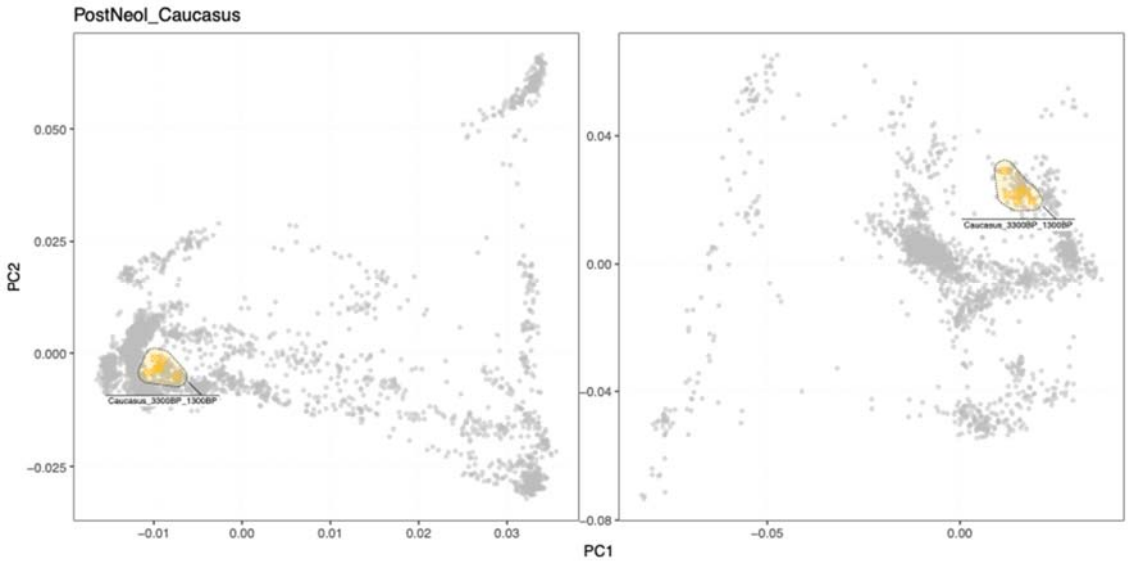


Fig. S3f.9 PCA for cluster group *PostNeol_Caucasus*. PCA positions of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

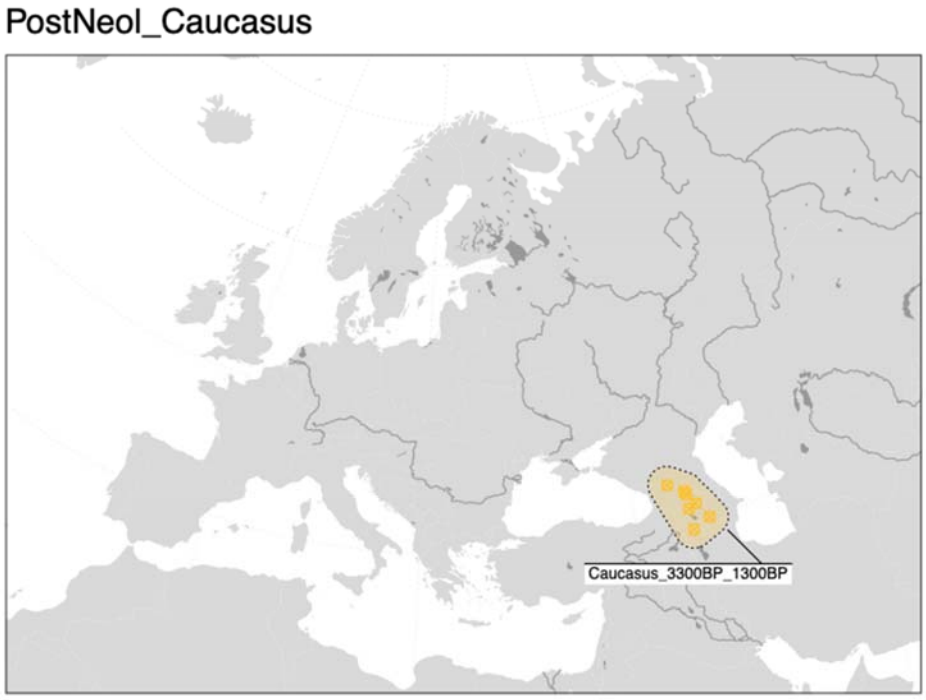


Fig. S3f.10 Geographic distribution of individuals in cluster group *PostNeol_Caucasus.* Geographic locations of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

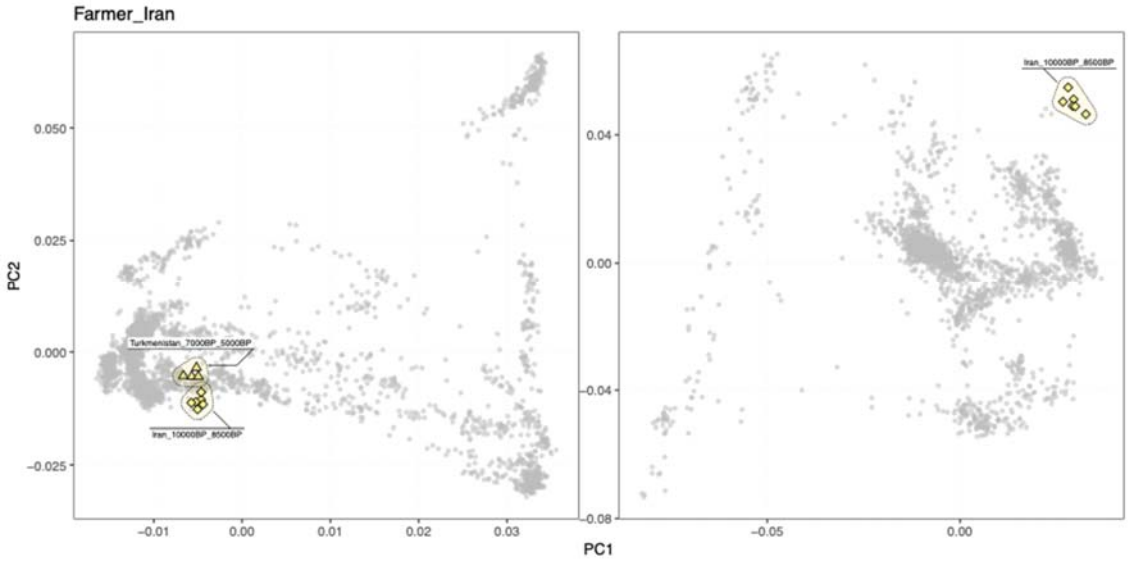


Fig. S3f.11 PCA for cluster group *Farmer_Iran*. PCA positions of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

Fig. S3f.12 Geographic distribution of individuals in cluster group *Farmer_Iran.* Geographic locations of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

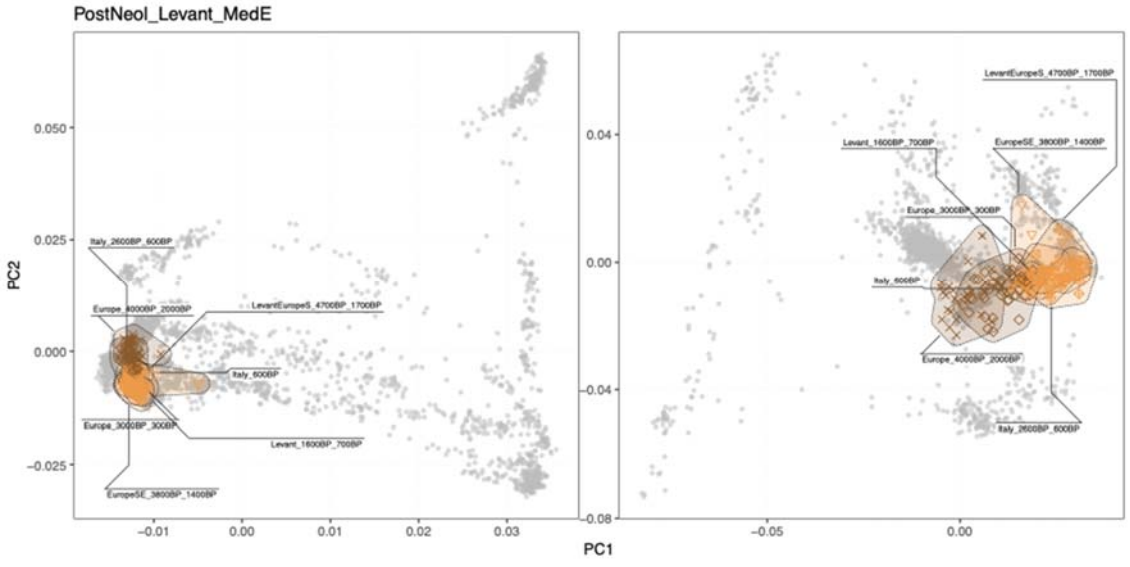


Fig. S3f.13 PCA for cluster group *PostNeol_Levant_MedE*. PCA positions of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

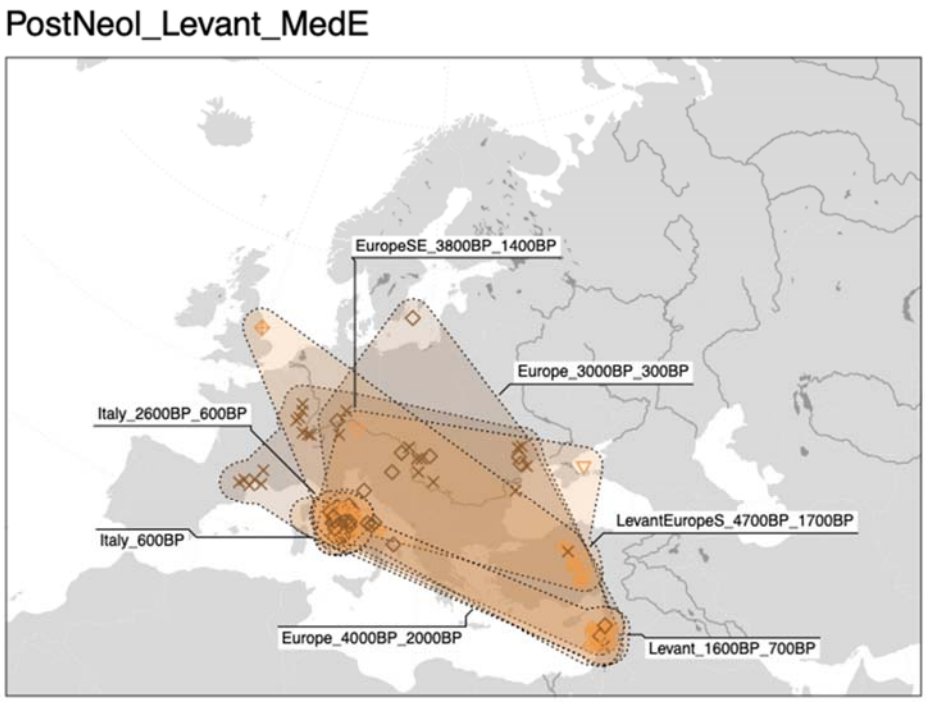


Fig. S3f.14 Geographic distribution of individuals in cluster group *PostNeol_Levant_MedE.* Geographic locations of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

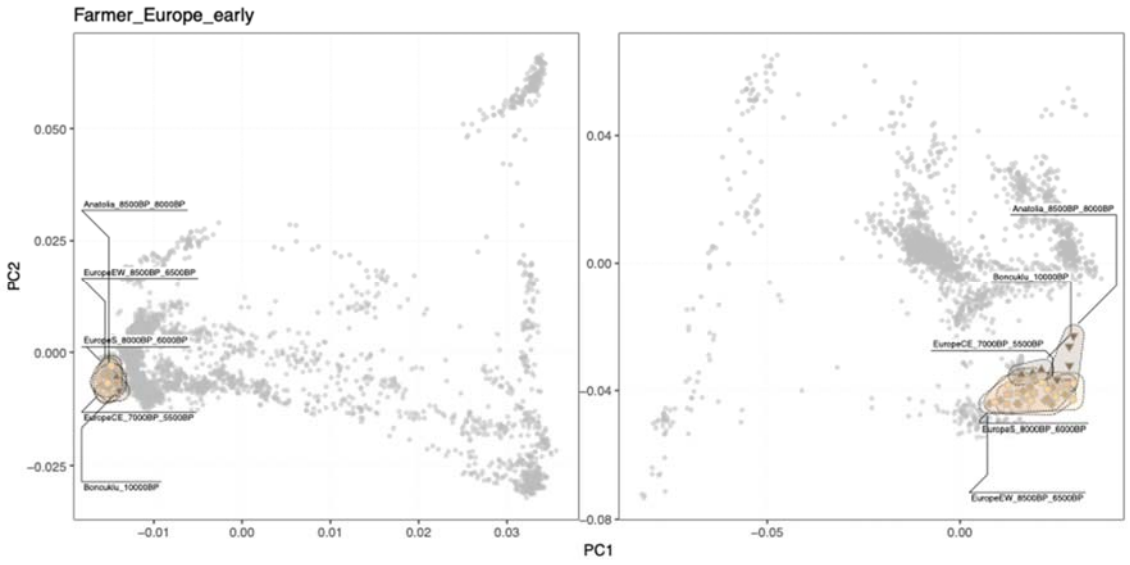


Fig. S3f.15 PCA for cluster group *Farmer_Europe_early*. PCA positions of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

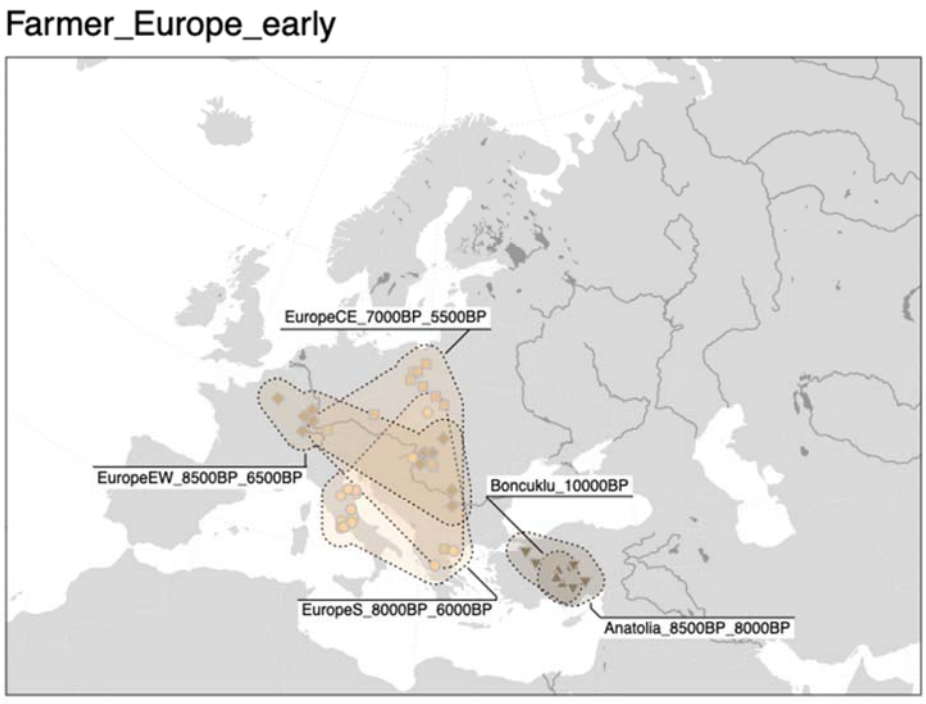


Fig. S3f.16 Geographic distribution of individuals in cluster group *Farmer_Europe_early.* Geographic locations of individuals within specific clusters are

highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

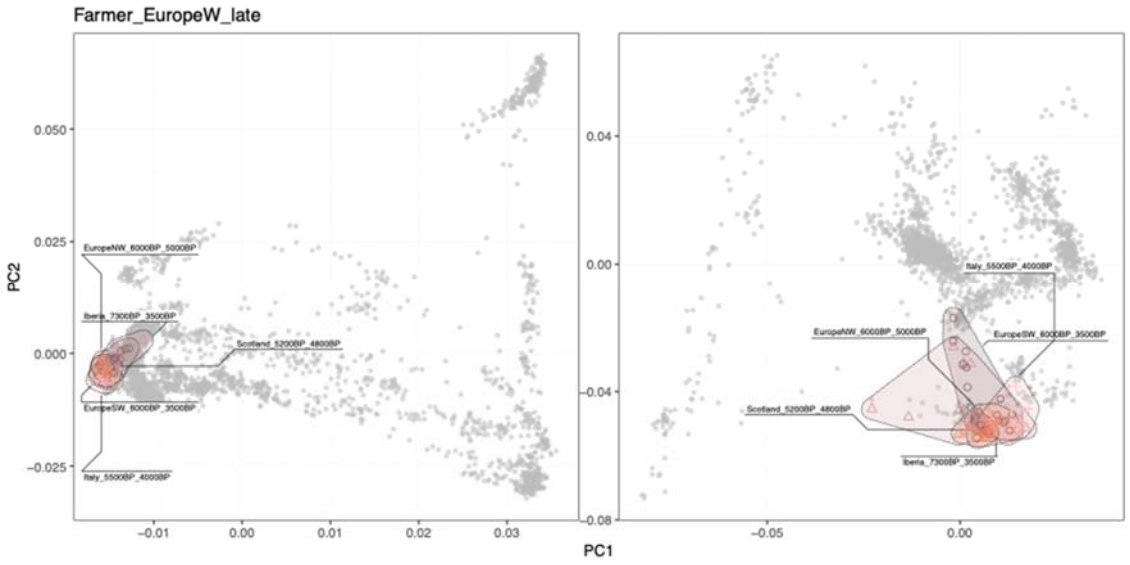


Fig. S3f.17 PCA for cluster group *Farmer_EuropeW_late*. PCA positions of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

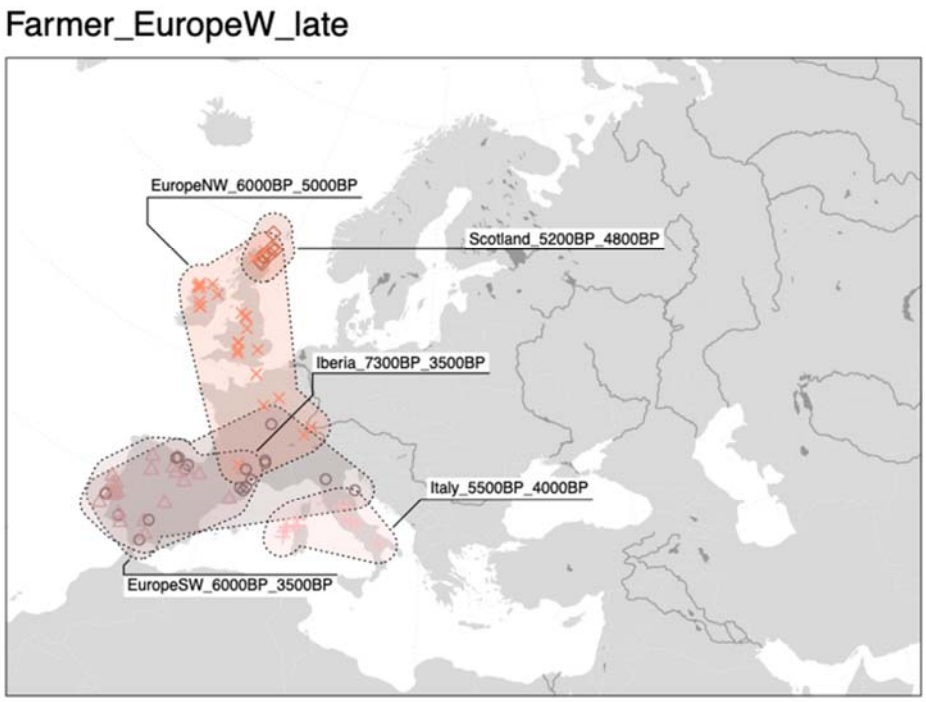


Fig. S3f.18 Geographic distribution of individuals in cluster group *Farmer_EuropeW_late.* Geographic locations of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

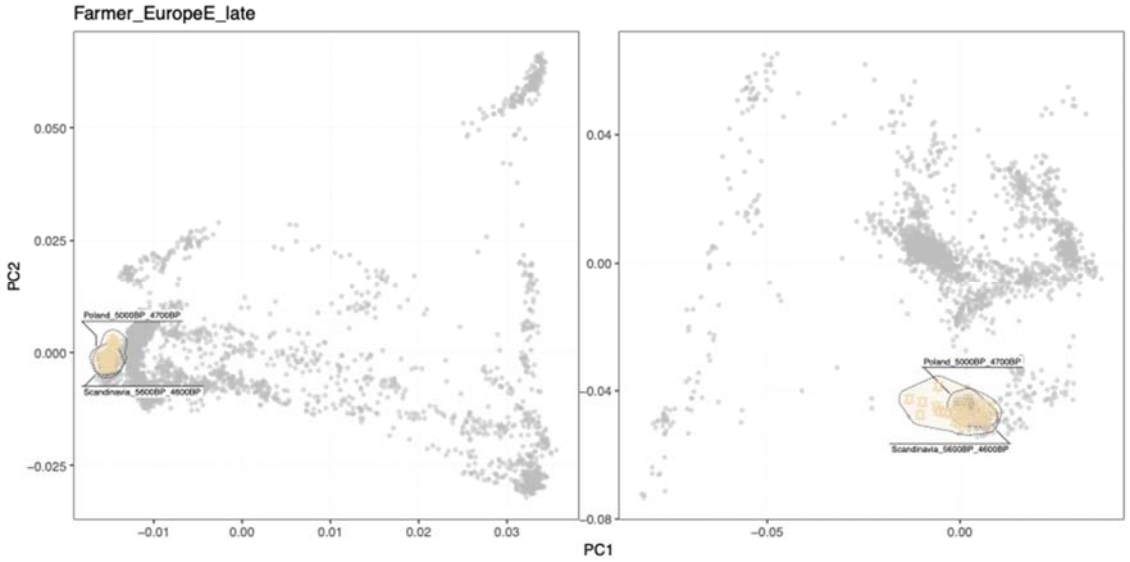


Fig. S3f.19 PCA for cluster group *Farmer_EuropeE_late*. PCA positions of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

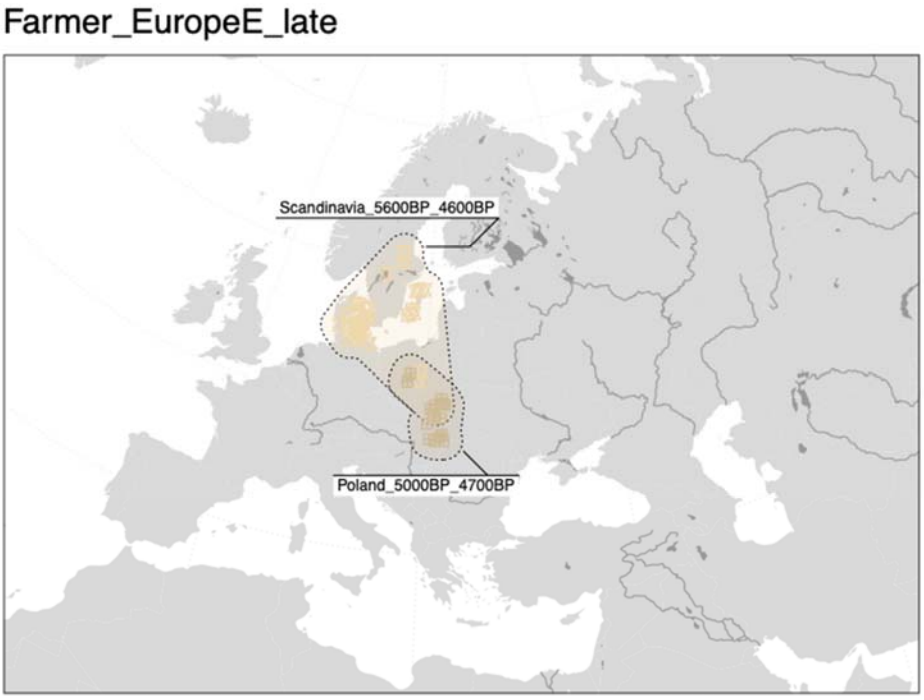


Fig. S3f.20 Geographic distribution of individuals in cluster group *Farmer_EuropeE_late.* Geographic locations of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

Eurasia_5000BP_200BP

This global cluster includes individuals from western Eurasia from the Bronze Age onwards. The individual genetic clusters are partitioned into a total of five cluster groups, including all clusters of individuals with “Steppe ancestry” related to Bronze Age Steppe pastoralists previously used in the literature (Fig. S3f.21-28).

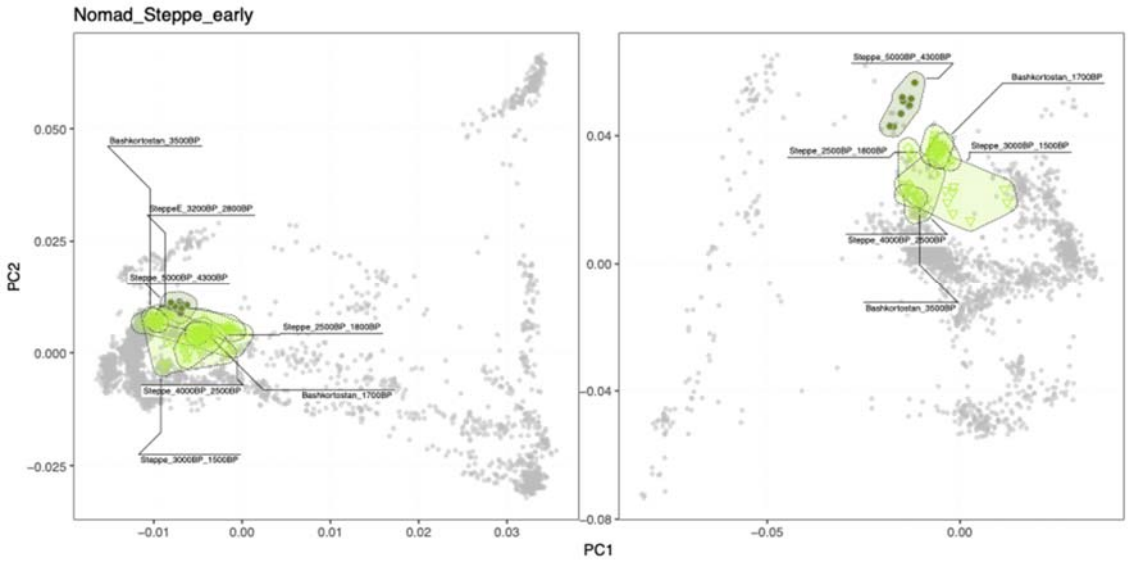


Fig. S3f.21 PCA for cluster group *Nomad_Steppe_early*. PCA positions of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

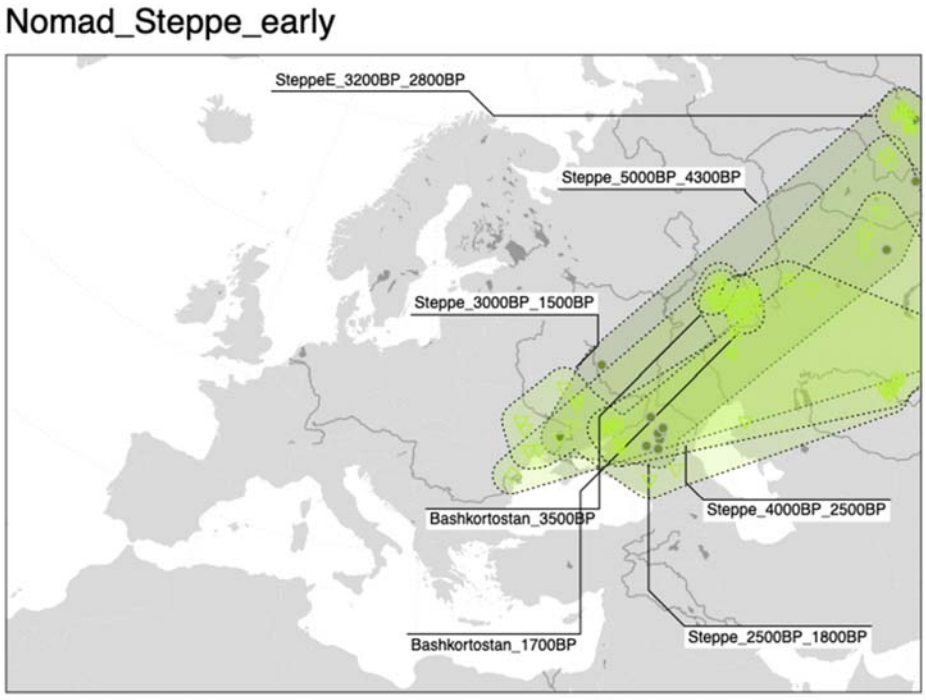


Fig. S3f.22 Geographic distribution of individuals in cluster group *Nomad_Steppe_early.* Geographic locations of individuals within specific clusters are

highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

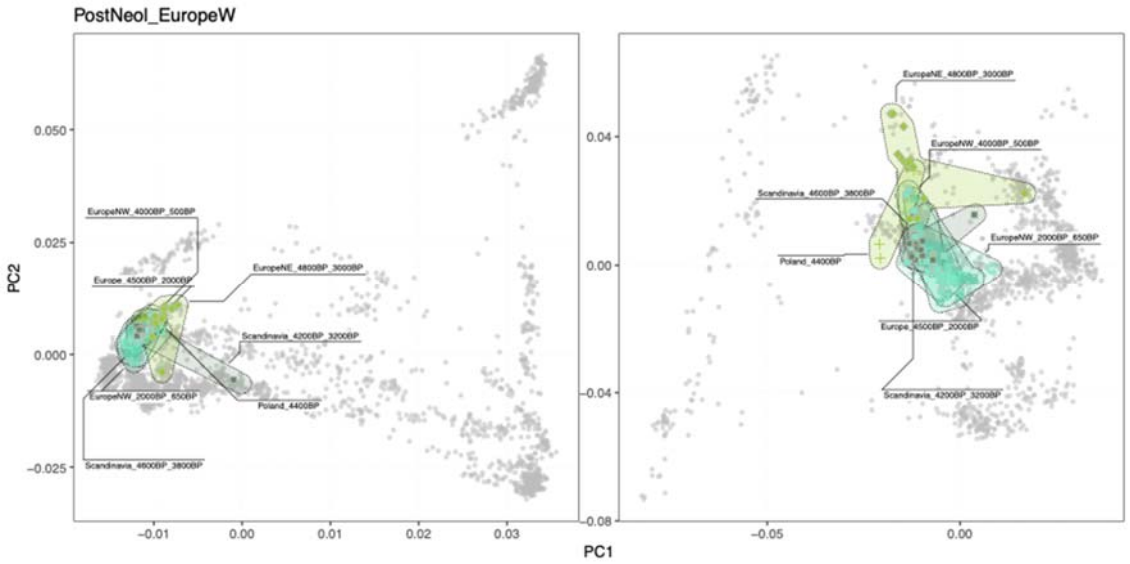


Fig. S3f.23 PCA for cluster group *PostNeol_EuropeW*. PCA positions of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

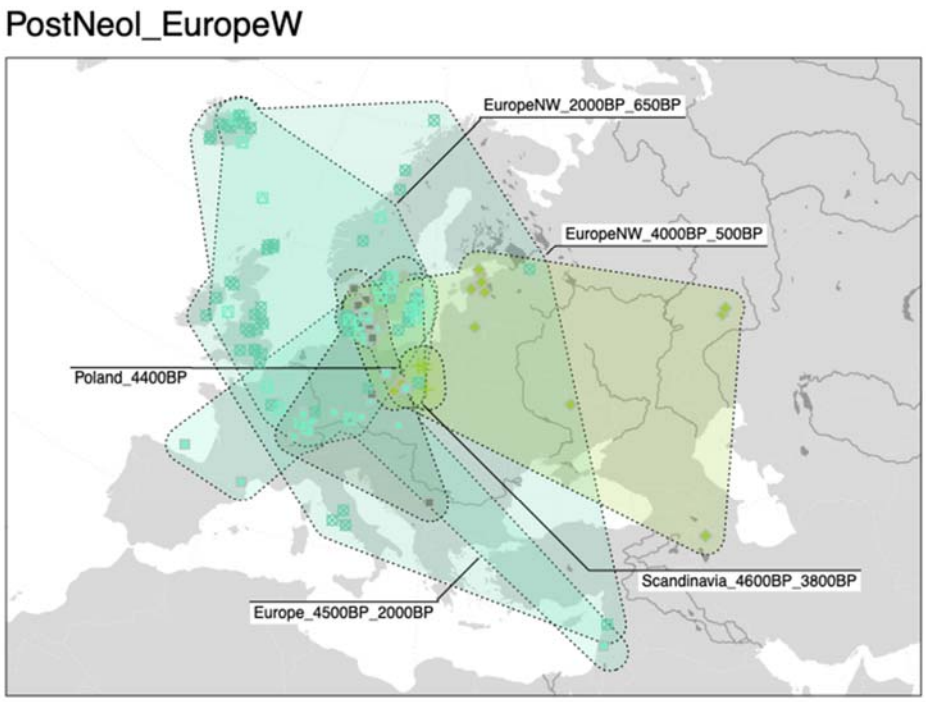


Fig. S3f.24 Geographic distribution of individuals in cluster group *PostNeol_EuropeW.* Geographic locations of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

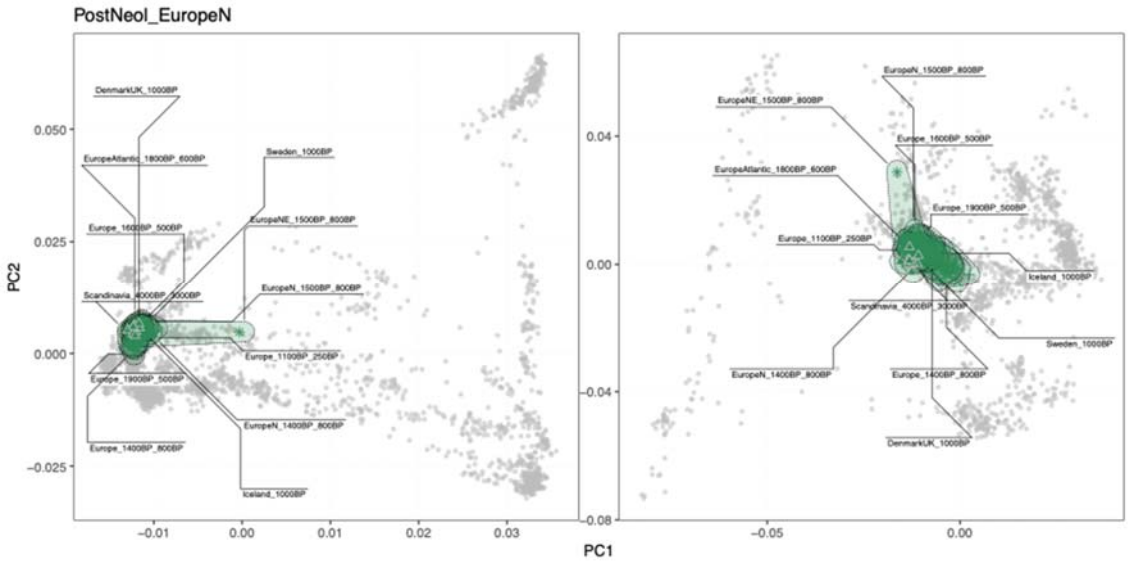


Fig. S3f.25 PCA for cluster group *PostNeol_EuropeN.* PCA positions of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

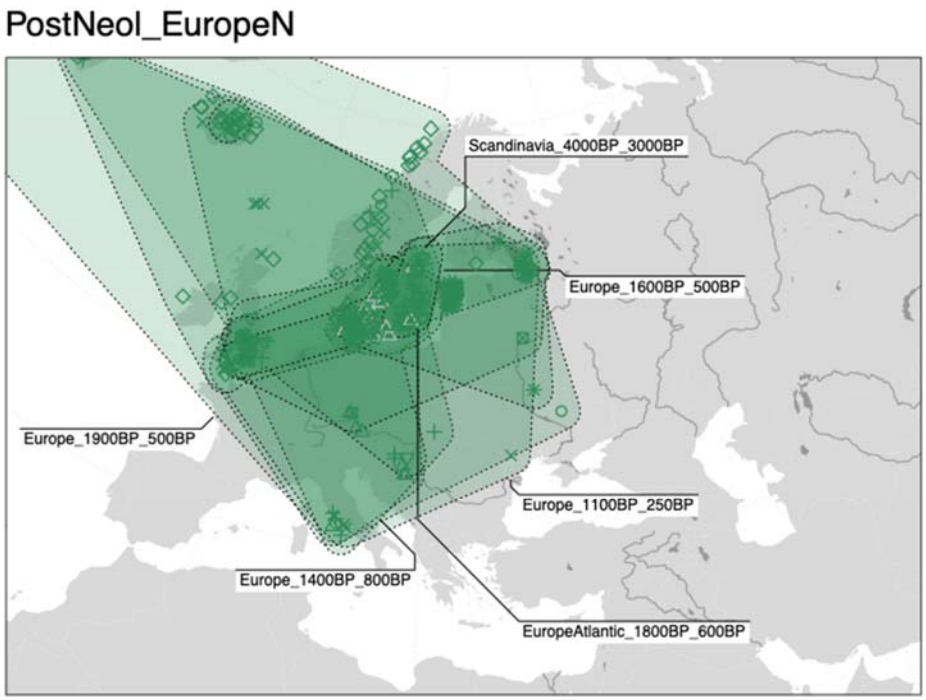


Fig. S3f.26 Geographic distribution of individuals in cluster group *PostNeol_EuropeN.* Geographic locations of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

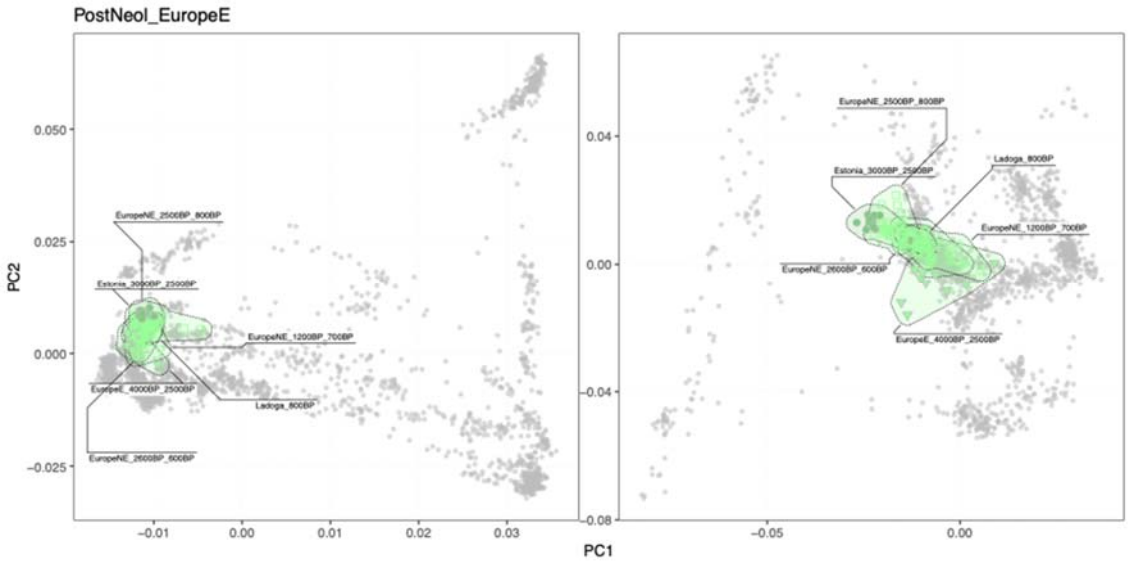


Fig. S3f.27 PCA for cluster group *PostNeol_EuropeE*. PCA positions of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

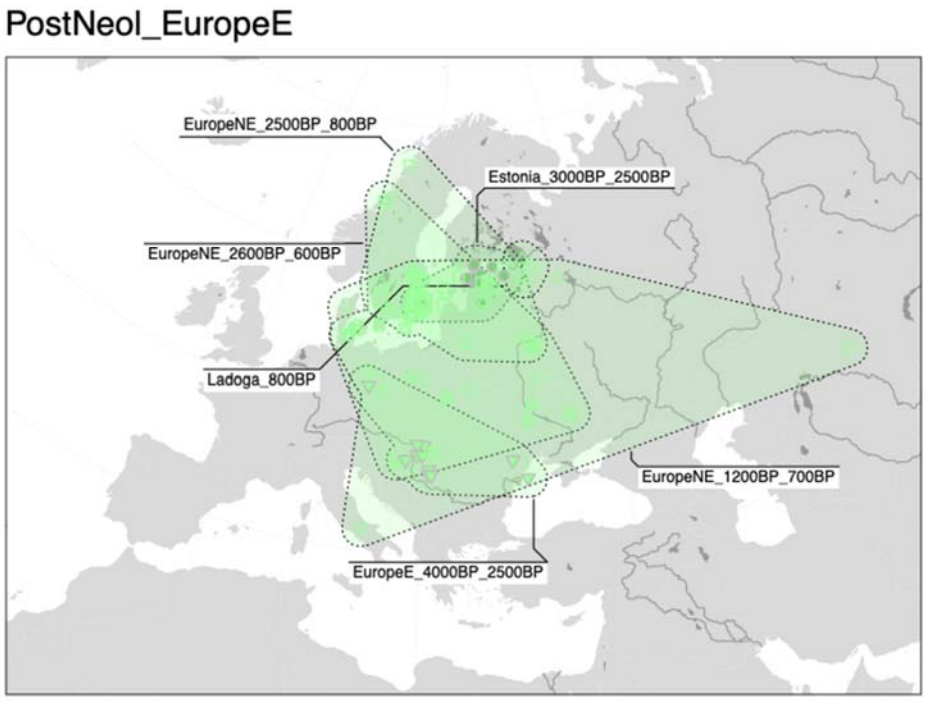


Fig. S3f.28 Geographic distribution of individuals in cluster group *PostNeol_EuropeE.* Geographic locations of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

Asia_45000BP_200BP

This global cluster includes diverse sets of clusters of individuals from the Southeast-, East- and North Asia, broadly characterised by “east Eurasian” ancestry. The individual genetic clusters are partitioned into a total of 12 cluster groups (Fig. S3f.29-44).

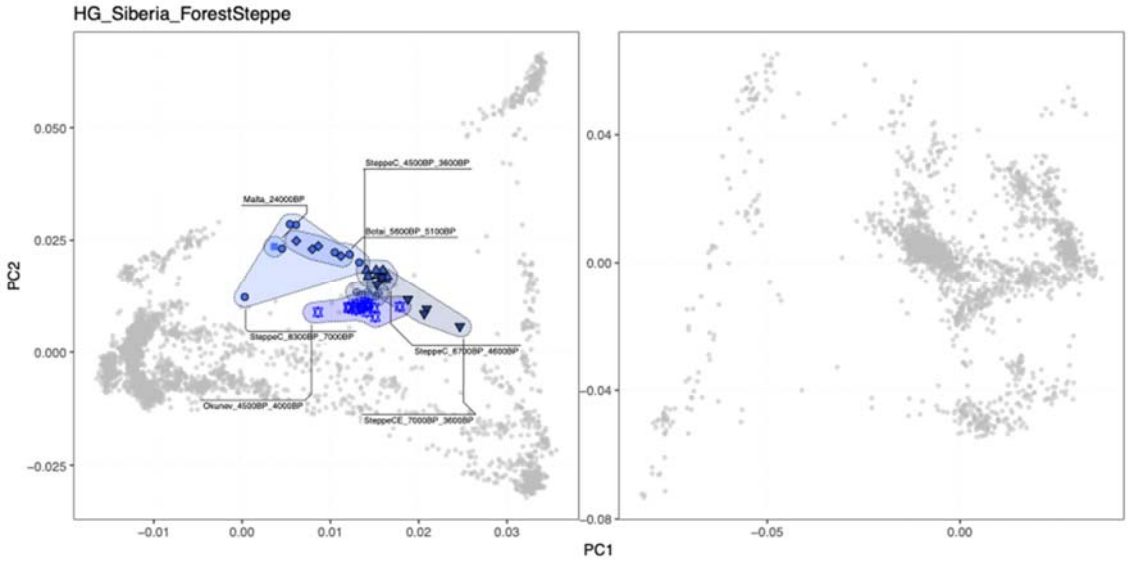


Fig. S3f.29 PCA for cluster group *HG_Siberia_ForestSteppe*. PCA positions of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

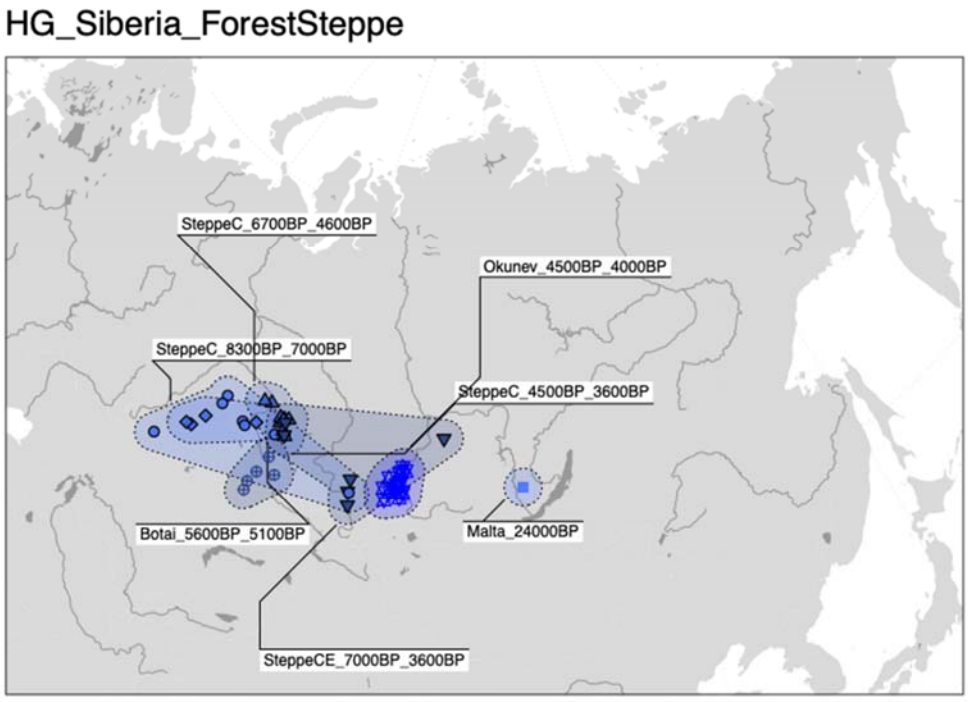


Fig. S3f.30 **Geographic distribution of individuals in cluster group *HG_Siberia_ForestSteppe.*** Geographic locations of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

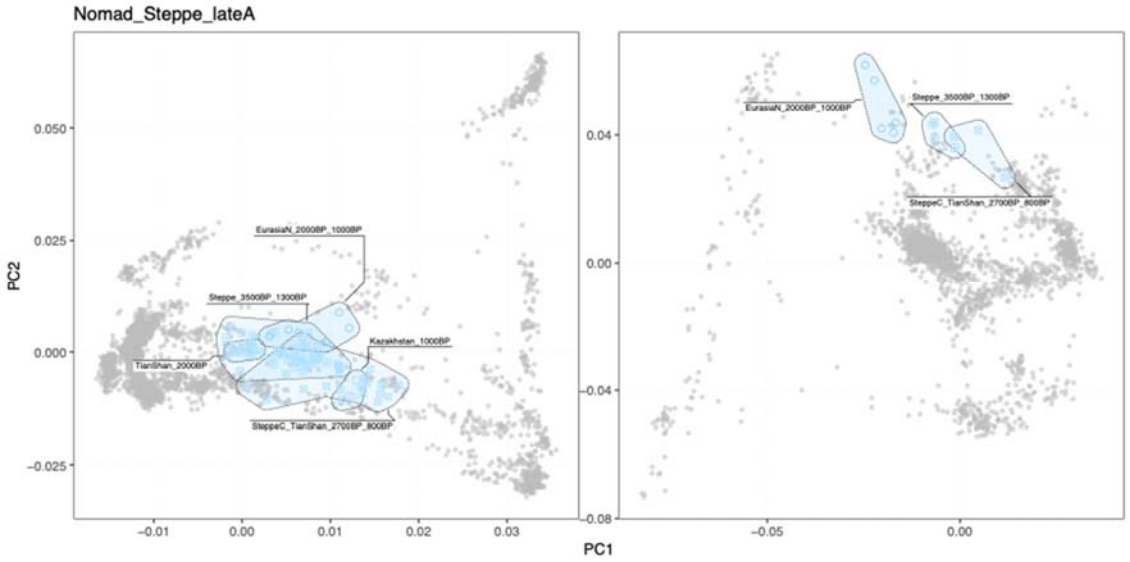


Fig. S3f.31 **PCA for cluster group *Nomad_Steppe_lateA***. PCA positions of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

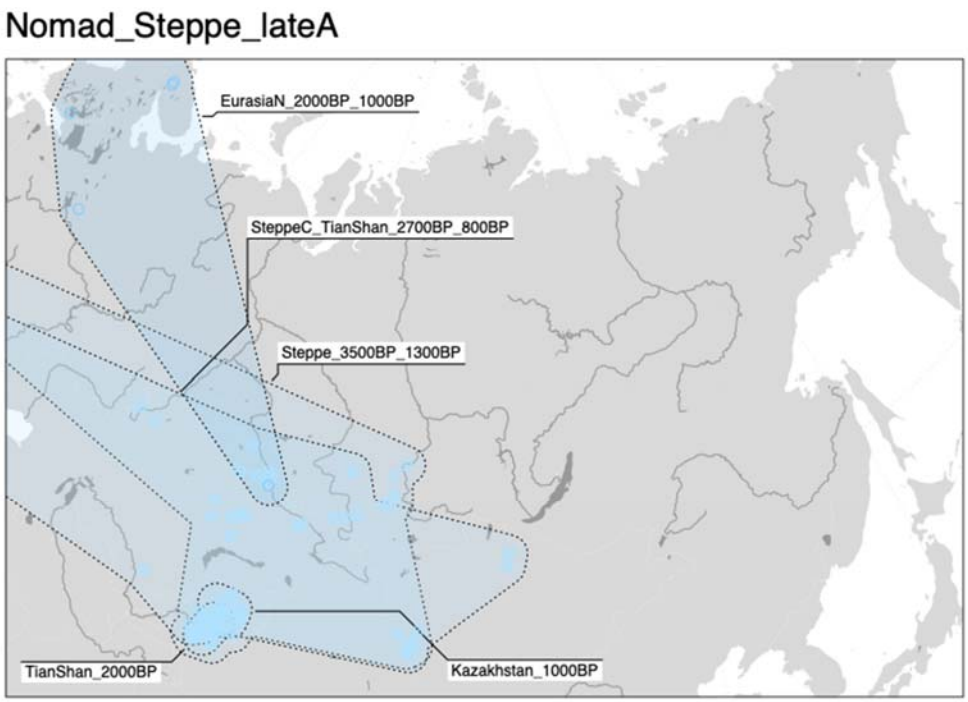


Fig. S3f.32 Geographic distribution of individuals in cluster group *Nomad_Steppe_lateA.* Geographic locations of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

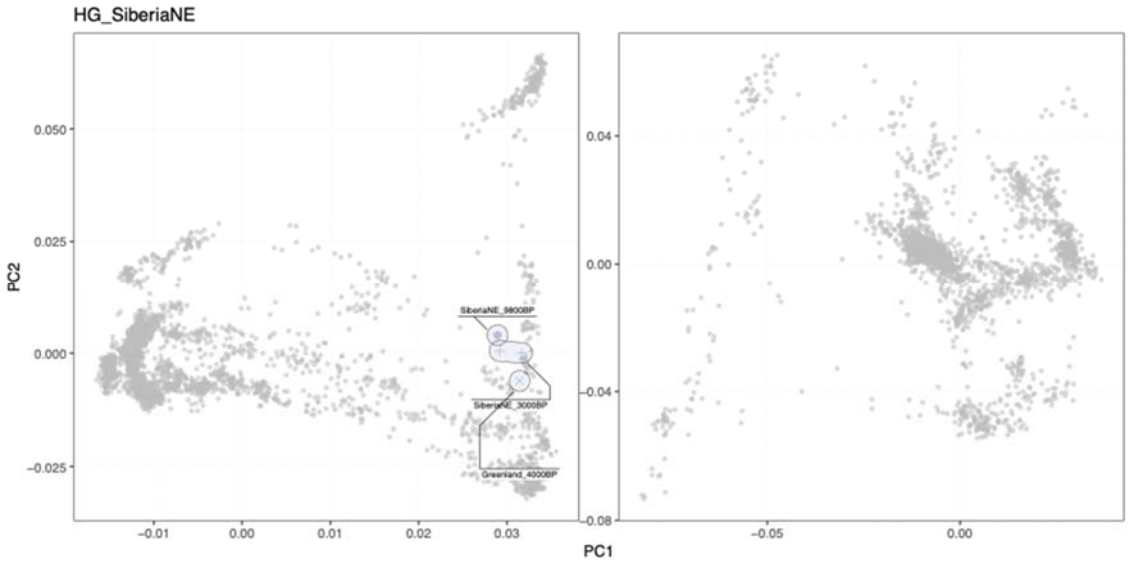


Fig. S3f.33 PCA for cluster group *HG_SiberiaNE*. PCA positions of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

Fig. S3f.34 Geographic distribution of individuals in cluster group *HG_SiberiaNE.* Geographic locations of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

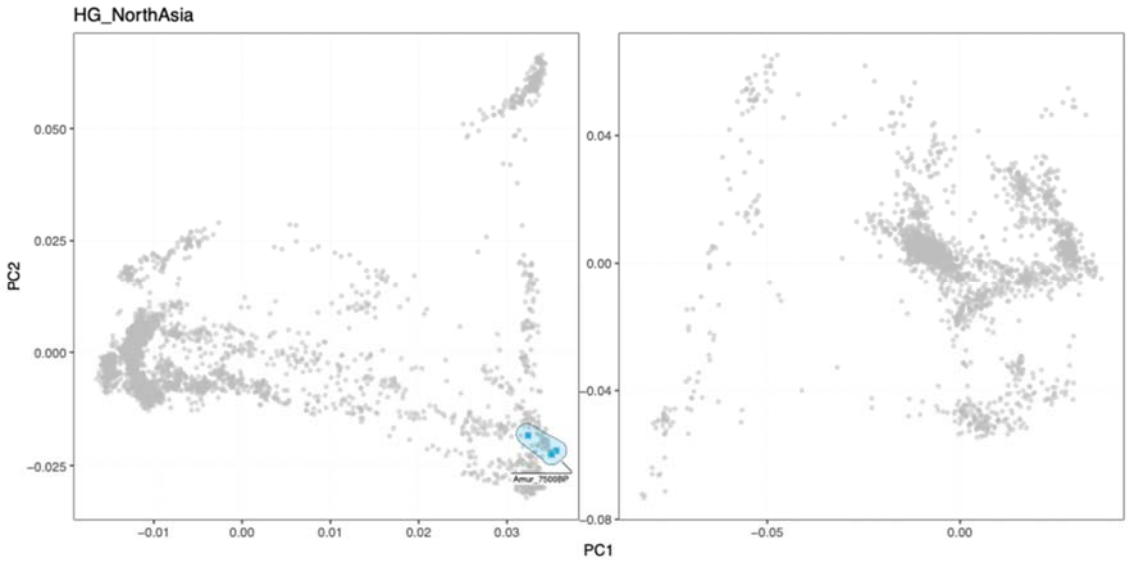


Fig. S3f.35 PCA for cluster group *HG_NorthAsia*. PCA positions of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

Fig. S3f.36 Geographic distribution of individuals in cluster group *HG_NorthAsia.* Geographic locations of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

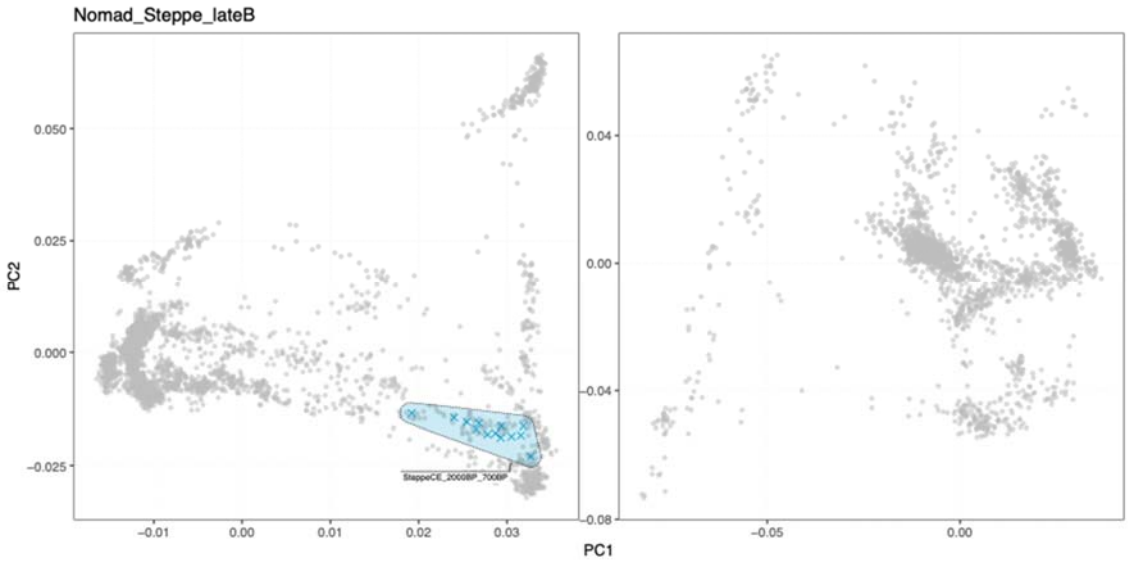


Fig. S3f.37 PCA for cluster group *Nomad_Steppe_lateB.* PCA positions of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

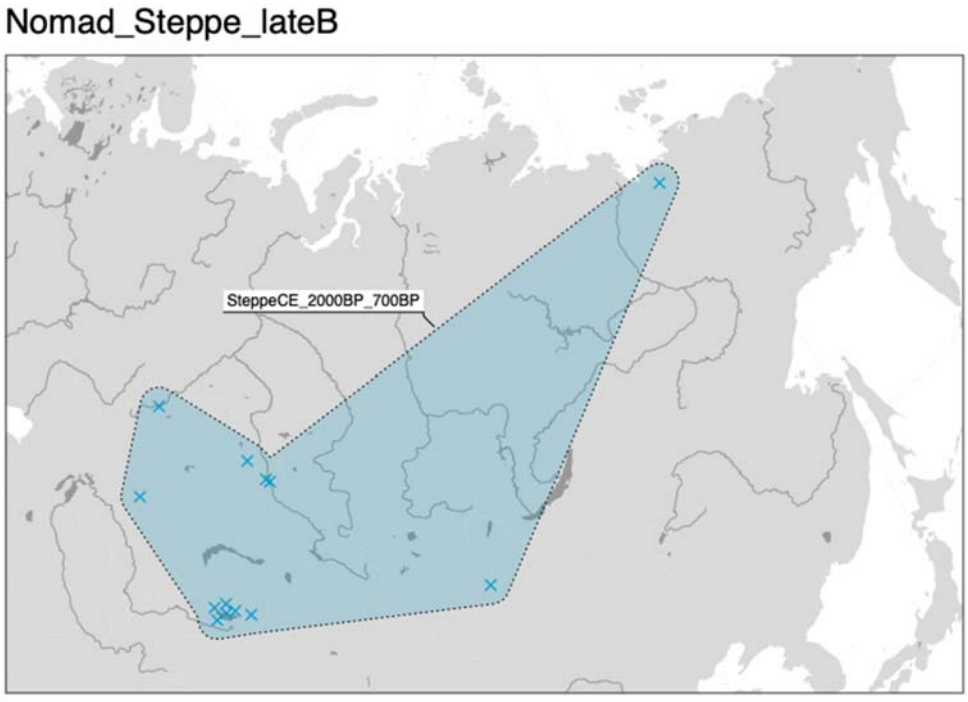


Fig. S3f.38 Geographic distribution of individuals in cluster group *Nomad_Steppe_lateB.* Geographic locations of individuals within specific clusters are

highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

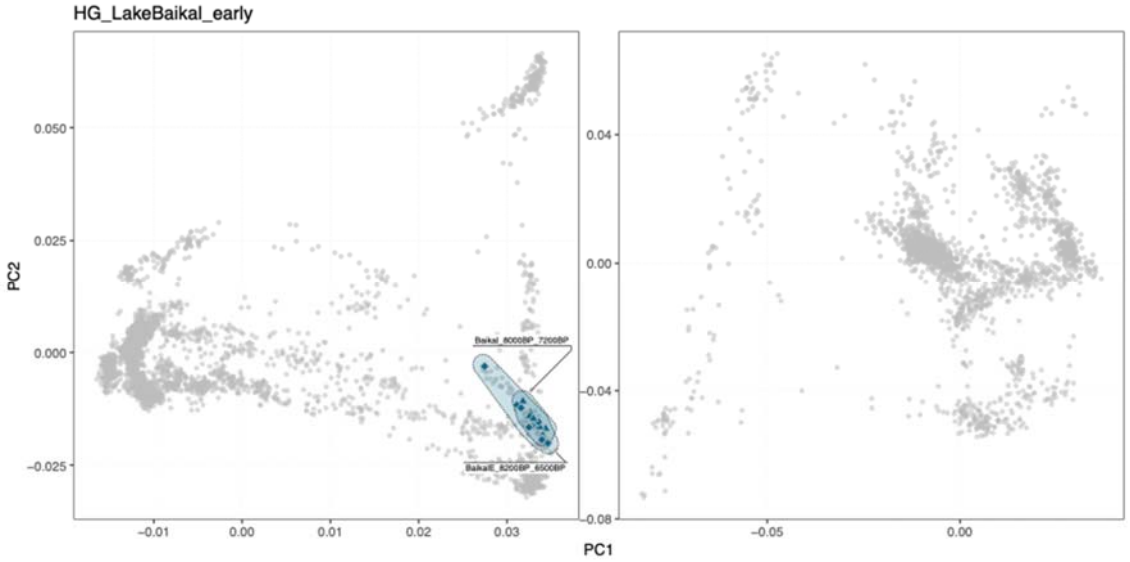


Fig. S3f.39 PCA for cluster group *HG_LakeBaikal_early*. PCA positions of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

Fig. S3f.40 Geographic distribution of individuals in cluster group *HG_LakeBaikal_early*. Geographic locations of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

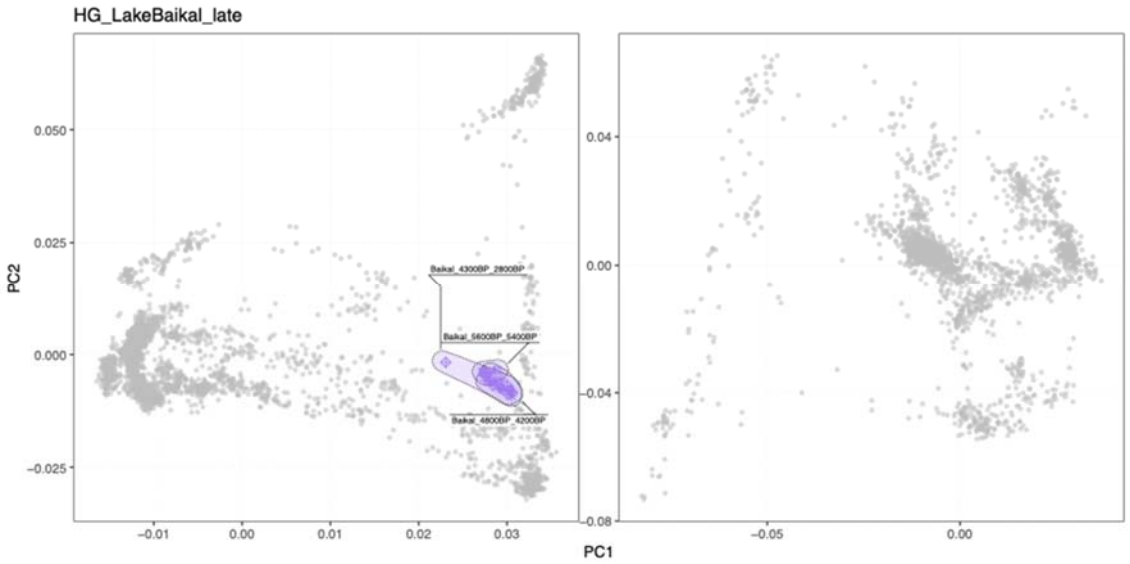


Fig. S3f.41 PCA for cluster group *HG_LakeBaikal_late*. PCA positions of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

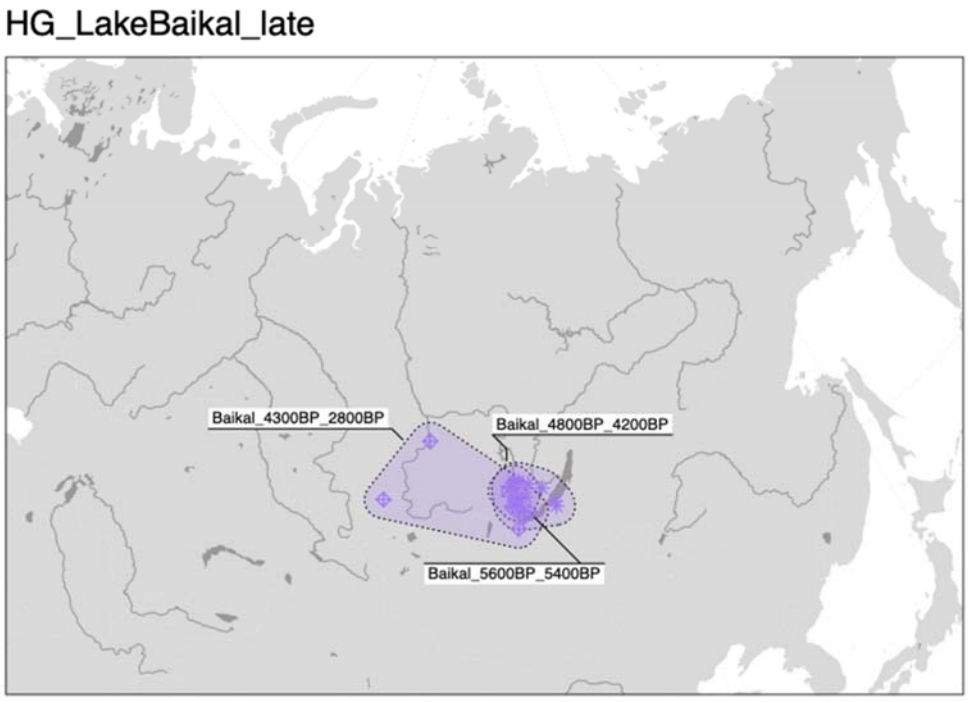


Fig. S3f.42 Geographic distribution of individuals in cluster group *HG_LakeBaikal_late.* Geographic locations of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

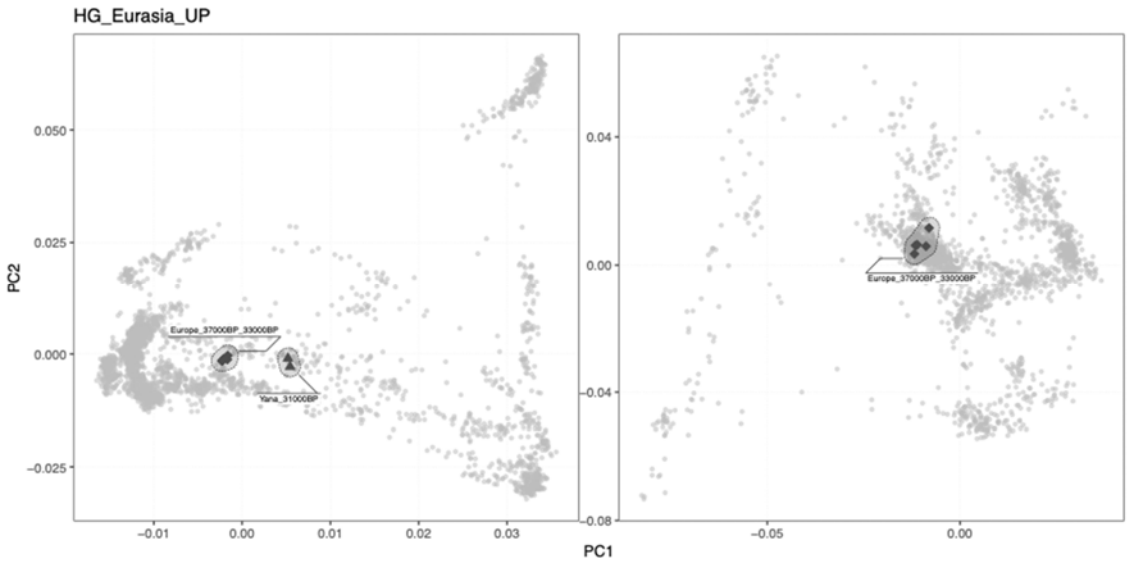


Fig. S3f.43 PCA for cluster group *HG_Eurasia_UP*. PCA positions of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

Fig. S3f.44 Geographic distribution of individuals in cluster group *HG_Eurasia_UP.* Geographic locations of individuals within specific clusters are highlighted with colored symbols, and connected with shaded hulls (from PCAs shown in Fig. S3d.7).

Runs of homozygosity and IBD sharing within clusters

We quantified genetic relatedness within clusters by investigating runs of homozygosity and pairwise IBD sharing among cluster individuals. The analyses showed broad differences in patterns of genetic relatedness between different cluster groups across Eurasia, associated with both spatiotemporal and subsistence contexts of the individuals. The highest amounts of IBD sharing and ROH were generally found in clusters of individuals from hunter-gatherer contexts. Individuals of comparable age from farming contexts showed lower sharing, consistent with overall higher effective population sizes in farming communities compared to forager groups (Fig. S3f.45-48).

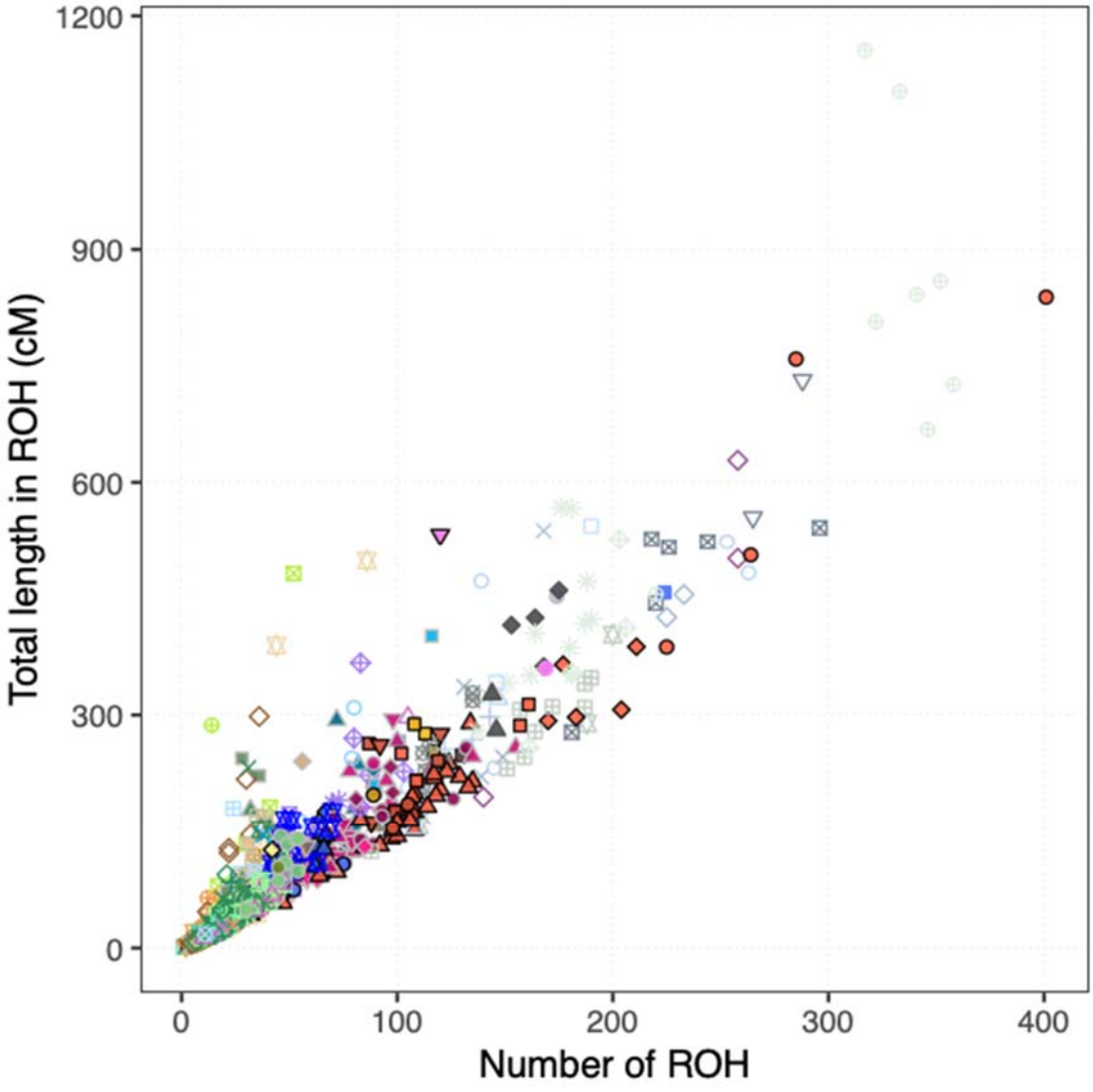


Fig. S3f.45. Number and total length of ROH segments. Plot shows the number and total length of ROH segments detected in the respective ancient individual. Symbol colour and shape indicated genetic cluster membership.

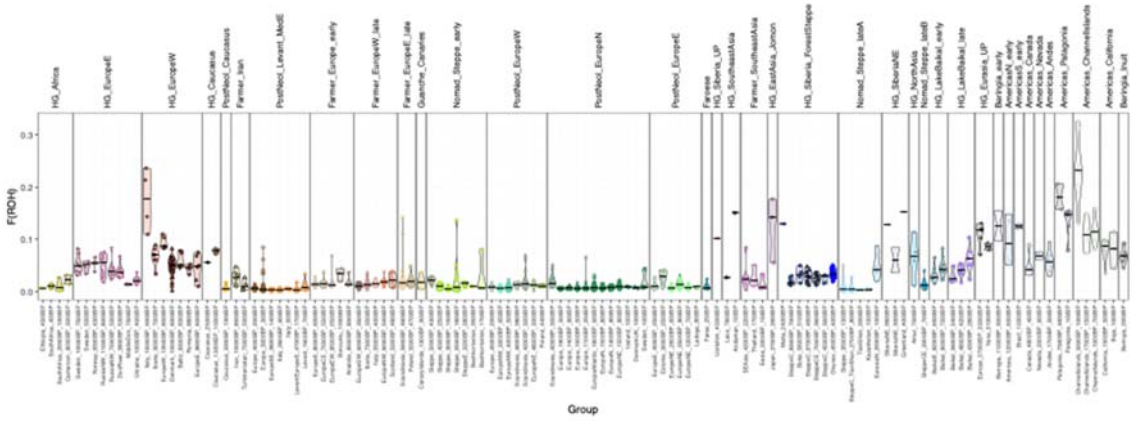


Fig. S3f.46. Distribution of F(ROH). Violin plots and symbols showing the distributions of F(ROH) within genetic clusters.

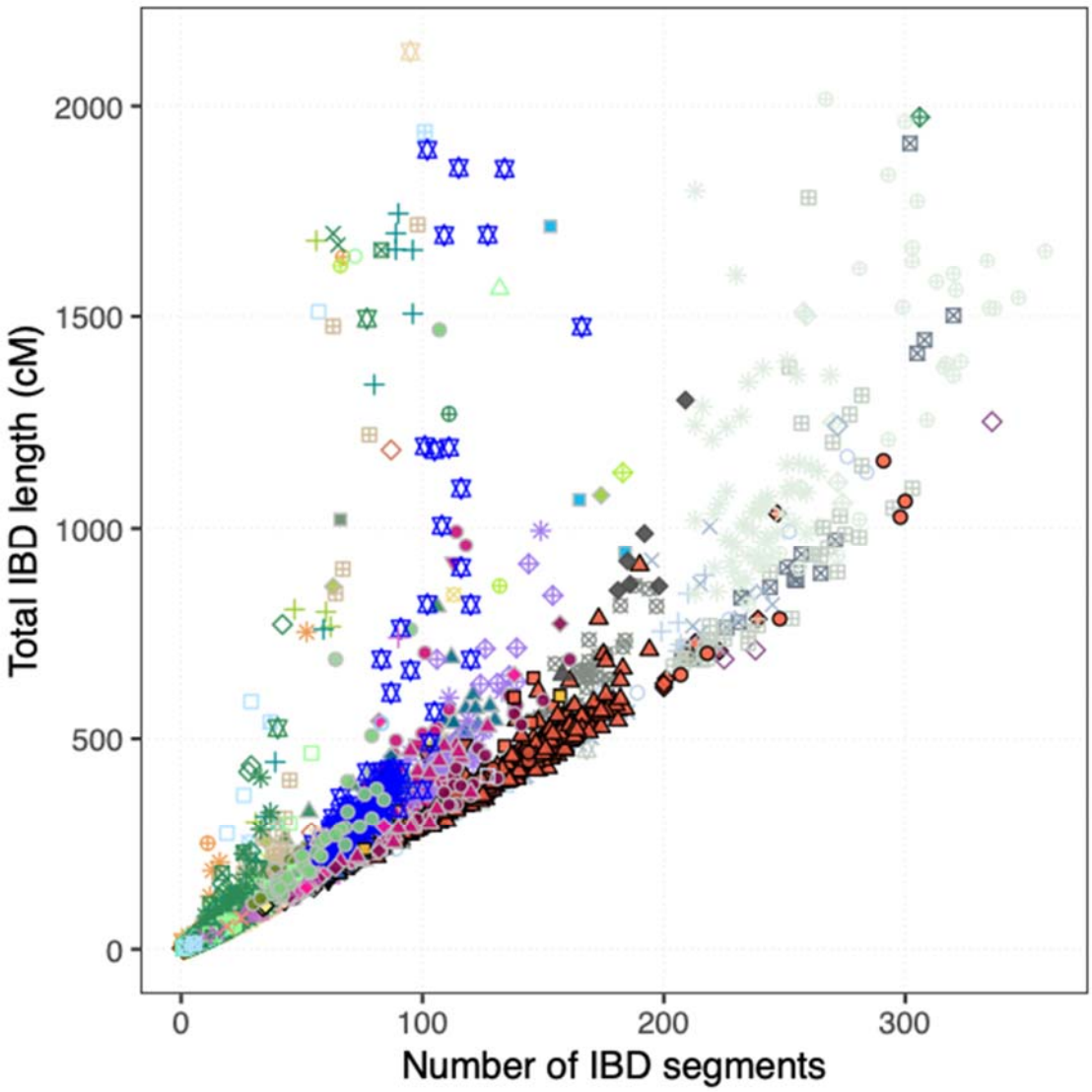


Fig. S3f.47. Number and total length of IBD segments. Plot shows the number and total length of IBD segments detected in the respective pair of individuals. Symbol colour and shape indicated genetic cluster membership.

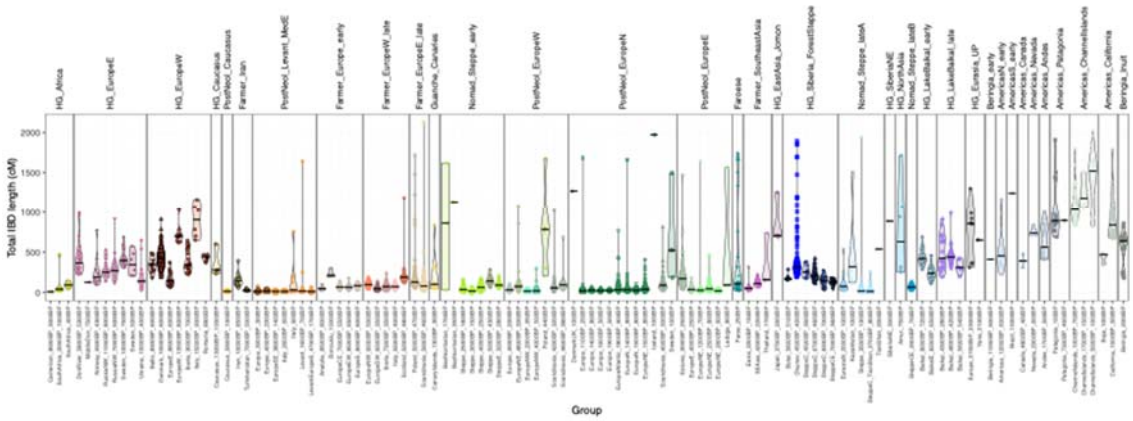


Fig. S3f.48. Violin plots and symbols showing the distributions of total IBD sharing within genetic clusters.

Mixture models

We used IBD "painting profiles" to model sets of target individuals as mixtures of putative source groups. To investigate how these IBD profiles are capturing underlying population structure, we compared their similarities using the "total-variation- distance" (TVD)[12,13] measure. We calculated pairwise TVD values for each pair of individual profiles in the combined ancient and modern dataset, as well as for average profiles aggregated across all individuals within a genetic cluster. Our results show that the painting profiles readily distinguish both broad- and fine-scale genetic differentiation among the individuals and genetic clusters (Fig. S3f.49,50).

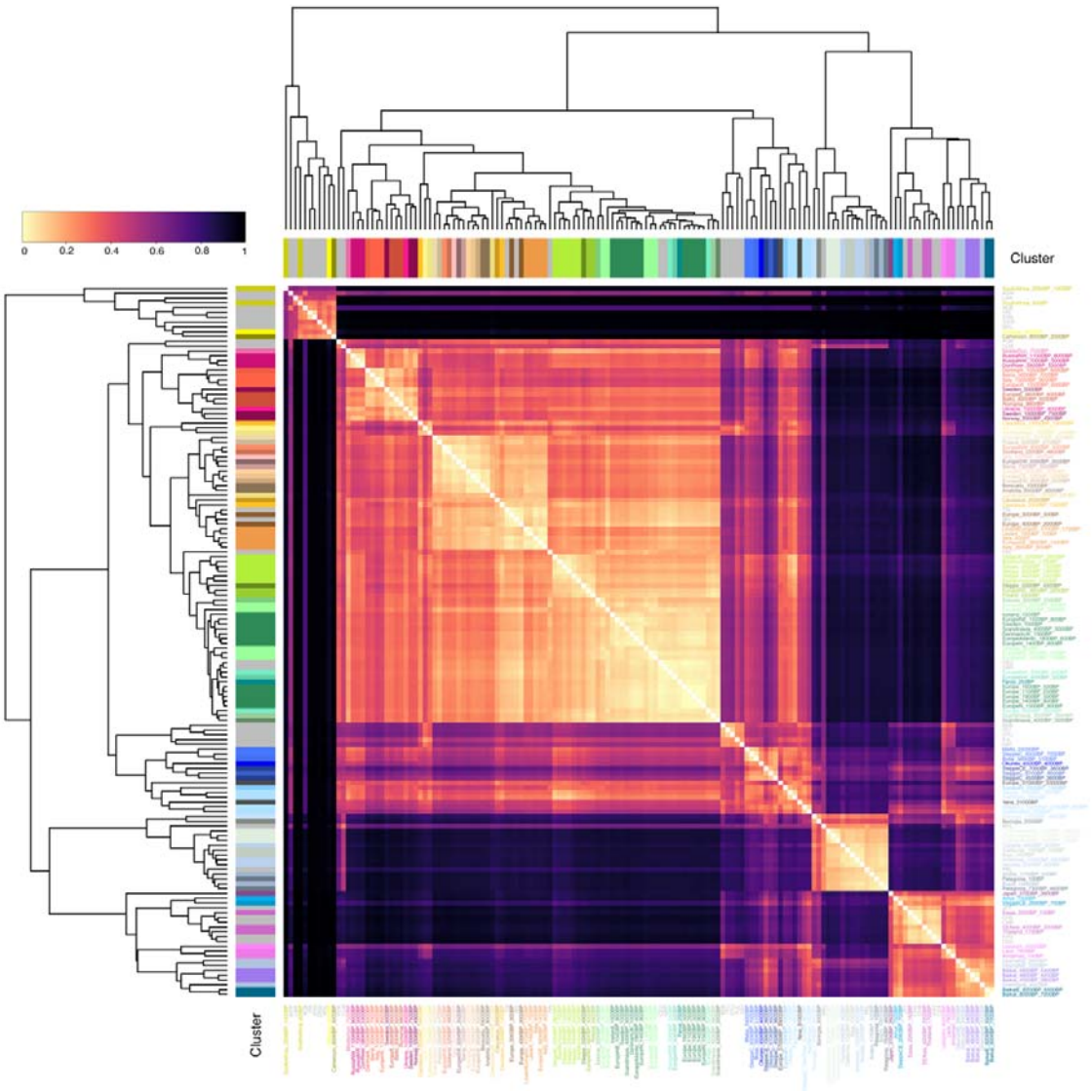


Fig. S3f.49. Cluster IBD painting profile distances. Heatmap showing pairwise distance between genetic cluster IBD painting profiles, measured using TVD. Colored bars indicate cluster membership.

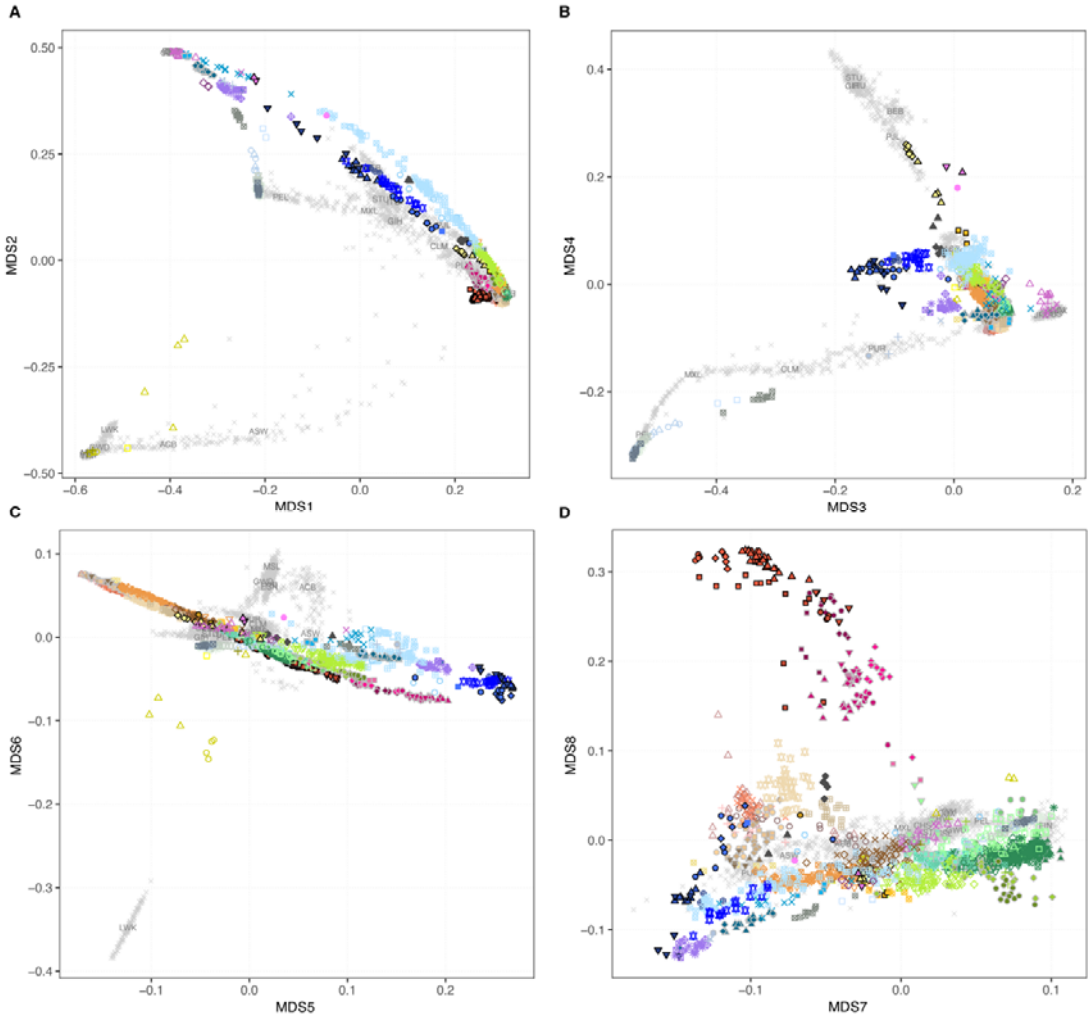


Fig. S3f.49. Genetic structure inferred from IBD painting profiles. (A)-(D) Plots show the first 8 dimensions of a multidimensional scaling (MDS) of individual painting profile TVDs across ancient and modern individuals. Genetic cluster membership for ancient individuals is indicated by symbol colour and shape. Present-day individuals are indicated with grey crosses, with labels indicating population median coordinates.

We used these painting profiles in supervised modelling of target individuals as mixtures from different sets of putative source groups (Fig. S3f.51; Supplementary Table VIII). In each source set analysis, we computed source group painting profiles by averaging the profiles of the included individuals within each source group. We then estimated the mixture and proportions of source profiles that best fits the profile observed in the target individuals, using non-negative least squares (Fig. S3f.52, 53; Supplementary Tables IX-XIV)

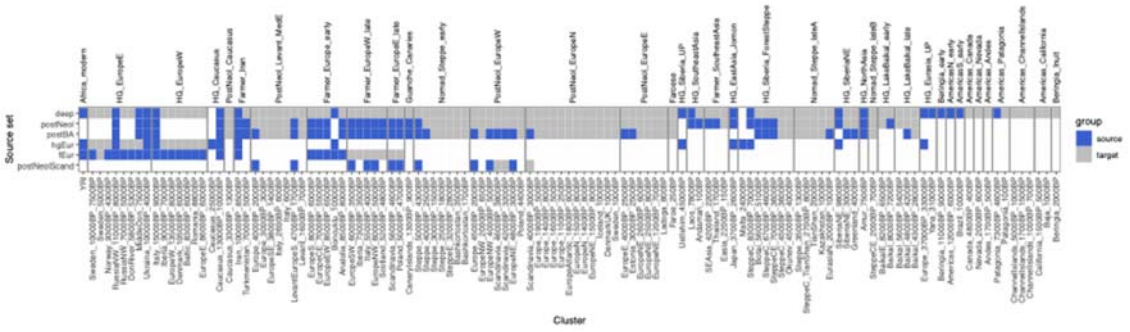
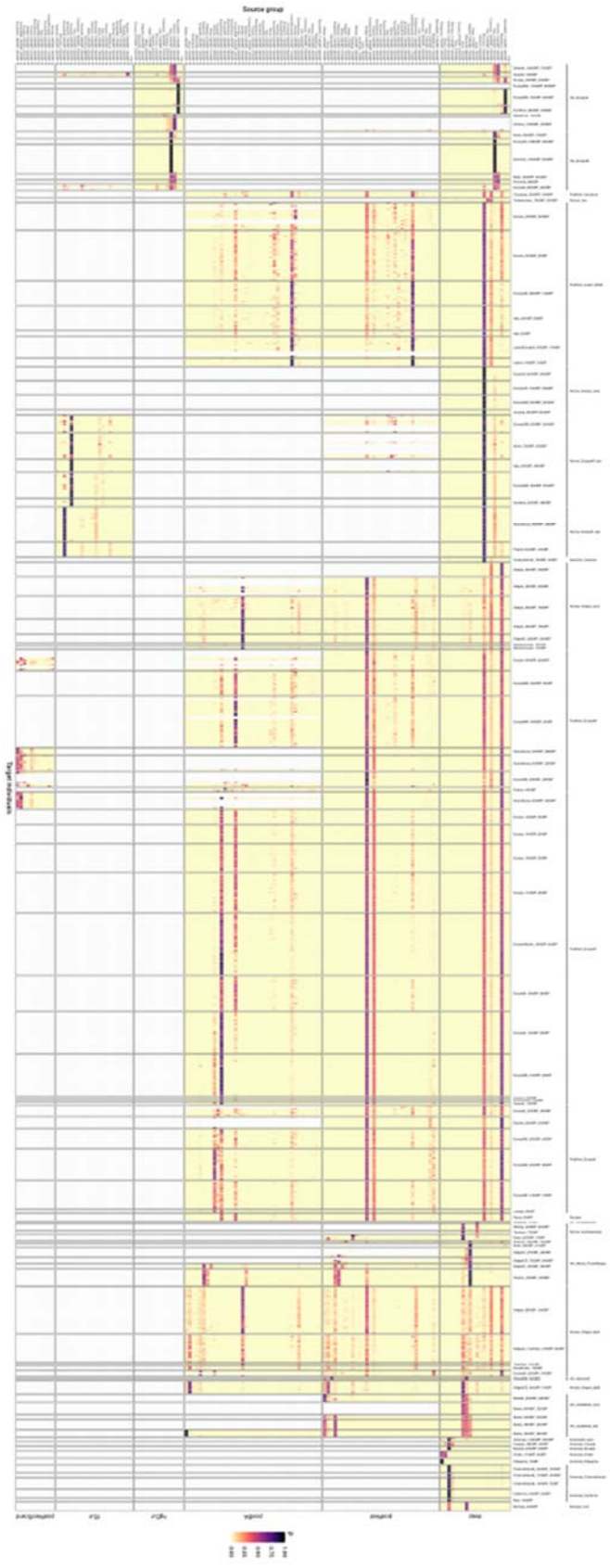


Fig. S3f.52. Mixture model source and target groups. Matrix showing the source and target groups used across the five source set analyses.

1914



1915

Fig. S3f.52. Mixture model results. Heatmap showing estimated ancestry proportions for target individuals (columns) from source groups (rows), across the five source set analyses.

References

1. Browning, B. L. & Browning, S. R. Detecting Identity by Descent and Estimating Genotype Error Rates in Sequence Data. *Am. J. Hum. Genet.* **93**, 840–851 (2013).
2. Browning, S. R. & Browning, B. L. Accurate Non-parametric Estimation of Recent Effective Population Size from Segments of Identity by Descent. *Am. J. Hum. Genet.* **97**, 404–418 (2015).
3. Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLOS Comput. Biol.* **9**, e1003118 (2013).
4. Greenbaum, G., Rubin, A., Templeton, A. R. & Rosenberg, N. A. Network-based hierarchical population structure analysis for large genomic data sets. *Genome Res.* **29**, 2020–2033 (2019).
5. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal* **Complex Systems**, 1695 (2006).
6. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 1–12 (2019).
7. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of Population Structure using Dense Haplotype Data. *PLoS Genet* **8**, e1002453 (2012).
8. Hofmanová, Z. *et al.* Early farmers from across Europe directly descended from Neolithic Aegeans. *Proc. Natl. Acad. Sci.* **113**, 6886–6891 (2016).
9. Hellenthal, G. *et al.* A Genetic Atlas of Human Admixture History. *Science* **343**, 747–751 (2014).
10. Soetaert, K., Meersche, K. V. den & Oevelen, D. van. *limSolve: Solving Linear Inverse Models*. (2009).
11. Eisenmann, S. *et al.* Reconciling material cultures in archaeology with genetic data: The nomenclature of clusters emerging from archaeogenomic analysis. *Sci. Rep.* **8**, 13003 (2018).
12. Leslie, S. *et al.* The fine-scale genetic structure of the British population. *Nature* **519**, 309–314 (2015).
13. Dorp, L. van *et al.* Evidence for a Common Origin of Blacksmiths and Cultivators in the Ethiopian Ari within the Last 4500 Years: Lessons for Clustering-Based Inference. *PLOS Genet.* **11**, e1005397 (2015).

3g) Selecting non-British individuals from the UK Biobank

Will Barrie[1] and Dan Lawson[2]

[1]Zoology Department, University of Cambridge, UK.

[2]School of Mathematics and Integrative Epidemiology Unit, University of Bristol, UK.

Introduction

The UK Biobank (UKB) contains approximately 40,000 individuals not born in the UK. Because many of these individuals are admixed or British, we set up a pipeline to (1) exclude genetically British-like individuals and (2) select individuals of a typical genetic ancestral background for each country, in order to investigate the genetic contribution of each ancient ancestry to modern European, Asian and African populations.

Methods

For individuals from each country in Europe, Asia and Africa but not the UK, (Data-Field 1647: Country of birth (UK/elsewhere) and Data-Field 20115: Country of Birth (non-UK origin)), we ran two density-based scans using Scikit-learn's [1] DBSCAN method (Density-Based Spatial Clustering of Applications with Noise) on a distance matrix constructed using the first 18 PCs[2], weighted by their Eigenvalues. This algorithm finds cores of high density within a distance matrix, which can be any shape, and can include nearby non-core points. The eps parameter can be adjusted to determine how strict the clustering is. This is preferable to using visual PC cut-offs, for which it is difficult to include higher PCs; and k-means clustering, which assumes clusters are convex and all points must be clustered.

In the first scan, designed to remove individuals with British-like ancestry who were born abroad, individuals from a given country were combined with 8,000 random white British individuals, and the clustering algorithm was run on the combined data. Any individuals born abroad who clustered with the white British were excluded (eps=60). For countries that are very similar to Britain in ancestry (e.g. Germany, Denmark) this is a balance between excluding individuals who are genuinely British (very common in the 'German' samples) but not biasing the samples away from British-like ancestry.

In the second scan, the remaining individuals were clustered, and the largest cluster was chosen to represent a typical ancestry for that country. The appropriate eps value (i.e. how strict the clustering should be) is a reflection of the genetic diversity of a country, and so was

adjusted manually to reflect this (Figure S3g.1). In a minority of cases, the major cluster was not the indigenous ancestral background, and so the second-largest was chosen (for example, in Kenya the largest cluster was individuals of Indian origin). All selections were visually verified. Countries that had no obvious main cluster (usually due to low sample numbers) were excluded; any country with 3 or fewer individuals was also excluded.

In order to select Irish individuals (Republic of Ireland and Northern Ireland), step 1 was skipped but step 2 was run with relatively tight parameters, in both cases excluding approximately 20% of individuals.

In order to test the effectiveness of the pipeline at selecting individuals of a similar ancestral background, we looked at the variance in the genome-wide painting proportions for each country. Countries with high variance would indicate recent admixture.

Results

This pipeline selected 24,511 individuals from 126 countries. These selected samples were painted using a reference/donor panel of ancient individuals (Supplementary Note S3h).

The countries that had high variance in ancestry proportions among individuals (and therefore likely that the DBSCAN was not effective in choosing individuals of a similar ancestral background) were Kazakhstan, Yemen, Egypt, Seychelles. Results for these countries should be interpreted with caution.

Discussion

The UKB represents an important source of data for white British people but also for people from other countries globally. Usually, researchers restrict themselves to the white British cohort, but here we develop a method to select individuals from other countries. This transforms the UKB from a resource that is informative about British ancestry to one that can be used to make inferences about populations worldwide.

Country	Number of individuals	Number in wb cluster	eps value	Final number selected
Kenya	1684	277	800	110

Netherlands	491	300	230	153
Switzerland	175	19	230	143
India	4012	358	230	3107
Belgium	158	75	230	70
Singapore	502	262	230	86
Palestine	60	13	700	28
Nigeria	1159	90	800	1016
Hungary	105	0	230	79
Czech Republic	126	2	230	107
Ghana	929	35	700	848
Sri Lanka	744	54	230	620
Egypt	313	82	600	223
Japan	266	12	230	242
Hong Kong	648	107	230	448
Germany	2136	1044	230	1045
Turkey	182	5	400	160
Iran	540	16	230	469
South Africa	1364	488	700	57
Angola	56	2	700	22
Cameroon	54	5	230	44
Pakistan	1439	47	230	1332
Zimbabwe	750	252	700	254
Channel Islands	121	94	230	24
Bangladesh	246	4	230	225
Tanzania	425	84	1000	25

Sierra Leone	230	5	800	199
Portugal	320	18	230	275
Uganda	616	73	700	126
Poland	637	2	230	619
China	413	26	230	371
Cyprus	328	114	230	160
Italy	821	16	230	789
Bulgaria	71	2	230	65
Israel	87	10	300	56
France	856	103	230	690
Malta	365	170	230	135
Myanmar (Burma)	124	23	400	31
Philippines	333	8	230	310
Iraq	337	17	400	298
Finland	158	2	230	154
Libya	110	42	800	35
Norway	134	19	230	104
Russia	159	0	300	124
Nepal	161	0	230	119
Denmark	231	137	230	86
Spain	355	2	230	339
Serbia/Montenegro	56	0	230	55
Algeria	92	1	600	68
Sicily	3	0	230	2
Afghanistan	112	1	500	103

Gibraltar	77	34	230	32
Lebanon	77	13	500	50
Sudan	117	21	800	61
Morocco	93	3	800	66
Greece	131	5	230	116
Austria	196	20	230	171
Ukraine	62	1	400	54
Congo	167	11	800	146
Lithuania	72	0	230	66
Vietnam	74	0	230	69
Romania	68	0	230	61
Malawi	111	25	800	10
Gambia	42	0	800	38
Equatorial Guinea	4	0	800	2
Thailand	104	6	300	87
Indonesia	62	7	400	35
Central African Republic	42	6	800	14
Sweden	216	9	230	196
Jordan	15	1	600	9
Croatia	46	0	230	45
Ethiopia	81	8	800	57
Somalia	119	2	800	78
Zambia	246	101	800	56
Tunisia	27	2	900	14
Rwanda	25	1	1000	19

Yemen	108	38	2000	56
Burundi	26	3	1000	19
Eritrea	54	4	800	44
Syria	31	0	800	27
Luxembourg	7	0	230	6
Cambodia	8	1	1500	7
Macau (Macao)	8	0	500	6
Seychelles	46	1	2000	23
Liberia	22	3	800	15
Kuwait	31	16	800	5
Taiwan	26	1	230	23
Niger	5	0	800	2
Georgia	4	0	400	3
Iceland	19	4	230	15
Macedonia	18	0	230	17
Ivory Coast	32	0	800	29
Mongolia	6	0	400	6
Kazakhstan	13	0	4000	13
Brunei	19	4	400	8
Latvia	53	0	300	52
Bosnia and Herzegovina	41	0	230	41
Guinea	11	0	800	6
Slovenia	11	0	230	11
Azerbaijan	6	0	500	5
Slovakia	35	0	230	30

Kyrgyzstan	4	0	2000	3
Estonia	15	0	230	14
Senegal	12	0	800	7
South Korea	26	0	230	24
Togo	10	0	230	9
Armenia	5	0	600	5
Albania	12	0	230	12
British Indian Ocean Territory	10	2	500	4
Kurdistan	2	0	600	2
North Korea	6	0	230	5
Laos	3	0	500	3
Lesotho	2	0	400	2
Serbia	2	0	230	2
Republic of Kosovo	6	0	230	6
Botswana	7	1	800	4
Uzbekistan	3	0	5000	3
Kashmir	3	0	300	3
Turkmenistan	1	0	230	1
Belarus	6	0	230	4
Tibet	1	0	230	1
Crete	1	0	230	1
Moldova	1	0	230	1
Tajikistan	1	0	230	1

Table S3g.1. Parameters for selection of individuals in the UKB born in a given country of a 'typical ancestral background'. This shows the initial number of individuals coded as being born in a country; the number removed because they clustered with the white British; the eps value for selecting the main cluster; and the number of individuals in

the final selected cluster. Countries with fewer than 3 individuals in the final cluster, or no obvious main cluster, were discarded. The eps value is dependent on the genetic diversity of the population being selected, and was chosen manually and visually checked. In most but not all cases the largest cluster was chosen.

References

1. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research* **12**, 2825–2830 (2011).
2. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

3h) Painting the UK BioBank

Will Barrie[1] and Dan Lawson[2]

[1]Zoology Department, University of Cambridge, UK.

[2]School of Mathematics and Integrative Epidemiology Unit, University of Bristol, UK.

Introduction

Here, we develop new methods to use Chromopainter [1] on a biobank scale to 'paint' modern genomes from the UK Biobank (UKB) using ancient genomes, grouped into reference populations, as donors. Painting was done following the pipeline of Margaryan et al. [2] based on GLOBETROTTER [3], and admixture proportions were estimated using Non-Negative Least squares. These results (technically most recent coalescences) are used as a proxy for ancestry. We store both genome-wide and local ancestry (i.e. per variant per individual) results.

Methods

Painting pipeline introduction

The process of painting consists of forming a reference/donor panel consisting of ancient individuals of as pure ancestry as possible, having undergone QC and clustering using

fineSTRUCTURE. The target/recipient panel and reference/donor panel are then filtered for variants, merged, and the target panel is painted using the reference panel as donors.

Reference/donor panel formation

We used imputed best guess haplotypes filtered for imputation information score (FORMAT/INFO) above 0.5. Samples were selected based on IBD-sharing, visual PCA inspection, and fineSTRUCTURE analysis (unsupervised clustering based on the coancestry matrix output of ChromoPainter; Figure S3h.1); low coverage, contaminated, and related individuals were excluded. The aim was to group samples into as pure 'source' populations as possible, while maintaining reasonable numbers in each population. We do not expect our filters for white British/non-british individuals to be perfect; furthermore, modelling modern Eurasians as a mixture of hunter-gatherer/Steppe/farmer is overly simplistic. Therefore, we also include ancient African and EastAsian reference populations to account for possible 'non-European' ancestry.

Ultimately, 318 individuals split into ten reference populations were used (Figure S3h.2, Figure S3h.3, Table S3h.1): western hunter-gatherer (WHG), eastern hunter-gatherer (EHG), Caucasus hunter-gatherer (CHG), FarmerAnatolian, FarmerEarly, FarmerMiddle, FarmerLate, Yamnaya, African and EastAsian. Populations are characterised by preferentially copying from individuals within the population, as well as being biologically and historically meaningful. This dataset is henceforth called the "present aDNA dataset".

The farmers are split into four separate populations due to their differing behaviour as donors (columns) in the fineSTRUCTURE analysis (Figure S3h.1). There is a cline in their degree of WHG admixture that roughly correlates with age, while some samples also show Steppe admixture. Given the nature of the splits, the differences between these groups should be interpreted with caution, and for most downstream analysis these groups are merged into a 'Farmer' ancestry.

Target/recipient panel formation

We used white British individuals from the UKB as reported in Bycroft et al.[4]; these are individuals who self-reported as white British and have British-like ancestry according to PCA. We also used individuals from the UKB of a typical ancestral background selected by country of origin (Supplementary Note S3g). We used phased haplotype data, downloaded

from https://www.ukbiobank.ac.uk. This totalled 408,884 white British individuals, and 24,511 non-British individuals. This dataset is henceforth called the “UKB dataset”.

SNP selection and merging of the panels

Due to computational considerations, the number of SNPs used in the painting was limited to those in the UKB Axiom Array; these SNPs were chosen to capture genome-wide variation, rare and coding variants, and variants relevant to specific phenotypes or regions of interest [4]. The present aDNA dataset and UKB datasets were merged and filtered for these variants using QCTOOL v2 (https://www.well.ox.ac.uk/~gav/qctool_v2/), and then filtered to exclude variants with a minor allele frequency below 1% using bcftools (http://samtools.github.io/bcftools/), leaving a total of 549,323 SNPs across chromosomes 1-22.

Painting process

ChromoPainter [1] uses an approach premised on the observation that markers on the same chromosome are inherited together unless separated by recombination; at the population level, this results in linkage disequilibrium (LD) between close markers that reflect a shared history of descent. The haplotype-based algorithm of ChromoPainter aims to harness this information, detecting shared haplotypes to reconstruct phased recipient genomes as chunks ‘copied’ from donors.

Considering the genealogy of a single locus, we can identify one or more closest relatives to that locus, henceforth called ‘nearest neighbours’; if viewed as a genealogy, these are the other leaves of the tree underneath the first coalescence. Therefore at each locus of each haplotype, there exists one or more nearest neighbours. ChromoPainter aims to identify these using an approximate method based on that introduced by Li and Stephens [5]: the Hidden Markov Model (HMM), which explicitly reconstructs the haplotype of a recipient/target individual as a series of chunks of genetic material donated by the other donor/reference individuals, using information on the types of the recipient and potential donor at each SNP. This approach is probabilistic, calculating the expectations of which haplotype acts as donor to a recipient as a function of position over an infinite number of paintings [1]. Although ChromoPainter was originally intended to use this information, in the form of a ‘co-ancestry matrix’, to ascertain fine-scale population structure and clustering (in the fineSTRUCTURE software package), the software can be used with pre-defined donor and recipient populations.

If the donor panel is formed of ancient individuals and the recipient individual is modern, the nearest neighbour should reflect some history of that locus. The ability of chromosome painting to accurately infer ancestry is expected to depend on the diversity between donor populations: more genetically similar populations and the algorithm will find it difficult to correctly identify the nearest neighbour(s). There is also the issue of 'masking' whereby haplotypes from older populations would have travelled through more recent populations before arriving in the modern population; this causes a genome-wide bias towards the more recent ancient populations, the effects of which are discussed below. Here, we use nearest neighbour as a proxy for local ancestry - i.e. which population that haplotype came from (which may not be a single unique population from our panel).

Chromosome painting cannot include the target of painting. Therefore, painting was done (following the pipeline of Margaryan et al. [2] based on GLOBETROTTER [3] by leaving out one individual at random (chosen independently for each chromosome) from each other donor population for all donor individuals. Target individuals from the UK Biobank were painted by similarly removing one individual at random from all donor populations. This ensures that individuals from the reference and UK Biobank are exchangeable.

Once we had a well-chosen set of ancient populations from the present aDNA panel, each individual was repainted twice leaving out themselves as a possible donor: first to learn the painting parameters Ne and μ, and then to learn a genome-wide individual-specific donor-prior. For each of the reference populations, the average amount of genome received from each donor individual was learnt. We then painted the modern individuals in the UKB panel using the reference populations and the learnt parameters and priors.

The probability that each recipient copied each donor population at every SNP was recorded. The genome-wide information for each recipient was also stored, in the form of (i) chunkcounts, the number of chunks copied from each donor population and (ii) chunk lengths, the sum of the lengths of the chunks copied from each population, weighted by their copying probability. Admixture proportions were then estimated using Non-Negative Least Squares (NNLS).

Painting at biobank scale

We used custom scripts to speed up this process (specifically reading from large phase files), to enable running for large numbers of recipients in parallel across multiple nodes, and

to store the local copying probabilities in a memory-efficient format in real time (all scripts available at https://github.com/will-camb/Nero/tree/master/scripts/cp_panel_scripts). The total CPU time for painting the UKB panel was approximately 550,000 CPU hours.

Results

Ancestry-PCs relationship

PCA is a dimensionality reduction technique that can be applied to genetic data, the results of which are useful as a means to visualise variation between individuals/groups, and are expected to reflect historical events that cause differences in ancestry due to drift, admixture etc. It is well established that PC1 vs PC2 vs PC3 generally separate African, European and East Asian populations. We ran multivariate linear regressions using ancestry components to predict UKB PCs [4]. Previous work has shown that the main UKB PCs that reflect British population structure are PCs 5 and 9, describing variation between English, Scottish and Welsh ancestry, and PCs 11 and 14 which further separate structure within Wales and England [6].

We found significant correlations between ancestry components and PC4 (R-squared=0.553) and PC5 (R-squared=0.376), as well as PC1 (R-squared=0.165) and PC7 (R-squared=0.130). We found that the high PC4 correlation with ancestry component was largely driven by a Steppe (Yamnaya/EHG) vs Farmer divide, both within Britain and internationally: high PC4 values are associated with high Steppe/low Farmer ancestry, while low PC4 values are associated with low Steppe/high Farmer ancestry.

Ancestry-geographic variation

Within the British Isles, all individuals were painted with similar proportions from each reference population, as expected when measuring coalescence tracts rather than direct admixture tracts and after a long time since admixture events; but, the differences in copying proportions showed significant geographic heterogeneity. We ran multivariate linear regressions, using longitude and latitude of place of birth ("Place of birth in UK – east co-ordinate" and "Place of birth in UK – north co-ordinate") to predict NNLS ancestry fractions. We found significant correlations for Yamnaya ancestry (R-squared=0.081), Farmer ancestry (R-squared=0.066), CHG ancestry (R-squared=0.015), WHG ancestry (R-squared=0.007), African ancestry (R-squared=0.011) and EHG ancestry (R-squared=0.01, longitude only). To visualise this, we assigned individuals to a county based on their UKB place of birth data,

and plotted the average admixture proportion per county for each ancestry, binned in ten equal interval quantiles using ArcGIS Online (www.arcgis.com; Figure 5, main text).

We found that Neolithic farmer ancestry was highest in southern and eastern England and lower in populations in Scotland, Wales and Cornwall. We found the opposite pattern in Yamnaya ancestry, representing the Steppe component, which has previously been shown to be higher in Scotland but not Wales [7]; we found this was highest in the Outer Hebrides. This Farmer/Yamnaya dichotomy broadly reflects an Anglo-Saxon/'Celtic' distribution. We are unable to date when these subtle population structures arose, but note that the Neolithic Anatolian-related farmer ancestry is already present in the British and Roman Iron Age but lower in Saxon individuals (Extended Data Figure 3, main text), meaning these patterns cannot be explained just by Saxon-related ancestry. They are likely a result of pre-Roman migration between 1000 and 875 BC which resulted in a slight increase in Early Farmer ancestry in England and Wales but not Scotland [8], although we note our results show a marked difference between Wales/Cornwall and England too. We also found higher levels of WHG-related ancestry in central and Northern England.

Looking at a continent-wide level, the hunter-gatherer ancestries display distinct structure in modern populations (Figure 5, main text). WHG-related ancestry is highest in present-day individuals from the Baltic States, Belarus, Poland and Russia; EHG-related ancestry is highest in Mongolia, Finland, Estonia and Central Asia; and CHG-related ancestry is maximised in countries east of the Caucasus, in Pakistan, India, Afghanistan and Iran, in accordance with previous results [9]. The CHG-related ancestry likely picks up both Caucasus hunter-gatherer and Iranian Neolithic signals, explaining the relatively high levels in south Asia [10]. Consistent with expectations [11,12], Neolithic Anatolian-related farmer ancestry is concentrated around the Mediterranean basin, with high levels in southern Europe, the Near East and North Africa, including the Horn of Africa but less in Northern Europe. A contrasting pattern was observed in Yamnaya-related ancestry decreasing from high levels in northern Europe, peaking in Ireland, Iceland, Norway and Sweden, but decreasing further south where Neolithic farmer ancestry still dominates. There is also evidence for its spread into southern Asia. These results provide a new level of detail on the modern distribution of ancient ancestries.

To better understand how countries varied in their ancestry proportions, we ran Scikit-learn's PCA [13] on the average admixture proportions per country. We then ran a hierarchical clustering algorithm on the first 4 PCs (explained variance=0.244), and built a dendrogram

(Figure S3h.4) [14]. For further analysis, we excluded countries in clusters dominated by African or East Asian ancestry, leaving 80 countries.

We used Sklearn's StandardScaler utility class to standardise each feature (zero mean and unit variance), then ran a PCA using the standardised average admixture proportions for each of the 80 remaining countries (https://github.com/erdogant/distfit), and plotted a biplot (PC1 vs PC2 with loadings for each feature plotted), which shows the correlations between ancestries (Figure S3h.5). This gives a more visual representation of how countries group by ancestry, and broadly reflects actual geography.

Discussion

The large ancient DNA panel established here combined with the UKB allows us to trace for the first time the fine-scale distribution of Mesolithic/Neolithic/Bronze Age ancestry components in modern British individuals, using DNA directly from ancient individuals. It also demonstrates the ancestry differences within an 'ethnic group' (white British) traditionally regarded as being relatively homogenous, highlighting the need for care over ancestry considerations when using resources like the UK Biobank.

Figures

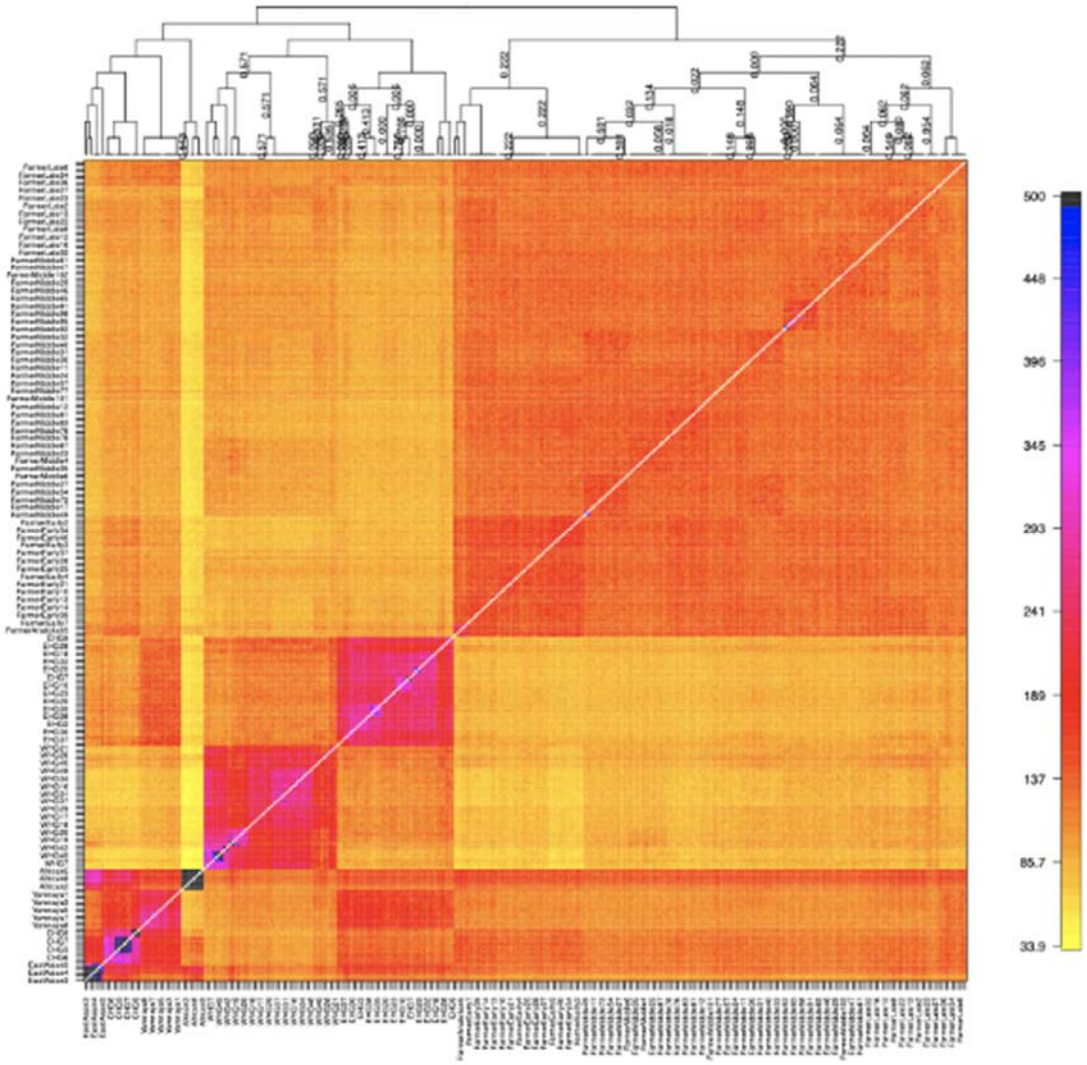


Figure S3h.1. Co-ancestry heatmap of selected ancient samples. The output of fineSTRUCTURE analysis of the ancient reference panel, showing copying proportions between ancient populations (columns=donors, rows=recipients). There is a cline in Hunter-Gatherer admixture in the Farmers, roughly correlating with age. For most downstream analyses, the Farmer populations were merged.

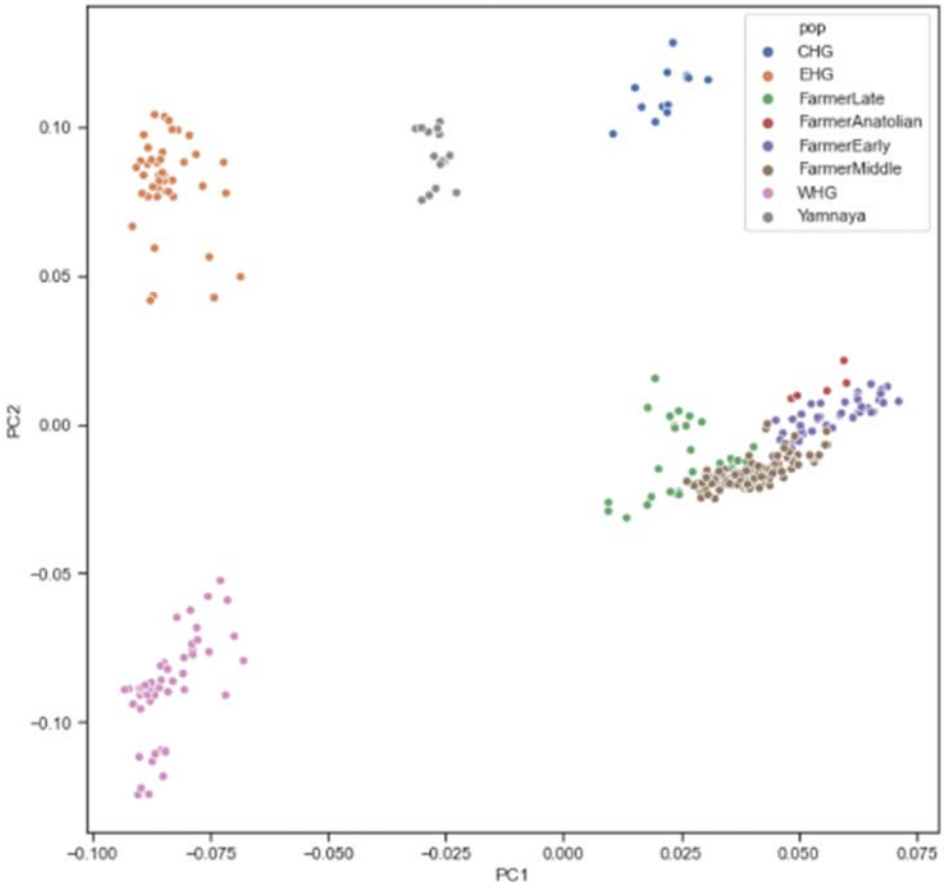


Figure S3h.2. PCA of ancient reference samples, coloured by assigned population. PC1 vs PC2 of a PCA of the ancient western Eurasian samples (excluding African and EastAsian), coloured by their assigned population used in the painting. As can be seen, populations are fairly distinct, with intermediate admixed individuals having been excluded. Some Farmers are admixed with Steppe and Hunter-Gatherer populations to differing degrees, but particularly among later individuals.

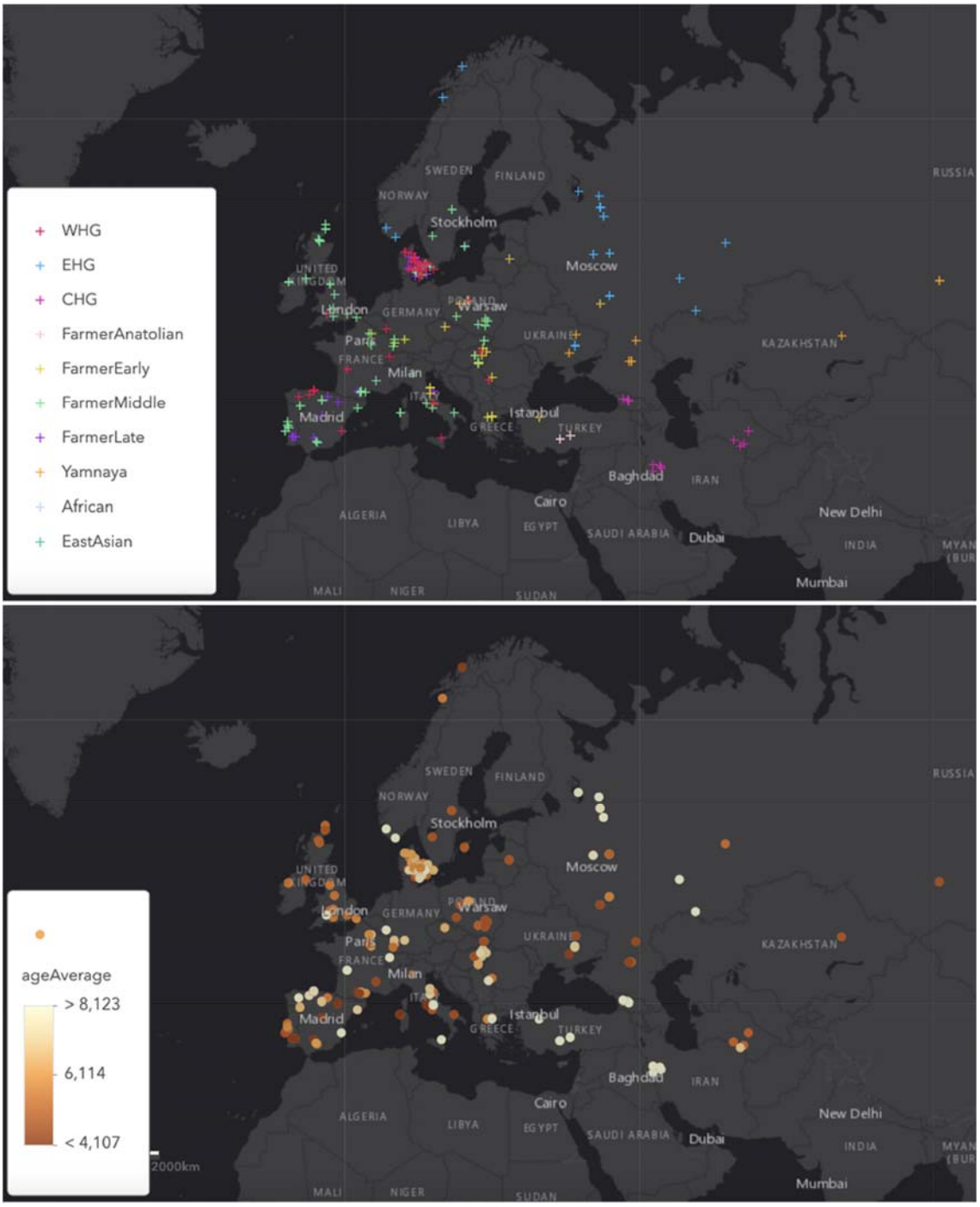


Figure S3h.3. Maps of ancient sample locations coloured by assigned reference population (above) and age (below). Not showing African and East Asian samples.

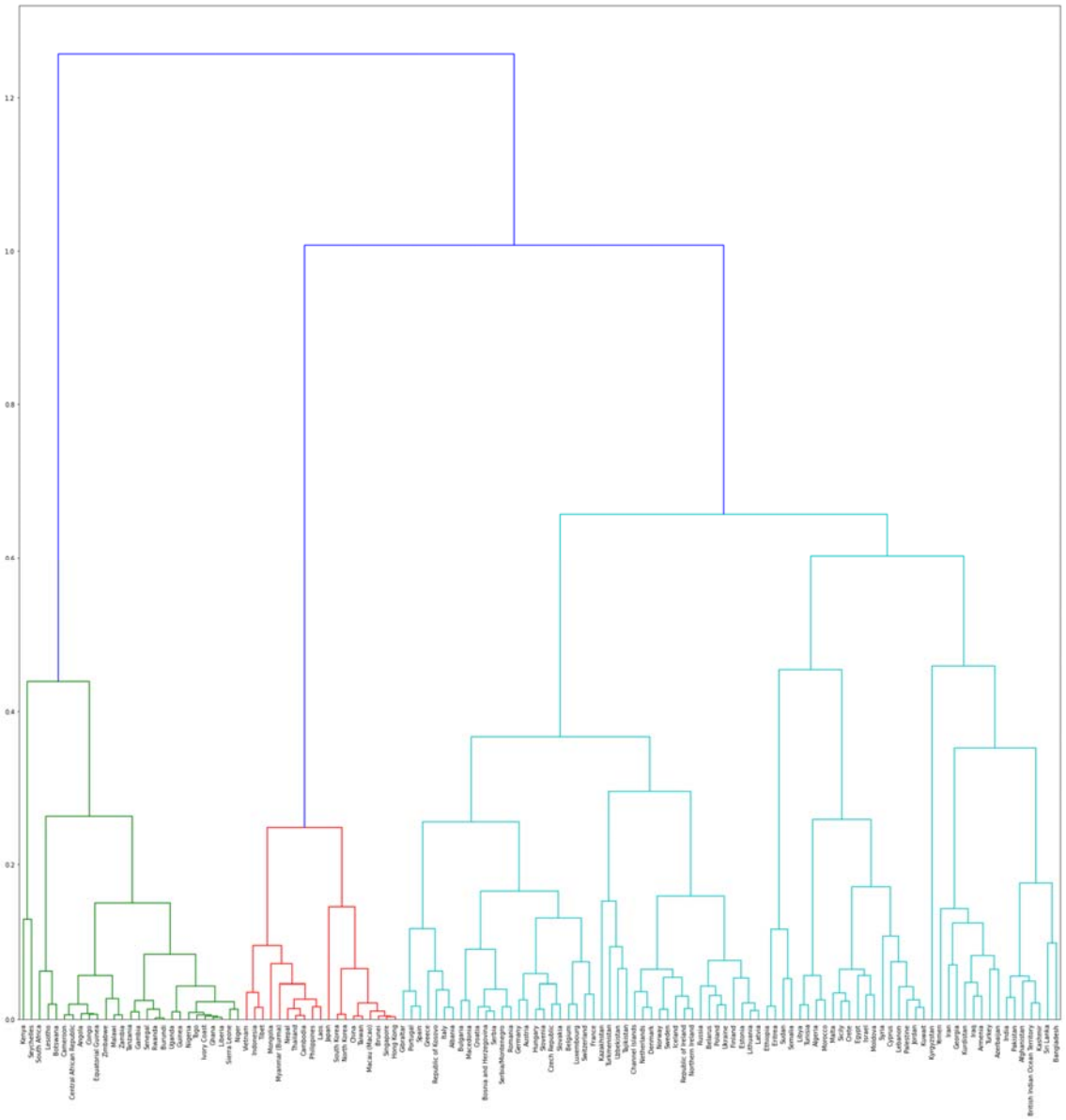


Figure S3h.4. Dendrogram based on hierarchical clustering of first 4 PCs of average admixture proportions per country. For further analysis, countries in clusters dominated by African (green) or East Asian (red) ancestry were dropped.

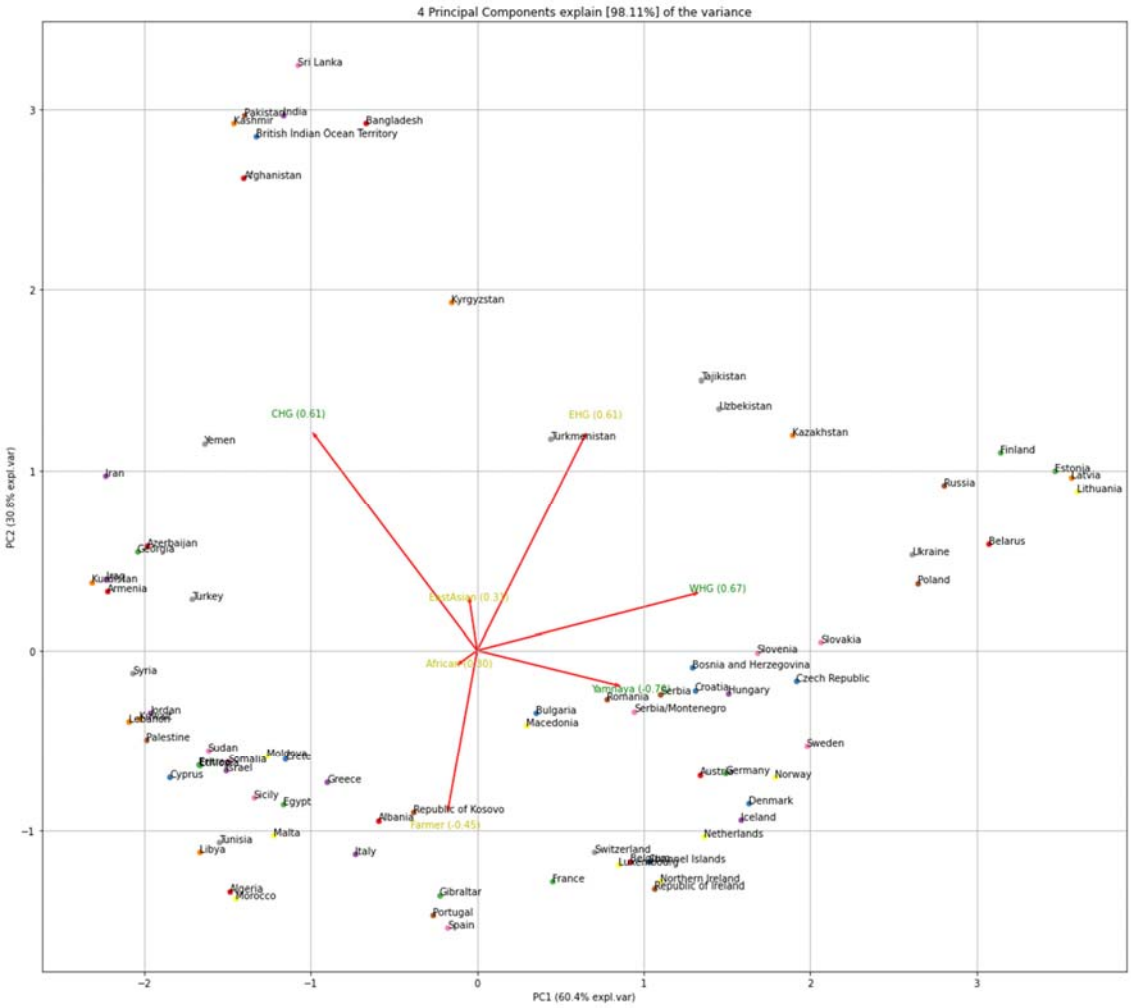


Figure S3h.5. PCA biplot of standardised average NNLS admixture proportion per country, based on 80 countries in Europe, West/Southern Asia, the Middle East and North Africa.

referencePopulation	sampleId	country	groupLabel	latitude	longitude	ageAverage	coverage	sex	clusterIBDFine
African	mfo01	South Africa	SouthAfrica_IronAge	-28.73	30.81	378.0	7.061854822000000	XX	6.1_SouthAfrica_400BP
African	bab01	South Africa	SouthAfrica_Neolithic	-29.54	31.22	2040.5	1.29865429	XY	6.2_SouthAfrica_2000BP_1000BP
African	ela01	South Africa	SouthAfrica_IronAge	-28.92	29.13	493.0	13.46552376	XX	6.1_SouthAfrica_400BP
African	I10871	Cameroon	Cameroon_Neolithic	5.86	10.08	7885.0	15.21262736	XY	6.3_Cameroon_8000BP_3000BP
African	I10873	Cameroon	Cameroon_Neolithic	5.86	10.08	3065.0	3.276739876	XY	6.3_Cameroon_8000BP_3000BP

African	I9133	South Africa	SouthAfrica_Neolithic	-31.98	18.52	1970.0	2.0755013510000000	XY	6.2_SouthAfrica_2000BP_1000BP
African	baa01	South Africa	SouthAfrica_Neolithic	-29.54	31.22	1908.5	13.50021278	XY	6.2_SouthAfrica_2000BP_1000BP
African	new01	South Africa	SouthAfrica_IronAge	-27.76	29.92	417.5	10.89613659	XX	6.1_SouthAfrica_400BP
CHG	WC1	Iran	Iran_Neolithic	34.61	47.11	9218.5	10.4300754	XY	2.1_Iran_10000BP_8500BP
CHG	AH4	Iran	Iran_Neolithic	34.19	48.37	9929.5	0.8674508960000000	XX	2.1_Iran_10000BP_8500BP
CHG	AH2	Iran	Iran_Neolithic	34.19	48.37	9930.5	0.6492785520000000	XY	2.1_Iran_10000BP_8500BP
CHG	AH1	Iran	Iran_Neolithic	34.19	48.37	9900.0	1.161365185	XX	2.1_Iran_10000BP_8500BP
CHG	DA380	Turkmenistan	Turkmenistan_Neolithic_Namazga	37.6	59.33	5180.5	0.495449798	XX	2.1_Turkmenistan_7000BP_5000BP
CHG	DA381	Turkmenistan	Turkmenistan_Neolithic_Namazga	37.19	61.03	5181.0	0.83822953	XY	2.1_Turkmenistan_7000BP_5000BP
CHG	NEO816	Iran	Iran_Neolithic	33.76	47.1	8700.0	0.940243388	XY	2.1_Iran_10000BP_8500BP
CHG	NEO281	Georgia	Georgia_Mesolithic	42.22	43.32	9724.0	3.607878115	XY	2.1_Caucasus_13000BP_10000BP
CHG	KK1	Georgia	Georgia_Mesolithic	42.28	43.28	9720.0	11.83484526	XY	2.1_Caucasus_13000BP_10000BP
CHG	SATP	Georgia	Georgia_Mesolithic	42.38	42.59	13255.0	1.18417508	XY	2.1_Caucasus_13000BP_10000BP
CHG	GD13a	Iran	Iran_Neolithic	34.45	48.12	9846.0	1.41728065	XX	2.1_Iran_10000BP_8500BP
CHG	DA383	Turkmenistan	Turkmenistan_Neolithic_Namazga	38.72	61.69	5150.0	0.7751095790000000	XX	2.1_Turkmenistan_7000BP_5000BP
CHG	NEO310	Turkmenistan	Turkmenistan_Neolithic	36.85	60.42	7150.0	1.278435989	XY	2.1_Turkmenistan_7000BP_5000BP
EHG	Karelia	Russia	Russia_Mesolithic	61.65	35.65	8279.5	1.692885466	XY	4.1_RussiaNW_11000BP_8000BP
EHG	NEO166	Russia	Russia_Neolithic	53.0	40.4	5668.0	2.1533629430000000	XY	4.1_DonRiver_5800BP_5300BP
EHG	NEO167	Russia	Russia_Neolithic	53.0	40.4	5657.0	0.588921338	XX	4.1_DonRiver_5800BP_5300BP
EHG	NEO170	Russia	Russia_Neolithic	53.0	40.4	5562.0	0.41889687700000000	XX	4.1_DonRiver_5800BP_5300BP
EHG	NEO171	Russia	Russia_Neolithic	53.0	40.4	5835.0	0.818232704	XX	4.1_DonRiver_5800BP_5300BP
EHG	VK531	Norway	Norway_Neolithic	69.47	18	4350.0	1.4595637520000000	XY	4.1_Norway_9300BP_4300BP
EHG	NEO173	Russia	Russia_Neolithic_Sredny	52.28	38.96	6345.0	0.689663925	XX	4.1_RussiaNW_7000BP_5000BP

EHG	NEO100	Russia	Russia_Mesolithic	51.57	53.68	9929.0	0.10793872200000000	XY	4.1_RussiaNW_11000BP_8000BP
EHG	NEO88	Russia	Russia_Mesolithic	56.67	38.02	7871.0	2.6217337880000000	XY	4.1_RussiaNW_11000BP_8000BP
EHG	H22	Norway	Norway_Mesolithic	58.83	6.33	9363.5	0.719978067	XX	4.1_Norway_9300BP_4300BP
EHG	stg001	Norway	Norway_Neolithic	67.76	14.85	5857.0	1.291071015	XY	4.1_Norway_9300BP_4300BP
EHG	NEO87	Russia	Russia_Mesolithic	56.67	38.02	8259.0	0.183898205	XX	4.1_RussiaNW_11000BP_8000BP
EHG	NEO17	Norway	Norway_Mesolithic	58.06	7.74	9146.0	1.0991194990000000	XY	4.1_Norway_9300BP_4300BP
EHG	Ukraine_N1	Ukraine	Ukraine_Neolithic	48.13	35.08	7250.0	0.167926522	XY	4.1_Ukraine_10000BP_4000BP
EHG	Latvia_MN2	Latvia	Latvia_Neolithic_CCC	56.28	25.13	5965.0	1.147611647	XX	4.1_RussiaNW_7000BP_5000BP
EHG	NEO160	Russia	Russia_Neolithic	53.0	40.4	5269.0	1.326576858	XX	4.1_DonRiver_5800BP_5300BP
EHG	NEO178	Russia	Russia_Neolithic	56.78	40.45	5322.0	0.3243332060000000	XY	4.1_RussiaNW_7000BP_5000BP
EHG	NEO163	Russia	Russia_Neolithic	53.0	40.4	5603.0	0.226384639	XY	4.1_DonRiver_5800BP_5300BP
EHG	NEO180	Russia	Russia_Neolithic	56.78	40.45	5947.0	0.3851877920000000	XY	4.1_RussiaNW_7000BP_5000BP
EHG	NEO687	Russia	Russia_Neolithic	57.58	58.2	5446.0	0.4408981110000000	XY	4.1_RussiaNW_7000BP_5000BP
EHG	NEO560	Russia	Russia_Neolithic	60.41	38.93	7919.0	1.548541653	XY	4.1_RussiaNW_11000BP_8000BP
EHG	NEO559	Russia	Russia_Neolithic	60.41	38.93	8268.0	1.228507174	XY	4.1_RussiaNW_11000BP_8000BP
EHG	NEO557	Russia	Russia_Neolithic	60.41	38.93	7917.0	0.7776540910000000	XY	4.1_RussiaNW_11000BP_8000BP
EHG	NEO556	Russia	Russia_Neolithic	60.41	38.93	7036.0	1.122201651	XX	4.1_RussiaNW_7000BP_5000BP
EHG	NEO555	Russia	Russia_Neolithic	60.41	38.93	8280.0	2.097005144	XY	4.1_RussiaNW_11000BP_8000BP
EHG	NEO539	Russia	Russia_Mesolithic	59.7	39.5	10060.0	0.291758629	XX	4.1_RussiaNW_11000BP_8000BP
EHG	NEO179	Russia	Russia_Neolithic	56.78	40.45	5467.0	0.6393595470000000	XX	4.1_RussiaNW_7000BP_5000BP
EHG	NEO536	Russia	Russia_Mesolithic	59.7	39.5	9541.0	0.190177354	XY	4.1_RussiaNW_11000BP_8000BP
EHG	NEO501	Ukraine	Ukraine_Mesolithic	48.2	35.22	10623.0	0.125455369	XY	4.1_Ukraine_10000BP_4000BP
EHG	NEO202	Russia	Russia_Mesolithic	61.27	38.91	10884.0	2.217735963	XY	4.1_RussiaNW_11000BP_8000BP
EHG	NEO197	Russia	Russia_Neolithic	56.78	40.45	5245.0	0.610890023	XY	4.1_RussiaNW_7000BP_5000BP

EHG	NEO195	Russia	Russia_Neolithic	56.78	40.45	5749.0	0.59759596	XY	4.1_RussiaNW_7000BP_5000BP
EHG	NEO174	Russia	Russia_Neolithic_Sredny	52.28	38.96	5306.0	1.152059756 0000000	XX	4.1_DonRiver_5800BP_5300BP
EHG	NEO193	Russia	Russia_Neolithic	56.78	40.45	5453.0	0.232823778	XY	4.1_RussiaNW_7000BP_5000BP
EHG	NEO184	Russia	Russia_Neolithic	56.78	40.44	5458.0	0.695336472	XX	4.1_RussiaNW_7000BP_5000BP
EHG	NEO185	Russia	Russia_Neolithic	56.78	40.45	7034.0	1.213238119 0000000	XY	4.1_RussiaNW_7000BP_5000BP
EHG	NEO194	Russia	Russia_Neolithic	56.78	40.45	5575.0	1.626304895	XY	4.1_RussiaNW_7000BP_5000BP
EHG	NEO187	Russia	Russia_Neolithic	56.78	40.45	4940.0	0.16188839	XX	4.1_RussiaNW_7000BP_5000BP
EHG	Sidelkino	Russia	Russia_Mesolithic	54.53	51.11	11258.5	2.996735907 0000000	XX	4.1_RussiaNW_11000BP_8000BP
EHG	NEO186	Russia	Russia_Neolithic	56.78	40.45	6922.0	0.368424063 00000000	XY	4.1_RussiaNW_7000BP_5000BP
EHG	NEO192	Russia	Russia_Neolithic	56.78	40.45	6841.0	0.269995424	XX	4.1_RussiaNW_7000BP_5000BP
EHG	NEO189	Russia	Russia_Neolithic	56.78	40.45	5648.0	0.614470507	XY	4.1_RussiaNW_7000BP_5000BP
EastAsian	IK002	Japan	Japan_Jomon	34.65	137.14	2569.0	1.89109494	XX	3.1_Japan_3700BP_2600BP
EastAsian	DA45	Mongolia	Mongolia_IronAge_XiongNu	42.53	105.18	2095.0	9.039659899	XY	3.1_SEAsia_4000BP_150BP
EastAsian	DA43	Mongolia	Mongolia_IronAge_XiongNu	42.53	105.18	2095.0	1.677926438	XY	3.1_SEAsia_4000BP_150BP
EastAsian	DA39	Mongolia	Mongolia_IronAge_XiongNu	48.02	101.35	1948.0	2.087185526	XY	3.1_SteppeCE_2000BP_700BP
EastAsian	DA38	Mongolia	Mongolia_IronAge_XiongNu	49.27	101.72	2124.5	2.85446773	XX	3.1_SteppeC_TianShan_2700BP_800BP
EastAsian	Funadomari_23	Japan	Japan_Jomon	45.38	141.04	3755.0	39.44124624	XX	3.1_Japan_3700BP_2600BP
EastAsian	Funadomari_5	Japan	Japan_Jomon	45.38	141.04	3755.0	3.821436778 0000000	XY	3.1_Japan_3700BP_2600BP
FarmerAnatolian	Bon001	Turkey	Anatolia_Neolithic_Aceramic	37.75	32.86	10032.0	0.163100502 00000000	XY	2.3_Anatolia_10000BP_8000BP
FarmerAnatolian	Bon002	Turkey	Anatolia_Neolithic_Aceramic	37.75	32.86	10078.0	6.692061812	XX	2.3_Anatolia_10000BP_8000BP
FarmerAnatolian	Bon004	Turkey	Anatolia_Neolithic_Aceramic	37.75	32.86	10076.0	0.242113723	XY	2.3_Anatolia_10000BP_8000BP
FarmerAnatolian	Tep002	Turkey	Anatolia_Neolithic	38.17	34.49	8585.0	0.707235454 0000000	XX	2.3_Anatolia_10000BP_8000BP
FarmerAnatolian	Tep004	Turkey	Anatolia_Neolithic	38.17	34.49	8295.0	0.467990841	XX	2.3_Anatolia_10000BP_8000BP
FarmerEarly	R3	Italy	Italy_Neolithic	41.96	13.54	7729.5	4.059641042	XX	2.3_EuropeS_8000BP_6000BP

FarmerEarly	R17	Italy	Italy_Neolithic	43.72	13.03	7223.5	0.56582164	XY	2.3_EuropeS_8000BP_6000BP
FarmerEarly	R19	Italy	Italy_Neolithic	43.72	13.03	7233.0	0.52086802	XY	2.3_EuropeS_8000BP_6000BP
FarmerEarly	R18	Italy	Italy_Neolithic	43.72	13.03	7298.5	0.6271842	XX	2.3_EuropeS_8000BP_6000BP
FarmerEarly	R16	Italy	Italy_Neolithic	43.72	13.03	7207.5	0.565514455	XX	2.3_EuropeS_8000BP_6000BP
FarmerEarly	R10	Italy	Italy_Neolithic	41.96	13.54	7629.0	1.321819486 0000000	XX	2.3_EuropeS_8000BP_6000BP
FarmerEarly	R2	Italy	Italy_Neolithic	41.96	13.54	7984.0	3.704063854 0000000	XX	2.3_EuropeS_8000BP_6000BP
FarmerEarly	R9	Italy	Italy_Neolithic	41.96	13.54	7496.0	4.04251971	XY	2.3_EuropeS_8000BP_6000BP
FarmerEarly	MDV248	France	France_Neolithic_LBK	49.42	4.01	7015.5	0.121146489 00000000	XY	2.3_Europe_8500BP_5500BP
FarmerEarly	ROS45	France	France_Neolithic_Grossgartach	48.5	7.47	6642.5	0.269292799 00000000	XY	2.3_Europe_8500BP_5500BP
FarmerEarly	ROS78	France	France_Neolithic_Grossgartach	48.5	7.47	6550.0	0.435416491	XY	2.3_Europe_8500BP_5500BP
FarmerEarly	Sch72-15	France	France_Neolithic_LBK	48.76	7.6	7036.5	0.235222248	XY	2.3_Europe_8500BP_5500BP
FarmerEarly	Schw432	France	France_Neolithic_LBK	48.76	7.6	7100.0	0.148141996	XX	2.3_Europe_8500BP_5500BP
FarmerEarly	NEO137	Hungary	Hungary_Neolithic_Koros	46.42	20.33	7591.0	0.198402858 00000000	XX	2.3_Europe_8500BP_5500BP
FarmerEarly	NEO140	Hungary	Hungary_Neolithic_Tisza	46.37	20.42	6718.0	0.145142912 00000000	XY	2.3_Europe_8500BP_5500BP
FarmerEarly	NEO145	Hungary	Hungary_Neolithic_Tisza	46.37	20.42	6744.0	0.218956151 00000000	XX	2.3_Europe_8500BP_5500BP
FarmerEarly	NEO147	Hungary	Hungary_Neolithic_Tisza	46.37	20.42	6724.0	0.284571059 00000000	XX	2.3_Europe_8500BP_5500BP
FarmerEarly	NEO695	Italy	Italy_Neolithic	43.08	13.06	7299.0	0.431411103 00000000	XX	2.3_EuropeS_8000BP_6000BP
FarmerEarly	NEO674	Romania	Romania_Neolithic	44.9	22.43	5570.0	0.237000547	XY	2.3_Europe_8500BP_5500BP
FarmerEarly	R8	Italy	Italy_Neolithic	41.96	13.54	7723.5	0.53180897	XX	2.3_EuropeS_8000BP_6000BP
FarmerEarly	kol6	CzechRepublic	Czech_Neolithic_Megalithic	50.03	15.2	6690.0	1.543500893	XX	2.3_EuropeCE_7000BP_5500BP
FarmerEarly	NEO655	Serbia	Serbia_Mesolithic	44.54	22.04	8668.0	0.225492674 00000000	XX	2.3_Europe_8500BP_5500BP
FarmerEarly	PL_N36	Poland	Poland_Neolithic_BKG	50.67	21.38	6250.0	1.669980973 0000000	XX	2.3_EuropeCE_7000BP_5500BP
FarmerEarly	NE5	Hungary	Hungary_Neolithic_ALP	47.17	20.83	7050.0	0.784933195	XY	2.3_EuropeCE_7000BP_5500BP
FarmerEarly	PL_N31	Poland	Poland_Neolithic_BKG	52.61	18.8	6137.5	3.003824151 0000000	XX	2.3_EuropeCE_7000BP_5500BP

FarmerEarly	NE1	Hungary	Hungary_Neolithic_ALP	47.85	21.15	7138.5	18.42090683	XX	2.3_EuropeCE_7000BP_5500BP
FarmerEarly	Stuttgart	Germany	Germany_Neolithic	48.78	9.18	7140.0	16.19254244	XX	2.3_EuropeCE_7000BP_5500BP
FarmerEarly	Bar31	Turkey	Anatolia_Neolithic	40.3	29.61	8278.5	3.6490902710000000	XY	2.3_Anatolia_10000BP_8000BP
FarmerEarly	Bar8	Turkey	Anatolia_Neolithic	40.3	29.61	8071.0	7.171025264	XX	2.3_Anatolia_10000BP_8000BP
FarmerEarly	NE6	Hungary	Hungary_Neolithic_LBK	47.17	19.83	7051.5	0.937998868	XY	2.3_EuropeCE_7000BP_5500BP
FarmerEarly	PL_N25	Poland	Poland_Neolithic_BKG	52.61	18.8	6250.0	2.304346802	XX	2.3_EuropeCE_7000BP_5500BP
FarmerEarly	PL_N26	Poland	Poland_Neolithic_BKG	50.67	21.38	6151.0	2.146927132	XY	2.3_EuropeCE_7000BP_5500BP
FarmerEarly	PL_N19	Poland	Poland_Neolithic_FBC	52.62	18.96	5462.0	1.6841604730000000	XX	2.3_EuropeCE_7000BP_5500BP
FarmerEarly	Pal7	Greece	Greece_Neolithic	40.51	22.5	6351.0	1.265496659	XX	2.3_EuropeCE_7000BP_5500BP
FarmerEarly	PL_N28	Poland	Poland_Neolithic_BKG	52.61	18.8	6073.5	1.75499685	XY	2.3_EuropeCE_7000BP_5500BP
FarmerEarly	Rev5	Greece	Greece_Neolithic	40.33	22.56	8301.0	1.133693538	XX	2.3_EuropeS_8000BP_6000BP
FarmerEarly	Klei10	Greece	Greece_Neolithic	40.26	21.74	6062.5	2.047213233	XY	2.3_EuropeS_8000BP_6000BP
FarmerEarly	NE2	Hungary	Hungary_Neolithic_ALP	47.52	21.59	7123.5	0.148202309	XX	2.3_Europe_8500BP_5500BP
FarmerEarly	PL_N27	Poland	Poland_Neolithic_BKG	52.61	18.8	6250.0	1.790218266	XY	2.3_EuropeCE_7000BP_5500BP
FarmerEarly	NE7	Hungary	Hungary_Neolithic_Lengyel	47.17	19.83	6374.0	0.909572155	XY	2.3_EuropeCE_7000BP_5500BP
FarmerLate	NEO119	France	France_Neolithic	44.47	4.77	4382.0	0.10885148	XX	2.2_EuropeSW_6000BP_3500BP
FarmerLate	NEO886	Denmark	Denmark_Neolithic	54.99	12.42	5457.0	0.271693685	XY	2.4_EuropeNE_5600BP_4600BP
FarmerLate	NEO925	Denmark	Denmark_Neolithic	54.77	10.68	4947.0	0.294332844	XX	2.4_EuropeNE_5600BP_4600BP
FarmerLate	NEO653	Spain	Iberia_BronzeAge	43.4	-4.71	3423.0	0.283734421	XX	2.2_EuropeSW_6000BP_3500BP
FarmerLate	TV3831	Portugal	Iberia_BronzeAge	37.94	-7.6	3550.0	0.9943057240000000	XY	2.2_Iberia_7300BP_3500BP
FarmerLate	COV20126	Spain	Iberia_Neolithic	37.41	-4.42	5588.0	0.303266123	XY	2.2_EuropeSW_6000BP_3500BP
FarmerLate	NEO896	Denmark	Denmark_Neolithic	54.97	12.49	5446.0	0.121712208	XX	2.4_EuropeNE_5600BP_4600BP
FarmerLate	NEO43	Denmark	Denmark_Neolithic	55.99	10.25	5067.0	0.108400524	XY	2.4_EuropeNE_5600BP_4600BP
FarmerLate	NEO39	Sweden	Sweden_Neolithic	55.57	13.04	5074.0	0.174678265	XY	2.4_EuropeNE_5600BP_4600BP

FarmerLate	TV32032extra	Portugal	Iberia_BronzeAge	37.94	-7.6	3550.0	0.865992675	XY	2.2_Iberia_7300BP_3500BP
FarmerLate	NEO744	Denmark	Denmark_Neolithic	55.86	11.59	5333.0	0.220878469	XY	2.4_EuropeNE_5600BP_4600BP
FarmerLate	MonteGato104	Portugal	Iberia_BronzeAge	38.02	-7.86	3535.0	1.236961027	XY	2.2_Iberia_7300BP_3500BP
FarmerLate	NEO830	Italy	Italy_Neolithic	43.38	13.55	5393.0	0.104506976	XX	2.2_EuropeSW_6000BP_3500BP
FarmerLate	NEO757	Denmark	Denmark_Neolithic	55.9	11.12	5452.0	0.129920601	XX	2.4_EuropeNE_5600BP_4600BP
FarmerLate	atp9	Spain	Iberia_BronzeAge	42.35	-3.52	3634.0	0.419041583	XX	2.2_EuropeSW_6000BP_3500BP
FarmerLate	QUIN234	France	France_BronzeAge	43.3	1.96	3600.0	0.120622268	XX	2.2_EuropeSW_6000BP_3500BP
FarmerLate	ROS102	France	France_Neolithic_Grossgartach	48.5	7.47	6550.0	0.151456852	XY	2.3_Europe_8500BP_5500BP
FarmerLate	NEO640	Poland	Poland_Neolithic_FBC	50.27	20.45	4902.0	0.20055724	XY	2.4_EuropeNE_5600BP_4600BP
FarmerLate	NEO609	Portugal	Iberia_Neolithic	38.68	-9.16	4333.0	0.134584406	XY	2.2_EuropeSW_6000BP_3500BP
FarmerLate	NEO599	Denmark	Denmark_Neolithic	55.55	11.68	5134.0	0.19019233	XY	2.4_EuropeNE_5600BP_4600BP
FarmerLate	NEO597	Denmark	Denmark_Neolithic	55.59	11.57	5210.0	0.177413333	XY	2.4_EuropeNE_5600BP_4600BP
FarmerLate	NEO595	Denmark	Denmark_Neolithic	55.79	11.29	5452.0	0.218556781	XY	2.4_EuropeNE_5600BP_4600BP
FarmerLate	ans016	Sweden	Sweden_Neolithic	57.34	18.26	4646.0	0.340856997 00000000	XY	2.4_EuropeNE_5600BP_4600BP
FarmerLate	NEO945	Denmark	Denmark_Neolithic	56.49	9.83	5445.0	1.384502704	XX	2.4_EuropeNE_5600BP_4600BP
FarmerLate	NEO25	Denmark	Denmark_Neolithic	56.43	10.79	4956.0	0.356367248 00000000	XX	2.4_EuropeNE_5600BP_4600BP
FarmerLate	ans005	Sweden	Sweden_Neolithic_Megalithic	57.34	18.26	5265.0	0.129055433	XX	2.4_EuropeNE_5600BP_4600BP
FarmerLate	NEO721	Spain	Iberia_Neolithic	40.44	-3.5	4170.0	0.358210677 00000000	XX	2.2_Iberia_7300BP_3500BP
FarmerLate	NEO943	Denmark	Denmark_Neolithic	55.46	9.69	4614.0	1.754137916	XY	2.4_EuropeNE_5600BP_4600BP
FarmerLate	NEO942	Denmark	Denmark_Neolithic	55.58	11.29	5491.0	0.890575166	XX	2.4_EuropeNE_5600BP_4600BP
FarmerLate	NEO753	Denmark	Denmark_Neolithic	55.77	12.21	5531.0	0.163364465	XX	2.4_EuropeNE_5600BP_4600BP
FarmerLate	esp005	Spain	Iberia_BronzeAge_Cogotas	42.0	-1	3370.0	2.457932356	XY	2.2_Iberia_7300BP_3500BP
FarmerLate	pir001	Spain	Iberia_BronzeAge_Argar	37.89	-4.78	3725.0	0.215607431 00000000	XY	2.2_EuropeSW_6000BP_3500BP
FarmerLate	por004	Spain	Iberia_Neolithic	42.35	-3.52	4955.0	0.129932409	XY	2.2_EuropeSW_6000BP_3500BP

FarmerLate	san216	Spain	Iberia_Neolithic	42.69	-2.73	5665.5	0.200001027	XY	2.2_EuropeSW_6000BP_3500BP
FarmerLate	ans003	Sweden	Sweden_Neolithic_Megalithic	57.34	18.26	5250.0	0.135654055	XX	2.4_EuropeNE_5600BP_4600BP
FarmerLate	ValeOuro10207	Portugal	Iberia_BronzeAge	38.06	-8.11	3550.0	0.248412062	XX	2.2_EuropeSW_6000BP_3500BP
FarmerMiddle	NEO121	France	France_Neolithic	44.47	4.77	4531.0	0.530892193	XY	2.2_EuropeSW_6000BP_3500BP
FarmerMiddle	NEO28	Denmark	Denmark_Neolithic	55.91	12.31	5459.0	0.922300627	XY	2.4_EuropeNE_5600BP_4600BP
FarmerMiddle	NEO36	Sweden	Sweden_Neolithic	55.57	13.04	5097.0	2.384739487	XX	2.4_EuropeNE_5600BP_4600BP
FarmerMiddle	NEO29	Denmark	Denmark_Neolithic	55.13	10.9	5489.0	0.530600133	XX	2.4_EuropeNE_5600BP_4600BP
FarmerMiddle	NEO23	Denmark	Denmark_Neolithic	55.6	11.31	5533.0	3.33772768	XY	2.4_EuropeNE_5600BP_4600BP
FarmerMiddle	WET370	France	France_Neolithic	48.06	7.3	5521.0	0.172865405	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	NEO866	Denmark	Denmark_Neolithic	54.87	11.84	5456.0	1.521286722	XY	2.4_EuropeNE_5600BP_4600BP
FarmerMiddle	NEO891	Denmark	Denmark_Neolithic	55.73	12.1	5661.0	0.595471133	XX	2.4_EuropeNE_5600BP_4600BP
FarmerMiddle	ans017	Sweden	Sweden_Neolithic_Megalithic	57.34	18.26	5080.0	7.08351096	XY	2.4_EuropeNE_5600BP_4600BP
FarmerMiddle	CO1	Hungary	Hungary_Neolithic_Baden	47.17	19.83	4750.0	0.8595327420000000	XX	2.3_EuropeS_8000BP_6000BP
FarmerMiddle	RISE489	Italy	Italy_Neolithic	45.26	10.38	4693.0	0.50758911	XY	2.2_EuropeSW_6000BP_3500BP
FarmerMiddle	NEO641	Poland	Poland_Neolithic_FBC	50.27	20.45	5132.0	0.26184454	XX	2.3_EuropeS_8000BP_6000BP
FarmerMiddle	NEO630	UK	Britain_Neolithic	58.73	-2.94	4898.0	0.211902615	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	NEO627	UK	Britain_Neolithic	58.73	-2.94	5132.0	0.44952267100000000	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	NEO626	UK	Britain_Neolithic	58.73	-2.94	5082.0	0.37193905200000000	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	NEO624	UK	Britain_Neolithic	58.73	-2.94	4897.0	2.099933966	XX	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	NEO717	UK	Britain_Neolithic	58.73	-2.94	4893.0	0.481287489	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	NEO935	Denmark	Denmark_Neolithic	55.56	12.02	5187.0	5.027509563	XY	2.4_EuropeNE_5600BP_4600BP
FarmerMiddle	NEO933	Denmark	Denmark_Neolithic	55.25	10.75	5337.0	0.522015892	XY	2.4_EuropeNE_5600BP_4600BP
FarmerMiddle	Aveline_1	UK	Britain_Neolithic	51.32	-2.75	5489.5	0.782428624	XX	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	NEO142	Hungary	Hungary_Neolithic_Tisza	46.37	20.42	6641.0	0.895629723	XX	2.3_Europe_8500BP_5500BP

FarmerMiddle	BurnGround	UK	Britain_Neolithic	51.84	-1.85	5770.0	0.410850116	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	CaveHa3_1	UK	Britain_Neolithic	54.07	-2.29	5264.0	0.112689685	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	Coldrum_1	UK	Britain_Neolithic	51.32	0.37	5430.0	0.547611201	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	Embo_1	UK	Britain_Neolithic	57.91	-4	5050.0	0.127757081	XX	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	Fussels_Lodge_1	UK	Britain_Neolithic	51.09	-1.73	5657.5	0.73024687	XX	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	Jubilee_cave	UK	Britain_Neolithic	54.08	-2.27	5462.5	0.116653425	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	Kelco_cave	UK	Britain_Neolithic	54.07	-2.29	5536.0	0.115486462	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	Ballynahatty	Ireland	Ireland_Neolithic	54.54	-5.96	5131.5	10.0157214	XX	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	PL_N18	Poland	Poland_Neolithic_FBC	52.62	18.96	5462.5	1.909978283	XX	2.4_EuropeNE_5600BP_4600BP
FarmerMiddle	PSS4693	France	France_Neolithic_Noyen	48.52	3.6	5438.5	0.740638747	XX	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	PL_N38	Poland	Poland_Neolithic_GAC	52.61	18.9	5033.0	1.772002823 0000000	XX	2.4_Poland_5000BP_4700BP
FarmerMiddle	Carsington_pasture_1	UK	Britain_Neolithic	53.08	-1.64	5538.5	9.748563794 000000	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	PL_N20	Poland	Poland_Neolithic_FBC	52.62	18.96	5462.0	0.802395212 0000000	XY	2.4_EuropeNE_5600BP_4600BP
FarmerMiddle	Mor6	France	France_Neolithic_LBK	48.82	7.63	7036.0	0.161199361 00000000	XX	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	ros3	Sweden	Sweden_Neolithic_FBC	60.26	16.41	4955.0	0.380939671 00000000	XX	2.4_EuropeNE_5600BP_4600BP
FarmerMiddle	RISE1161	Poland	Poland_Neolithic_GAC	48.7	21.2	4757.0	1.366034772	XX	2.4_Poland_5000BP_4700BP
FarmerMiddle	RISE1165	Poland	Poland_Neolithic_GAC	48.7	21.2	4742.5	2.189830631	XY	2.4_Poland_5000BP_4700BP
FarmerMiddle	RISE1166	Poland	Poland_Neolithic_GAC	48.7	21.2	4907.5	3.058742732	XX	2.4_Poland_5000BP_4700BP
FarmerMiddle	RISE1249	Poland	Poland_Neolithic_GAC	50.2	21.4	4736.5	1.010443069 0000000	XX	2.4_Poland_5000BP_4700BP
FarmerMiddle	RISE1248	Poland	Poland_Neolithic_GAC	50.2	21.4	4725.0	0.799515376 0000000	XX	2.4_Poland_5000BP_4700BP
FarmerMiddle	RISE1246	Poland	Poland_Neolithic_GAC	50.2	21.4	4715.0	0.549865829	XX	2.4_Poland_5000BP_4700BP
FarmerMiddle	RISE1252	Poland	Poland_Neolithic_GAC	50.8	21.5	4725.0	0.430947658 00000000	XY	2.4_Poland_5000BP_4700BP
FarmerMiddle	RISE1250	Poland	Poland_Neolithic_GAC	50.6	21.7	4725.0	0.493666575	XY	2.4_Poland_5000BP_4700BP
FarmerMiddle	RISE1254	Poland	Poland_Neolithic_GAC	51.1	17.1	4725.0	0.437402634	XY	2.4_Poland_5000BP_4700BP

FarmerMiddle	Gok2	Sweden	Sweden_Neolithic_FBC	58.18	13.41	4850.0	1.220935916	XX	2.4_EuropeNE_5600BP_4600BP
FarmerMiddle	NEO847	UK	Britain_Neolithic	51.7	-2.3	5463.0	1.777915755	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	ros005	Sweden	Sweden_Neolithic_FBC	60.26	16.41	4740.0	0.886088897	XY	2.4_EuropeNE_5600BP_4600BP
FarmerMiddle	PEI 2.00	France	France_Neolithic_Campaniforme	43.14	2.25	4385.5	0.302517672	XY	2.2_EuropeSW_6000BP_3500BP
FarmerMiddle	R1014	Italy	Italy_Neolithic	41.37	13.29	4950.0	0.615266206	XY	2.2_Italy_7000BP_4000BP
FarmerMiddle	R104	Italy	Italy_LateAntiquity	41.89	12.48	1450.0	0.879014223	XY	2.2_Italy_7000BP_4000BP
FarmerMiddle	RISE1159	Poland	Poland_Neolithic_GAC	48.7	21.2	4730.0	27.46258284	XX	2.4_Poland_5000BP_4700BP
FarmerMiddle	RISE1170	Poland	Poland_Neolithic_GAC	48.7	21.2	4748.5	3.79009955	XX	2.4_Poland_5000BP_4700BP
FarmerMiddle	RISE1168	Poland	Poland_Neolithic_GAC	48.7	21.2	4676.0	18.93418868	XY	N/A
FarmerMiddle	mur	Spain	Iberia_Neolithic_Almagra	42.35	-3.52	7136.0	3.4675614490000000	XY	2.2_Iberia_7300BP_3500BP
FarmerMiddle	lai001	UK	Britain_Neolithic	59.13	-3.05	5180.0	0.225980982	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	mid001	UK	Britain_Neolithic_Megalithic	59.13	-3.05	5450.0	0.282625993	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	mid002	UK	Britain_Neolithic_Megalithic	57.75	-3.92	5180.0	0.25578715400000000	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	prs002	Ireland	Ireland_Neolithic_Megalithic	54.25	-8.56	5675.0	5.870842597	XX	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	prs003	Ireland	Ireland_Neolithic_Megalithic	54.25	-8.56	5600.0	0.22378720800000000	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	prs006	Ireland	Ireland_Neolithic_Megalithic	54.25	-8.56	5330.0	0.26352444200000000	XX	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	prs009	Ireland	Ireland_Neolithic_Megalithic	54.25	-8.56	5310.0	7.571866381	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	NEO823	Italy	Italy_BronzeAge	40.88	16.73	4665.0	0.404577629	XY	2.2_Italy_7000BP_4000BP
FarmerMiddle	prs010	Ireland	Ireland_Neolithic_Megalithic	54.25	-8.56	5530.0	0.23205746800000000	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	prs013	Ireland	Ireland_Neolithic_Megalithic	54.25	-8.56	5320.0	4.952996849	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	prs016	Ireland	Ireland_Neolithic_Megalithic	54.25	-8.56	5560.0	8.754512586	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	ans008	Sweden	Sweden_Neolithic_Megalithic	57.34	18.26	5135.0	2.027487661	XY	2.4_EuropeNE_5600BP_4600BP
FarmerMiddle	CB13	Spain	Iberia_Neolithic_Cardial	41.37	1.89	7348.0	0.931947851	XX	2.2_Iberia_7300BP_3500BP
FarmerMiddle	atp016	Spain	Iberia_Neolithic	42.35	-3.52	5039.5	13.20832827	XX	2.2_Iberia_7300BP_3500BP
FarmerMiddle	atp12-1420	Spain	Iberia_Neolithic	42.35	-3.52	4895.5	2.528221016	XY	2.2_Iberia_7300BP_3500BP

FarmerMiddle	c40331	Spain	Iberia_Neolithic	37.37	-4.25	5649.5	0.293459982	XY	2.2_Iberia_7300BP_3500BP
FarmerMiddle	prs012	Ireland	Ireland_Neolithic_Megalithic	54.25	-8.56	5660.0	0.25153112700000000	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	LugarCanto44	Portugal	Iberia_Neolithic	39.42	-8.82	5950.0	2.016550504	XX	2.2_Iberia_7300BP_3500BP
FarmerMiddle	RISE1241	Poland	Poland_Neolithic_GAC	50.6	21.7	4752.5	0.859098485	XY	2.4_Poland_5000BP_4700BP
FarmerMiddle	R22	Italy	Italy_Neolithic	40.81	8.44	3895.5	0.776104455	XX	2.2_Italy_7000BP_4000BP
FarmerMiddle	BERG157-2	France	France_Neolithic_BORSMichelsberg	43.22	2.41	6050.0	0.346274491	XX	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	BERG157-7	France	France_Neolithic_BORSMichelsberg	43.22	2.41	6131.5	0.267013925	XY	2.2_EuropeSW_6000BP_3500BP
FarmerMiddle	CabecoArruda117B	Portugal	Iberia_Neolithic	39.11	-8.66	5050.0	0.376607974	XY	2.2_Iberia_7300BP_3500BP
FarmerMiddle	BLP10	France	France_Neolithic_Michelsberg	49.39	3.74	6052.0	0.18285663600000000	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	BUCH2	France	France_Neolithic_Cemy	48.24	4.11	6250.0	0.36476039200000000	XY	2.2_EuropeSW_6000BP_3500BP
FarmerMiddle	NEO812	France	France_Neolithic_Cardial	43.32	2.42	6545.0	6.542360034	XX	2.2_Iberia_7300BP_3500BP
FarmerMiddle	CRE20D	France	France_Neolithic_ChasseenAncien	43.21	3.13	6151.0	0.256045807	XX	2.2_EuropeSW_6000BP_3500BP
FarmerMiddle	LugarCanto42	Portugal	Iberia_Neolithic	39.42	-8.82	5950.0	3.006333862	XX	2.2_Iberia_7300BP_3500BP
FarmerMiddle	LU339	Portugal	Iberia_Neolithic	41.71	-6.93	6797.5	4.60334026	XY	2.2_Iberia_7300BP_3500BP
FarmerMiddle	LD270	Portugal	Iberia_Neolithic	41.71	-6.93	6336.0	4.064587193	XY	2.2_Iberia_7300BP_3500BP
FarmerMiddle	LD1174	Spain	Iberia_Neolithic	37.41	-4.42	6415.0	3.558801721	XX	2.2_Iberia_7300BP_3500BP
FarmerMiddle	CabecoArruda122A	Portugal	Iberia_Neolithic	39.11	-8.66	5050.0	1.782958508	XY	2.2_Iberia_7300BP_3500BP
FarmerMiddle	CovaMoura364	Portugal	Iberia_Neolithic	38.75	-9.22	4900.0	0.794100402	XY	2.2_Iberia_7300BP_3500BP
FarmerMiddle	Es97-1	France	France_Neolithic_Michelsberg	50.92	1.71	6004.5	0.294790665	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	NEO790	Denmark	Denmark_Neolithic	55.71	12.27	5663.0	0.685227719	XY	2.4_EuropeNE_5600BP_4600BP
FarmerMiddle	CovaMoura9B	Portugal	Iberia_Neolithic	38.75	-9.22	4900.0	2.611737333	XX	2.2_Iberia_7300BP_3500BP
FarmerMiddle	R6	Italy	Italy_Neolithic	41.96	13.54	7159.5	0.604714196	XY	2.2_Italy_7000BP_4000BP
FarmerMiddle	bal004	UK	Britain_Neolithic_Megalithic	57.77	-3.9	5190.0	1.56791032700000000	XX	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	R24	Italy	Italy_Neolithic	40.81	8.44	5450.0	0.549967737	XX	2.2_Italy_7000BP_4000BP
FarmerMiddle	R25	Italy	Italy_Neolithic	40.81	8.44	4150.0	0.53976087	XX	2.2_Italy_7000BP_4000BP

FarmerMiddle	R26	Italy	Italy_Neolithic	40.81	8.44	4150.0	0.525953541	XY	2.2_Italy_7000BP_4000BP
FarmerMiddle	R27	Italy	Italy_Neolithic	40.81	8.44	4150.0	0.70739547	XY	2.2_Italy_7000BP_4000BP
FarmerMiddle	R29	Italy	Italy_Neolithic	40.81	8.44	4150.0	0.559161215	XY	2.2_Italy_7000BP_4000BP
FarmerMiddle	R28	Italy	Italy_Neolithic	40.81	8.44	4150.0	0.728827575	XX	2.2_Italy_7000BP_4000BP
FarmerMiddle	Dolmen Ansião 96B	Portugal	Iberia_Neolithic	39.75	-8.81	5450.0	1.962153759	XY	2.2_Iberia_7300BP_3500BP
FarmerMiddle	R4	Italy	Italy_Neolithic	41.96	13.54	4865.0	3.6761982310000000	XY	2.2_Italy_7000BP_4000BP
FarmerMiddle	R5	Italy	Italy_Neolithic	41.96	13.54	4839.5	1.5029056810000000	XX	2.2_Italy_7000BP_4000BP
FarmerMiddle	LugarCanto41	Portugal	Iberia_Neolithic	39.42	-8.82	5950.0	1.06714512	XX	2.2_Iberia_7300BP_3500BP
FarmerMiddle	BERG02-2	France	France_Neolithic_BORSMichelsberg	43.22	2.41	5870.0	0.344146565	XX	2.2_EuropeSW_6000BP_3500BP
WHG	NEO855	Denmark	Denmark_Mesolithic	56.4	10.72	6302.0	1.3829770000000000	XY	4.2_Denmark_10500BP_6000BP
WHG	NEO856	Denmark	Denmark_Mesolithic	56.37	10.64	6777.0	0.56205807	XX	4.2_Denmark_10500BP_6000BP
WHG	NEO679	Sweden	Sweden_Mesolithic	55.39	13.48	6834.0	0.164673359	XX	4.2_Denmark_10500BP_6000BP
WHG	NEO683	Denmark	Denmark_Mesolithic	55.4	9.83	7529.0	1.81852777	XX	4.2_Denmark_10500BP_6000BP
WHG	NEO938	Spain	Iberia_Mesolithic	43.4	-4.71	7878.0	0.47515145700000000	XX	4.2_Iberia_9000BP_7000BP
WHG	NEO694	Spain	Iberia_Mesolithic	38.73	-0.46	9217.0	0.28475883400000000	XY	4.2_Iberia_9000BP_7000BP
WHG	NEO853	Denmark	Denmark_Mesolithic	55.55	10.62	6047.0	1.964862968	XX	4.2_Denmark_10500BP_6000BP
WHG	NEO852	Denmark	Denmark_Mesolithic	56.03	10.26	6308.0	0.189591791	XX	4.2_Denmark_10500BP_6000BP
WHG	NEO733	Denmark	Denmark_Mesolithic	55.77	11.39	6824.0	1.3166819510000000	XX	4.2_Denmark_10500BP_6000BP
WHG	NEO791	Denmark	Denmark_Mesolithic	55.33	11.15	7048.0	2.492448215	XY	4.2_Denmark_10500BP_6000BP
WHG	NEO941	Denmark	Denmark_Mesolithic	56.71	10.17	6372.0	0.135296816	XX	4.2_Denmark_10500BP_6000BP
WHG	NEO751	Denmark	Denmark_Mesolithic	56.87	9.22	6343.0	0.297822065	XY	4.2_Denmark_10500BP_6000BP
WHG	NEO932	Denmark	Denmark_Mesolithic	55.25	11.23	7499.0	2.7601622000000000	XX	4.2_Denmark_10500BP_6000BP
WHG	NEO749	Denmark	Denmark_Mesolithic	55.85	12.56	7070.0	1.905133435	XY	4.2_Denmark_10500BP_6000BP
WHG	NEO747	Denmark	Denmark_Mesolithic	55.85	12.56	6729.0	0.24942735800000000	XY	4.2_Denmark_10500BP_6000BP
WHG	NEO745	Denmark	Denmark_Mesolithic	55.85	12.56	6790.0	0.447895875	XY	4.2_Denmark_10500BP_6000BP
WHG	NEO960	Denmark	Denmark_Mesolithic	55.58	11.58	5926.0	0.150141897	XY	4.2_Denmark_10500BP_6000BP

WHG	NEO759	Denmark	Denmark_Mesolithic	55.4	12.37	9028.0	2.948103712	XY	4.2_Denmark_10500BP_6000BP
WHG	syltholm	Denmark	Denmark_Mesolithic	54.65	11.35	7709.5	2.291787968	XX	4.2_Denmark_10500BP_6000BP
WHG	NEO648	Spain	Iberia_Mesolithic	43.4	-4.71	7539.0	1.844283075	XY	4.2_Iberia_9000BP_7000BP
WHG	Cheddar_man	UK	Britain_Mesolithic	51.28	-2.77	10300.0	2.054202123	XY	4.2_EuropeW_13500BP_8000BP
WHG	PL_N22	Poland	Poland_Neolithic_BKG	52.61	18.9	6291.0	1.491551035	XX	4.2_EuropeE_8600BP_6000BP
WHG	KO1	Hungary	Hungary_Neolithic_Koros	47.56	20.72	7660.0	1.014600016	XY	4.2_EuropeE_8600BP_6000BP
WHG	Canes1	Spain	Iberia_Mesolithic	43.36	-4.72	7115.0	1.6461215340000000	XX	4.2_Iberia_9000BP_7000BP
WHG	Chan	Spain	Iberia_Mesolithic	42.73	-7.03	9131.0	5.008215765	XX	4.2_Iberia_9000BP_7000BP
WHG	Bichon	Switzerland	Switzerland_Mesolithic	47.1	6.87	13665.0	7.692393112	XY	4.2_EuropeW_13500BP_8000BP
WHG	Loschbour	Luxembourg	Luxembourg_Mesolithic	49.81	6.4	8050.0	18.23029647	XY	4.2_EuropeW_13500BP_8000BP
WHG	Brana	Spain	Iberia_Mesolithic	42.91	-5.38	7815.0	3.019525774	XY	4.2_Iberia_9000BP_7000BP
WHG	R11	Italy	Italy_Mesolithic	41.96	13.54	11908.0	0.957641603	XY	4.2_Italy_15000BP_9000BP
WHG	NEO669	Serbia	Serbia_Mesolithic	44.56	22.03	7950.0	0.23932044	XX	4.2_EuropeE_8600BP_6000BP
WHG	R7	Italy	Italy_Mesolithic	41.96	13.54	10681.5	3.153769086	XY	4.2_Italy_15000BP_9000BP
WHG	ST3	Italy	Italy_Mesolithic	37.85	14.7	14800.0	0.475034886	XY	4.2_Italy_15000BP_9000BP
WHG	PER1150503	France	France_Mesolithic	45.77	0.33	9067.0	0.315139903	XX	4.2_EuropeW_13500BP_8000BP
WHG	PER3023	France	France_Mesolithic	45.77	0.33	9067.0	0.161333797	XX	4.2_EuropeW_13500BP_8000BP
WHG	R15	Italy	Italy_Mesolithic	41.96	13.54	9124.5	3.070164483	XY	4.2_Italy_15000BP_9000BP
WHG	NEO91	Denmark	Denmark_Mesolithic	55.39	12.31	9122.0	1.176549838	XY	4.2_Denmark_10500BP_6000BP
WHG	NEO646	Spain	Iberia_Mesolithic	43.4	-4.71	8273.0	1.590267827	XX	4.2_Iberia_9000BP_7000BP
WHG	NEO645	Denmark	Denmark_Mesolithic	55.91	11.09	5870.0	0.211986752	XY	4.2_Denmark_10500BP_6000BP
WHG	NEO598	Denmark	Denmark_Mesolithic	55.95	11.9	6075.0	0.7272044320000000	XX	4.2_Denmark_10500BP_6000BP
WHG	NEO19	Denmark	Denmark_Mesolithic	56.27	10.47	8163.0	3.262849417	XY	4.2_Denmark_10500BP_6000BP
WHG	NEO586	Denmark	Denmark_Mesolithic	56.37	10.57	7031.0	0.201188562	XY	4.2_Denmark_10500BP_6000BP
WHG	NEO583	Denmark	Denmark_Mesolithic	56.37	10.57	6981.0	0.17639908900000000	XX	4.2_Denmark_10500BP_6000BP
WHG	NEO570	Denmark	Denmark_Mesolithic	56.4	10.72	6369.0	2.861753026	XX	4.2_Denmark_10500BP_6000BP

WHG	NEO589	Denmark	Denmark_Mesolithic	55.33	11.15	7478.0	7.410700578000000	XY	4.2_Denmark_10500BP_6000BP
WHG	NEO254	Denmark	Denmark_Mesolithic	55.4	10.13	10463.0	0.41962998	XY	4.2_Denmark_10500BP_6000BP
WHG	NEO123	Denmark	Denmark_Mesolithic	54.96	11.85	8182.0	0.286386228	XX	4.2_Denmark_10500BP_6000BP
WHG	NEO122	Denmark	Denmark_Mesolithic	54.96	11.85	8146.0	0.564966744	XX	4.2_Denmark_10500BP_6000BP
WHG	NEO568	Denmark	Denmark_Mesolithic	56.81	9.18	6586.0	1.9814869470000000	XY	4.2_Denmark_10500BP_6000BP
Yamnaya	poz81	Poland	Poland_Neolithic_CWC	52.29	17.55	4705.0	1.92879788	XY	1.2_EuropeNE_4800BP_3000BP
Yamnaya	RISE509	Russia	Siberia_BronzeAge_Afanasievo	54.36	90.92	4732.0	4.52834127	XX	1.2_Steppe_5000BP_4300BP
Yamnaya	RISE511	Russia	Siberia_BronzeAge_Afanasievo	54.36	90.92	4744.0	5.20403929	XX	1.2_Steppe_5000BP_4300BP
Yamnaya	RISE546	Russia	Russia_BronzeAge_Yamnaya	46.54	43.7	4850.0	0.125905828	XY	1.2_Steppe_5000BP_4300BP
Yamnaya	RISE547	Russia	Russia_BronzeAge_Yamnaya	46.54	43.7	4710.5	0.6864666010000000	XY	1.2_Steppe_5000BP_4300BP
Yamnaya	RISE548	Russia	Russia_BronzeAge_Yamnaya	46.54	43.7	4850.0	0.9108783580000000	XY	1.2_Steppe_5000BP_4300BP
Yamnaya	RISE550	Russia	Russia_BronzeAge_Yamnaya	46.56	43.68	4934.5	0.440260727	XY	1.2_Steppe_5000BP_4300BP
Yamnaya	RISE555	Russia	Russia_BronzeAge	48.72	44.5	4627.0	0.237337432	XY	1.2_Steppe_5000BP_4300BP
Yamnaya	Yamnaya	Kazakhstan	Kazakhstan_BronzeAge_Yamnaya	49.13	75.85	4902.5	26.39165529	XY	1.2_Steppe_5000BP_4300BP
Yamnaya	MJ-09	Ukraine	Ukraine_BronzeAge_Catacomb	47.43	34.27	4285.5	0.19947548700000000	XX	1.2_Steppe_5000BP_4300BP
Yamnaya	MJ-06	Ukraine	Ukraine_BronzeAge_Yamnaya	49.32	35.37	4629.5	0.161999877	XX	1.2_EuropeNE_4800BP_3000BP
Yamnaya	NEO175	Russia	Russia_Neolithic_Sredny	52.28	38.96	4607.0	0.416698274	XY	1.2_Steppe_5000BP_4300BP
Yamnaya	Latvia_LN1	Latvia	Latvia_Neolithic_CWC	56.28	25.13	4833.0	0.197755635	XX	1.2_EuropeNE_4800BP_3000BP
Yamnaya	RISE552	Russia	Russia_BronzeAge_Yamnaya	46.62	43.33	4446.0	2.458824579	XY	1.2_Steppe_5000BP_4300BP
Yamnaya	RISE240	Russia	Russia_BronzeAge_Yamnaya	46.58	43.68	4706.0	0.173195772	XX	1.2_Steppe_5000BP_4300BP

Table S3h.1. Metadata and grouping of ancient individuals into reference populations.

References

1. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet.* **8**, e1002453 (2012).

2. Margaryan, A. *et al.* Population genomics of the Viking world. *Nature* **585**, 390–396 (2020).

3. Hellenthal, G. *et al.* A genetic atlas of human admixture history. *Science* **343**, 747–751 (2014).

4. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

5. Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233 (2003).

6. Sarmanova, A., Morris, T. & Lawson, D. J. Population stratification in GWAS meta-analysis should be standardized to the best available reference datasets. *bioRxiv* 2020.09.03.281568 (2020) doi:10.1101/2020.09.03.281568.

7. Galinsky, K. J., Loh, P.-R., Mallick, S., Patterson, N. J. & Price, A. L. Population Structure of UK Biobank and Ancient Eurasians Reveals Adaptation at Genes Influencing Blood Pressure. *Am. J. Hum. Genet.* **99**, 1130–1139 (2016).

8. Patterson, N. *et al.* Large-scale migration into Britain during the Middle to Late Bronze Age. *Nature* (2021) doi:10.1038/s41586-021-04287-4.

9. Sikora, M. *et al.* The population history of northeastern Siberia since the Pleistocene. *Nature* **570**, 182–188 (2019).

10. Shinde, V. *et al.* An Ancient Harappan Genome Lacks Ancestry from Steppe Pastoralists or Iranian Farmers. *Cell* **179**, 729–735.e10 (2019).

11. Hofmanová, Z. *et al.* Early farmers from across Europe directly descended from Neolithic Aegeans. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 6886–6891 (2016).

12. Feldman, M. *et al.* Late Pleistocene human genome suggests a local origin for the first farmers of central Anatolia. *Nat. Commun.* **10**, 1218 (2019).

13. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

14. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

3i) Building a population history model of Europeans and using it to assign local ancestry to all haplotypes in modern and ancient European samples

Alice Pearson[1], Richard Durbin[1,2]

[1]Department of Genetics, University of Cambridge, UK.
[2]Wellcome Sanger Institute, Wellcome Genome Campus, UK.

Introduction

We built a quantitative admixture graph model that represents the major ancestry flows contributing to modern European genomes over the last 50,000 years. Within this model we placed chromosome sampling points from appropriate populations and times that resemble European and West Asian samples from the imputed present aDNA dataset. Based on the model, we developed an approach using genealogical nearest neighbours [1] in tree sequences inferred using the Relate program [2] to estimate the path backwards in time through the population structure taken by each modern and ancient haplotype at each position in the genome. Throughout we evaluate our methods by comparing results from the real data to those from simulations of the model.

Method and Results

Model of population structure

We first filtered the present aDNA dataset to include 1015 ancient genomes most relevant to modern European genetic history. This included samples that have a West Eurasian archeological location and lying on EHG-WHG-CHG-Farmer clines in Principle Component Analysis and excluded very old West Eurasian samples and archaics. These were merged with all 503 present-day European 1000 Genomes Project samples to make up our dataset, totalling 1518 diploid samples. Figure **S3i.**1 shows a schematic of our model that describes the evolution of population structure in Europe during the last 50 ky. We used parameters subjectively estimated from Principal Component Analysis and Jones *et al.* [3], and constructed a simulator for this model in msprime (a coalescent simulator) [4]. Shortly after

the expansion of anatomically modern humans into Eurasia there is an early population split 45 kya between the Northern Europeans (NE) who continued travelling north west into Europe and West Asians (WA) who stayed more locally in the Levant and South Caucasus area. The WA population then splits to form the Caucasus Hunter-Gatherers (CHGs) and the Anatolian Farmers (Ana) 24kya. 18kya the Western Hunter-Gatherers (WHGs) and Eastern Hunter-Gatherers (EHGs) begin to diverge within the NE. At that point, the four genetic strands that make up present day European ancestry are distinct in our model and subsequently admixture between these components describes the formation of modern European gene pool; we show the formation of the Neolithic Farmers (Neo) from admixture between WHGs and the Ana 6kya, the formation of the Yamnaya Steppe people (Yam) from admixture of EHGs and CHGs during the early Bronze Age 5.4 kya and finally the formation of the Bronze Age gene pool 4.2kya as a two-way mix between the Yam and the Neo. From the Bronze Age up to present-day, there is simply exponential growth in population size.

The sampling distribution in the model is based on the numbers and average ages of sample groups in the present aDNA data set, where Iron Age, Viking and Late Antiquity samples younger than 2500 years BP are grouped with the present day genomes and those older are grouped with Bronze Age samples.

Local Ancestry Using Tree Sequences:

In the absence of recombination, a set of sample chromosomes from different genomes are related to each other by a single tree, identifiable by the sharing of derived mutations. A recombination event between any two chromosomes changes the tree topology. As you move from one end of the sample chromosomes to the other, a sequence of changing trees is observed, each encoding the genealogy of a segment of DNA (sample haplotypes) and each tree change reflecting one or more recombination events. The Relate software [2] aims to infer the true underlying tree sequence from genotype data. We worked with three types of tree sequence data; tree sequences simulated directly from the model using msprime (model simulated), tree sequences inferred by Relate from genotype data simulated from the model (Relate simulated) and tree sequences inferred by Relate from the real genotype data (Relate MesoNeo).

With ancient samples and admixture events, the first coalescence alone is insufficient for understanding the full ancestry of a given haplotype. This is because the first coalescence event may occur at a time younger than the age of some sampled groups, in which case the older sampled individuals could not be found as the closest relatives and therefore the true local ancestry of haplotypes may not be correctly established. Similarly, with some sampled populations formed via admixture of other sampled populations, the closest relatives to a haplotype may by chance be from the admixed population alone even if the age of first coalescence is old enough to capture all sampled groups, in which case again the true local ancestry of a haplotype may not be found. The path that a haplotype takes backwards in time from the present day to the root of the model is more informative about its local ancestry as its relationship to all relevant historical and admixing populations is established. If we assume an infinite-sites model, all alleles only appear once in history by mutation and all take a single path from the present day to the root.

The method involves taking a focal haplotype and its marginal tree from the tree sequence describing its genealogical relationship to all other sample haplotypes. From that haplotype we traverse up the tree, jumping to successive parent nodes towards the root i.e parent, grandparent, great-grandparent etc. We want to identify what population each of these nodes (ancestors) occurs in, which in turn describes what path the haplotype has taken backwards in time. In simulated tree sequences, we can simply obtain (using tskit) the population identity of each node but in tree sequences inferred with Relate this is not possible. We therefore adapted the concept of Genealogical Nearest Neighbours [1]. At each node we record its age and the number of reference samples belonging to each ancestral group in the leaves below that node as a proportion of all the reference samples in the leaves below that node (not including the focal leaf itself or leaves seen at previously analysed nodes further down the tree). I refer to the distribution of reference ancestry proportions and age at a node as GNN*x* (Genealogical Nearest Neighbours at *x*) where *x* is the *x*th node examined towards the root. The ordered collection of all GNNx distributions of all *x* nodes examined during a tree traversal describes the population identity of each node and therefore the path that the focal haplotype has taken to the root. The key is to consider the GNNs together, not independently.

We therefore first define the set of ancient reference samples from our dataset that are genetically diagnostic of each of the 7 ancestral groups. These samples form tight clusters

within broader ancestry clouds in the PCA. In order to assign paths to millions of sample lineages we implemented a supervised machine learning method using the Python Keras package with a TensorFlow backend [5]. The input to the network is the ordered collection of GNN distributions plus the node age for the first five informative nodes traversed towards the root from a single sample haplotype, configured as a 5 x 8 matrix. Informative nodes are those that have at least one leaf from the reference set of ancient samples. If we reach the root of the tree in less than five informative nodes, then the remaining rows of the matrix are filled with -15 as padding. The output of the network is a numerical label 0-6 characterising the path taken by the haplotype backwards in time. These paths can be seen clearly in Figure **S3i.**1 and are described in Table **S3i.**1. Given the sampling distribution, there is not always an informative node in the area of the model where the four ancestries are distinct, 200-600 generations ago. Labels 5, 6 and 0 capture these situations where we have limited information to assign the full path and instead only a partial path or none at all.

Path label	**Populations of inheritance**
0	None can be determined
1	Neolithic Farmers/Anatolian/West Asian
2	Yamnaya/CHG/West Asian
3	Neolithic Farmers/WHG/Northern European
4	Yamnaya/EHG/Northern European
5	Northern European
6	West Asian

Table S3i.1: The populations that carried a haplotype through time by inheritance, ordered backwards in time that make up each path.

A large amount of training data can be generated by simulation of tree sequences from the model. We simulated a tree sequence and corresponding VCF file of chromosome 3 and applied Relate to the VCF file to obtain a Relate simulated tree sequence. We then extracted GNN distributions, traversing from all Present-day, Bronze Age, Neo and Yam samples at evenly spaced trees from the Relate simulated tree sequence. Path labels at corresponding sites were extracted from the model simulated tree sequence. WHG, EHG, CHG, and Ana can be given paths 1-4 from their population identity alone and therefore are not involved in the network training. Our total training set consisted of 4,000,000 GNN matrices and labels.

85% of true labels were full paths 1-4. The network was trained using a categorical cross-entropy function, Adam optimisation and a batch size of 30. Training took place over 30 epochs. For testing we generated another Relate Simulated tree sequence and extracted 1,000,000 Relate GNNs and simulated true labels. The network displayed 75% accuracy. Figure **S3i.**B is a confusion matrix comparing the classed labels to the true labels.

Relate tree sequences were inferred for all chromosomes of our merged dataset, using a mutation rate of 1.25×10^{-8} per bp per generation and fine-scale human recombination maps. We assigned paths for all samples at every site using the trained network. We annotated the merged VCF files with the path assignments as a FORMAT tag with ID=AP standing for Ancestral Path.

Admixture fractions for hybrid groups can be estimated from the local ancestry assignments as the proportion of sites taking each of the relevant paths out of all sites in all samples of a hybrid group. We calculated this genome-wide from the VCF files containing the new AP format tag. The results are shown in Figure **S3i.**3.

Figures:

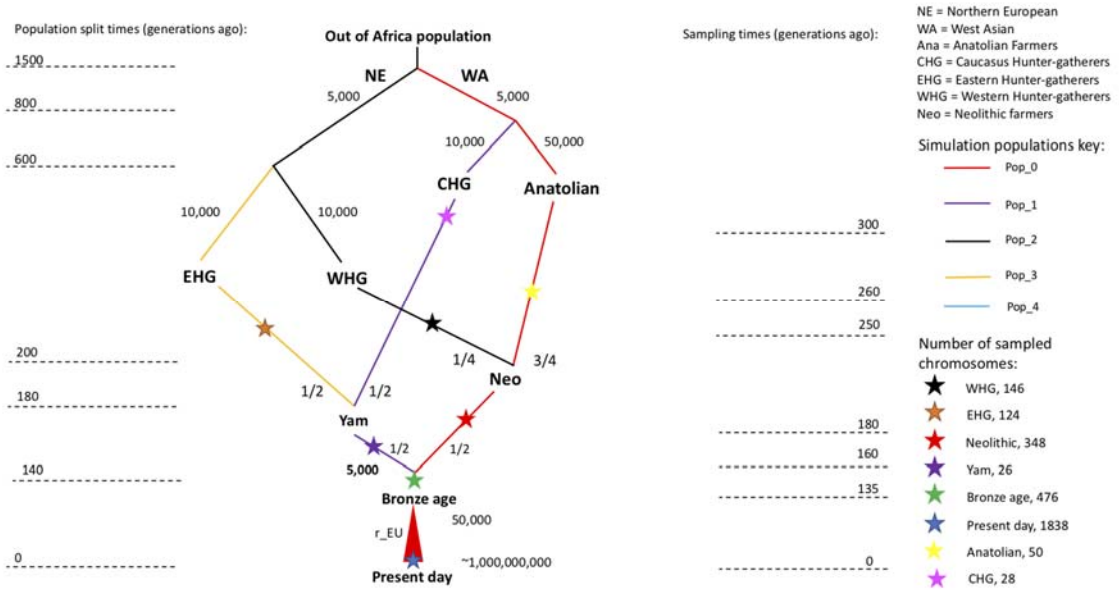


Figure S3i.1: A schematic of the model of population structure in Europe that was built in msprime. Moving down the figure is forwards in time and the population split times and admixture times are given in generations ago. Coloured lines represent the four populations declared in the simulation that extend through time. Sampled populations and times are marked with a star and the number of chromosomes sampled is given in the key.

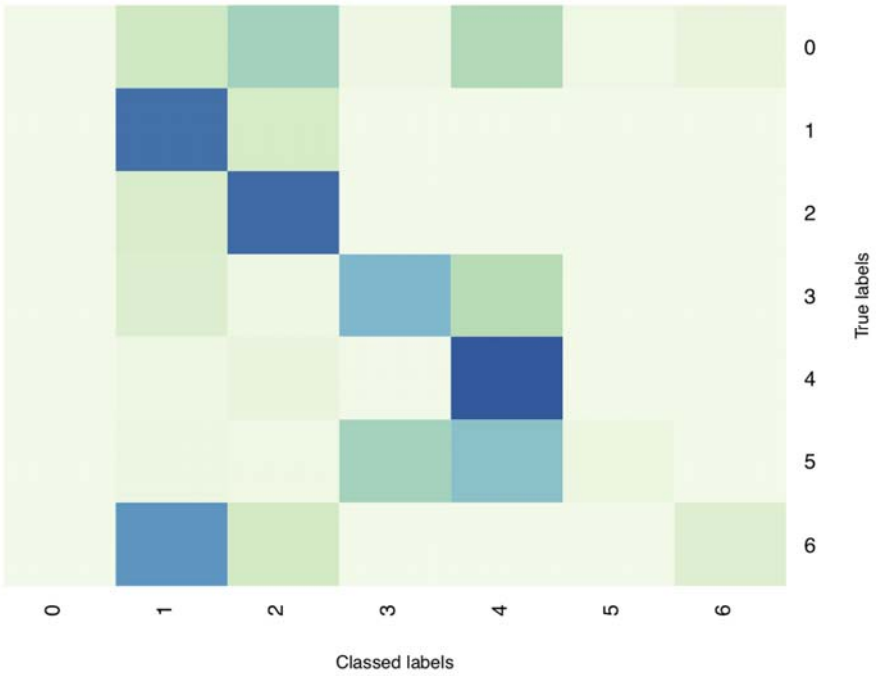


Figure S3i.2: Heatmap of the confusion matrix between the true labels and classed labels for Relate simulated test data. This is normalised by the sum of rows i.e of the true labels, how many the classifier correctly predicted.

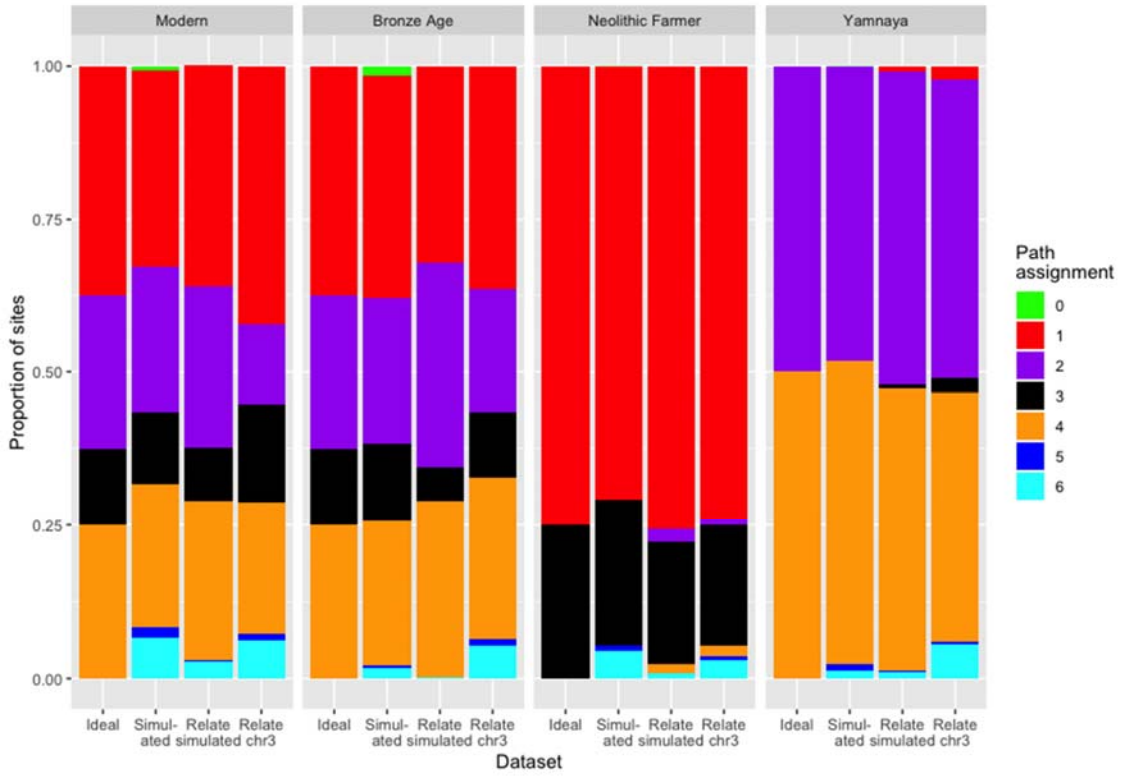


Figure S3i.3: Admixture fractions displayed as bars divided by the proportion of all sites assigned each path 0-6. For each admixed population there are four bars from four different tree sequences: *Ideal* is our best guess at the admixture fractions given previously published results and shows the fractions we used as input in msprime simulation. *Simulated* are the proportions of sites assigned each path extracted from the msprime simulated tree sequences which we used to train the neural, therefore partial paths are included where the node distribution means full paths cannot be determined. *Relate simulated* are the proportions of sites classed as each path in a Relate tree sequence inferred from simulated data, when GNNs are extracted from this tree sequence and classified as paths by the neural network. *Relate chr3* are the proportions of sites classed as each path in a Relate tree sequence inferred from the real (ancient genomic) chromosome 3 data, when GNNs are extracted from this tree sequence and classified as paths by the neural network.

References

1. Kelleher, J. *et al.* Inferring whole-genome histories in large population datasets. *Nat. Genet.* **51**, 1330–1338 (2019).

2. Speidel, L., Forest, M., Shi, S. & Myers, S. R. A method for genome-wide genealogy estimation for thousands of samples. *Nat. Genet.* **51**, 1321–1329 (2019).

3. Jones, E. R. *et al.* Upper Palaeolithic genomes reveal deep roots of modern

Eurasians. *Nat. Commun.* **6**, 1–8 (2015).

4. Kelleher, J., Etheridge, A. M. & McVean, G. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLoS Comput. Biol.* **12**, e1004842 (2016).

5. Chollet, F. *et al.* Keras. https://keras.io (2015).

4) Natural selection and trait evolution

4a) Estimating allele frequency trajectories of trait-associated variants

Evan K. Irving-Pease[1], Aaron J. Stern[2], Rasmus Nielsen[1,2], Fernando Racimo[1]

[1]Lundbeck Foundation GeoGenetics Centre, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark
[2]Department of Integrative Biology, University of California, Berkeley

Introduction

Genome-wide association studies (GWAS) of present-day human populations have identified large numbers of genetic variants associated with complex traits. However, the extent to which these variants have been under positive selection during recent human evolution is unclear. We aimed to model the allele frequency trajectories and selection coefficients of GWAS variants through time, using genomic data from both present-day populations and ancient individuals sampled across West Eurasia during the Holocene. We used the software *CLUES* [1], which supports inference of allele frequency trajectories from marginal trees sampled from a reconstruction of the ancestral recombination graph (ARG) [2] for a set of genomic sequences, in combination with genotype likelihoods from serially sampled ancient DNA (aDNA).

To account for population structure in our samples, we applied a novel chromosome painting technique (Supplementary Note S3i). This technique is based on inference of a sample's nearest neighbours in the marginal trees of an ARG that contains labelled individuals. In our case, the labelling corresponds to ancestral populations that predate the main episodes of admixture in West Eurasia (Supplementary Note S3i). This method allows us to accurately assign ancestral population labels to haplotypes found in both ancient and present-day individuals. By conditioning our selection analyses on these haplotype backgrounds, we can infer the selection trajectories of GWAS risk alleles in a manner that is approximately invariant to change in the admixture proportions through time. These ancestry specific allele trajectories reveal many novel aspects about the dynamic interplay between selection and admixture in West Eurasia throughout the Holocene.

Methods

The computational pipeline to perform all analyses was written in the `snakemake` workflow management system [3]. For a full list of all the software and versions used, see Table S4a.1. The directed acyclic graph (DAG) of the computational pipeline is shown in Figure S4a.1. All pipeline code, custom scripts and a conda environment to replicate the analyses are available in the GitHub repository (https://github.com/ekirving/mesoneo_paper).

SNP Ascertainment

GWAS SNPs

We ascertained -a list of GWAS targets by downloading version v1.0.2 (2020-06-04) of the NHGRI-EBI GWAS Catalog [4]; containing 187,403 GWAS associations for 3,735 traits. To account for the varying significance thresholds used in the 4,007 published studies included in the catalogue, we restricted our analysis to SNPs with a genome-wide significance threshold of p<5e-8. We further filtered the catalogue to retain only single-nucleotide polymorphisms (SNPs) with a valid dbSNP Reference SNP identifier (rsID); resulting in 121,795 GWAS associations for 70,224 rsIDs. For each of the retained associations, we retrieved the trait ontology hierarchy by querying the EMBL-EBI Ontology Lookup Service [5]. We then queried the Ensembl REST API [6], to retrieve metadata about each rsID; including chromosome and position in the GRCh37 assembly, ancestral allele, and nearest genes.

Control SNPs

To determine the extent to which GWAS variants are enriched for selection, we paired each GWAS SNP with a unique “Control SNP”. Control SNPs were ascertained by selecting all biallelic SNPs within the imputed dataset (Supplementary Note 2) and excluding any that fell within +/- 50 kb of a GWAS SNP or a gene region. Gene annotations for GRCh37 were downloaded from Ensembl (release 87) [7]. Control SNPs were grouped into bins based on their derived allele frequency (DAF), rounded to the nearest 1%, and paired randomly (without replacement) with GWAS SNPs in the same chromosome and DAF bin.

Simulated Neutral SNPs

To measure the effects of demography on the modelled allele frequency trajectories, we frequency paired GWAS SNPs with neutral SNPs, simulated under the demography used to train the chromosome painting model (Supplementary Note 3i). Neutral simulations were performed with *msprime* [8], for genomes of length 198 Mbp (i.e., the approximate length of chr3), with sample sizes and ages broadly equivalent to those in the empirical aDNA dataset.

1000G ARG

We built genome-wide genealogies for all samples in the 1000 Genomes Project (1000G) Phase 3 release [9] using the software *Relate* (1.1.3) [2].

Data pre-processing

Prior to inference of the ARG, we converted VCF files into HAPS format, removed all non-biallelic SNPs, polarised SNPs against the ancestral allele calls from the Ensembl Compara 71 database (ens-staging2:3306) [10], filtered sites using the 1000 Genomes StrictMask (20140520) and generated SNP annotations using `Relate`.

ARG Inference

We jointly inferred genome-wide genealogies for all 1000G Phase 3 samples, using `Relate`, assuming a mutation rate of 1.25e-8 and an Ne of 30,000. From this ARG, we extracted subtrees containing samples belonging to three European (EUR) populations: (i) Finnish in Finland (FIN); (ii) British in England and Scotland (GBR); and (iii) Toscani in Italia (TSI). We used these subtrees to jointly reinfer branch lengths and to infer a population size history for the EUR metapopulation. Lastly, we remapped all SNPs which had been pruned during inference of the ARG onto the branch-length calibrated EUR subtrees.

Modifications to CLUES

Here we describe several modifications made to *CLUES* (see the GitHub wiki page for more information https://github.com/35ajstern/clues/wiki).

ARG sampling using Relate

Instead of sampling ARGs using *ARGweaver* [11], we sample ARGs using *Relate* [2] for scalability reasons. `Relate` differs from *ARGweaver* in that it assumes a continuous-time coalescent process (vs discrete-time for *ARGweaver*); hence we modified the hidden Markov model (HMM) used by *CLUES* to (1) take time steps every generation, *vs* over a smaller number (~10-50) of timesteps; and (2) within time steps, the probability density of coalescence is calculated using the approach of [12] and [13], *vs* the discrete-time lines-of-descent approach used in [11] and [1].

Ancient DNA samples

We also introduced a new feature that allows the user to specify a time series of ancient genotype likelihoods which are incorporated into the HMM. We incorporate these samples by, for a given timestep *t,* including (i.e. multiplying by) a Binomial(n=2, p=X_t) emission probability for each ancient sampled during timestep *t* in the HMM, where X_t is the latent allele frequency during timestep *t.* In this particular application, we supplanted genotype likelihoods with genotype posterior probabilities (which should be identical under a uniform prior); we confirmed through tests that this did not yield any systematic biases.

Selection Analysis

CLUES with Modern 1000G data

For each modelled SNP, we used `Relate` to draw 100 samples from the MCMC posterior distribution of trees at that locus. Trees were sampled assuming a mutation rate of 1.25e-8 [2,14] and using the population size history from the EUR calibrated subtrees.

We ran *CLUES* to infer allele frequency trajectories and selection coefficients from modern 1000G data, using: (i) the 100 sampled trees from *Relate* (`--times`); (ii) the inferred EUR population size history (`--coal`); (iii) with trajectories polarised the by the derived allele (`--A1`); (iv) a terminal frequency equal to the DAF of each SNP in EUR (`--popFreq`); and (v) constrained selection to a single epoch spanning the last 15,000 years (`--timeBins`).

We ran these models for all GWAS and Control SNPs present in the imputed dataset for which `Relate` was able to confidently map a mutation to the inferred trees (n=73,232).

CLUES with aDNA Time Series

We also ran *CLUES* in an alternative mode, excluding the modern ARG data, and replacing them with aDNA time series data, using: (i) the time series of aDNA genotype probabilities (`--ancientSamps`) (ii) the inferred EUR population size history (`--coal`); (iii) a terminal frequency equal to the DAF of each SNP in EUR (`--popFreq`); and (iv) constraining selection to a single epoch spanning the last 15,000 years (`--timeBins`).

We ran these models for all GWAS and Control SNPs present in the imputed dataset, irrespective of their mappability in the `Relate` analyses (n=73,988), as well as for all Simulated SNPs (n=11,665).

CLUES with aDNA Ancestral Paintings

We also ran four additional models for each GWAS and Control SNP, in which we conditioned the time series of aDNA genotype probabilities on one of four specific ancestral haplotype pathways (Supplementary Note S3i):

1. ANA (Anatolian Farmers -> Neolithic)
2. CHG (Caucasus Hunter-gatherers -> Yamnaya)
3. WHG (Western Hunter-gatherers -> Neolithic)
4. EHG (Eastern Hunter-gatherers -> Yamnaya)

In a minority of cases, the chromosome painting model assigned a haplotype to one of two basal pathways: (i) North European ancestry (WHG and EHG); and (ii) West Asian ancestry (CHG and ANA). In such cases, we included these haplotypes in the conditional analyses for both downstream pathways.

All other particulars of these models were identical to the earlier aDNA analyses, except that genotypes were passed to *CLUES* in haploid mode (`--ancientHaps`), even when both haplotypes in an individual shared the same painting.

For the simulated SNPs, we ran these analyses on two different datasets: (i) a simulation labelled with the true pathways of each haplotype; and (ii) a simulation in which the pathways were inferred by the chromosome painting model.

Reference and mapping bias filters

To address issues of mapping biases that may cause artifactual changes in allele frequency at individual sites, we also constructed a causal model for distinguishing direct effects of age on allele frequency from indirect effects mediated by read depth, read length, and/or error rates (Supplementary Note S4b). We then filtered out SNPs in which more than half of the signal for temporal allele frequency change was driven by non-biological artefacts ($0.5 \geq \mathrm{F_j} < 1.0$). Furthermore, we implemented an additional mapping bias test, in which we compared the inferred present-day frequency of all SNPs based on (i) a *CLUES* model of the aDNA time-series data which was conditioned on the present-day frequency of each variant in the three EUR populations (see above); and (ii) a simpler *CLUES* model containing the aDNA time-series data alone. We then filtered out all SNPs in which the addition of the modern data both increased the significance of the selection test and resulted in an absolute present-

day frequency difference of > 0.1 between the two models. We also filtered out SNPs in which the observed pattern of genotypes in modern individuals was inconsistent with the marginal trees inferred from the surrounding haplotypes, as determined by `Relate` [2].

Genome-wide selection

We aggregated the results of all *CLUES* models (n=73,988) and converted the likelihood-ratio scores into p-values, using the chi-squared distribution with one degree of freedom. To identify genome-wide selection peaks, we used *harvester* (0.1) [15] with a relaxed input filter (`-inlimit 1e-3`) and a minimum peak height equal to the Bonferroni corrected p-value threshold (`-peak-limit 5.87`) for the number of tests in each grouping, after firstly applying a positional correction for the sparseness of the SNP ascertainment (because the published GWAS SNPs had already been pruned for linkage disequilibrium in their original studies). We then merged any adjacent peaks that were less than 100 kb apart in the same chromosome.

We also obtained a list of putatively selected SNPs and their p-values, inferred in an earlier aDNA study [16]. To ascertain novel sweep loci in our study, we used `harvester` to infer peaks from the published p-values. Additionally, we used *CLUES* to infer new p-values for all genome-wide significant SNPs (p<5e-8; n=381) from the earlier study [16], as not all SNPs found to be significant in that study were present in our GWAS/Control ascertainment.

Table S4a.1. Software and versions used in the allele trajectory pipeline

Software	Version	URL	Reference
bcftools	1.10.2	https://github.com/samtools/bcftools	17
bedtools	2.29.2	https://github.com/arq5x/bedtools2	18
biopython	1.76	https://github.com/biopython/biopython	19
clues	36cb7de	https://github.com/35ajstern/clues	1
conda	4.9.0	https://github.com/conda/conda	20
harvester	0.1	https://genomics.ut.ee/en/tools/manhattan-harvester	15
msprime		https://github.com/tskit-dev/msprime	8
numpy	1.17.0	https://github.com/numpy/numpy	21
pandas	1.0.4	https://github.com/pandas-dev/pandas	22

pysam	0.15.3	https://github.com/pysam-developers/pysam	[23]
python	3.6.7	https://www.python.org	[24]
r-base	3.6.1	https://www.r-project.org/	[25]
r-bedr	1.0.7	https://github.com/cran/bedr	[26]
r-dplyr	0.8.0.1	https://github.com/tidyverse/dplyr	[27]
r-ggplot2	3.1.1	https://github.com/tidyverse/ggplot2	[28]
r-ggrastr	0.2.1	https://github.com/VPetukhov/ggrastr	[29]
r-ggrepel	0.8.2	https://github.com/slowkow/ggrepel	[30]
r-ggridges	0.5.1	https://github.com/wilkelab/ggridges	[31]
r-stringr	1.4.0	https://github.com/tidyverse/stringr	[32]
relate	1.1.3	https://myersgroup.github.io/relate	[2]
scipy	1.4.1	https://github.com/scipy/scipy	[33]
snakemake	5.12.3	https://github.com/snakemake/snakemake	[3]

2662

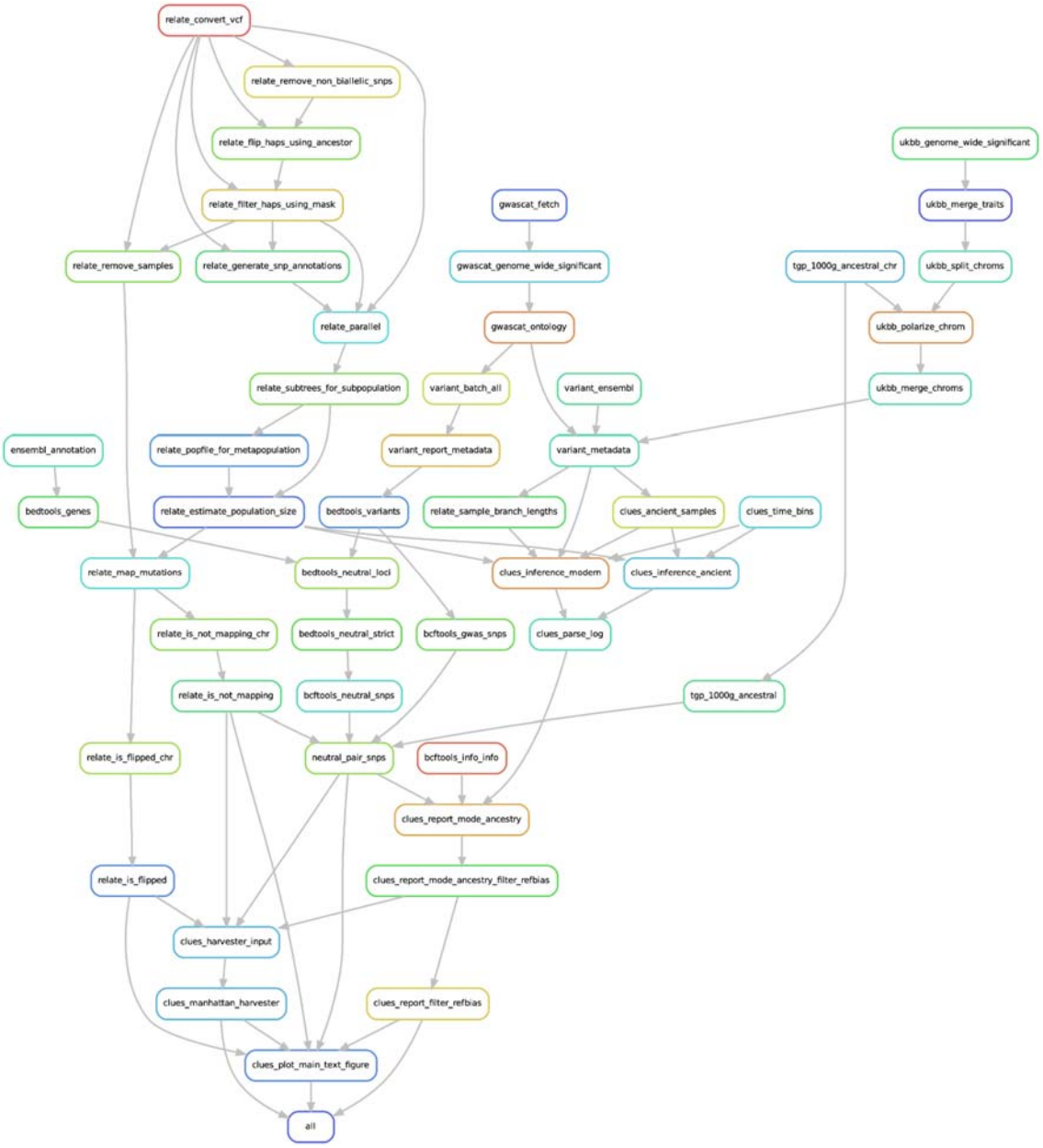


Figure S4a.1. Directed acyclic graph (DAG) of the allele frequency trajectory analysis pipeline

Results

Selection in 1000G EUR

CLUES analysis of all GWAS (n=35,592) and Control group SNPs (n=35,592) in the 1000G Phase 3 populations FIN, GBR, and TSI identified zero genome-wide significant SNPs (p<5e-8), and only one GWAS group SNP with a p-value below the Bonferroni corrected significance threshold (p<1.40e-6) (see Fig. S4a.2; Supplemental Table XV. Despite the

overall lack of genome-wide significant SNPs, the GWAS group was significantly enriched for evidence of selection when compared to the Control group (Wilcoxon signed-rank test, p-value<2.2e-16).

The only significant SNP was rs1375131 (chr2:135954797), an intron variant in *ZRANB3*, which is associated with mosquito bite size [34]. Non-significant SNPs within the surrounding peak region include the lactase persistence SNP rs4988235 (*MCM6*; p=9.3e-6), which has been widely reported as a target of strong selection in West Eurasians [35,36]. Among the GWAS SNPs, there was limited evidence for non-significant selection peaks, with partial overlap between these regions and those previously reported as genome-wide significant in other studies. Among the Control SNPs, no evidence of selection was identified.

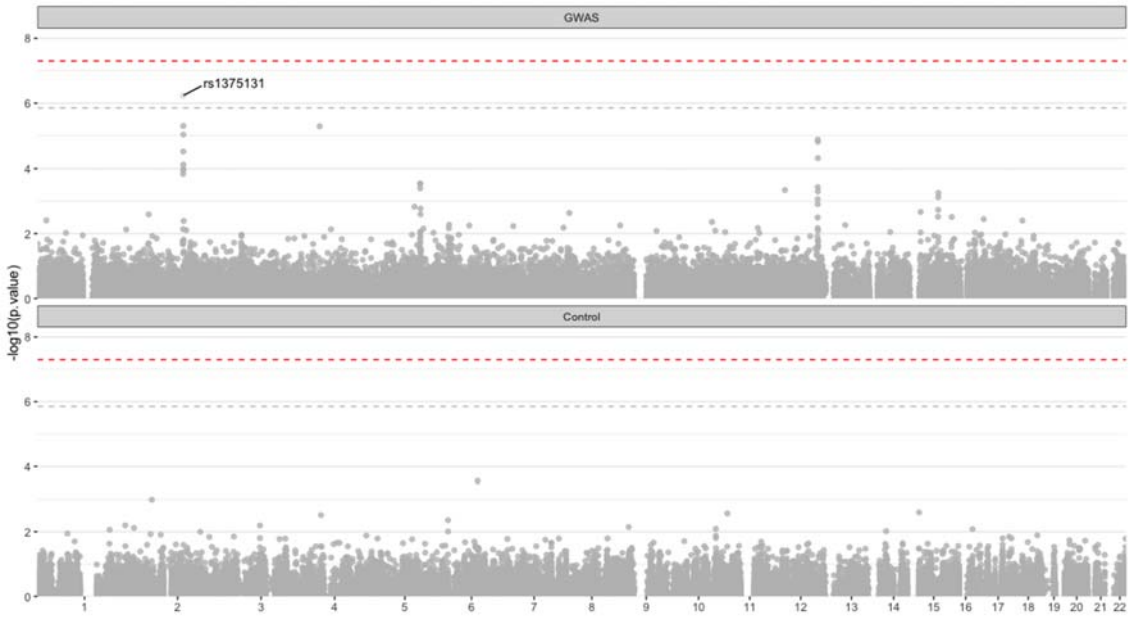


Figure S4a.2. Manhattan plot of the p-values from running CLUES on an ARG containing all samples in FIN, GBR, and TSI from 1000G Phase 3, for (a) GWAS SNPs from the GWAS Catalog; and (b) Control SNPs, frequency paired with the GWAS SNPs.

Selection in aDNA Time Series

CLUES analysis of all GWAS (n=33,323) and Control group SNPs (n=33,323) in the aDNA time-series dataset identified 127 genome-wide significant SNPs (p<5e-8); 127 in the GWAS group and 0 in the Control group. Using a Bonferroni corrected significance threshold, we detected 207 significant SNPs (p < 1.50e-6); 202 in the GWAS group (97.6%) and 5 in the Control group. Within the GWAS group, we identified 11 Bonferroni corrected significant selection peaks (see Fig. S4a.3; Supplementary Table XVI), of which 6 overlapped with

those previously characterised in [16]. No significant selection peaks were detected in the Control group.

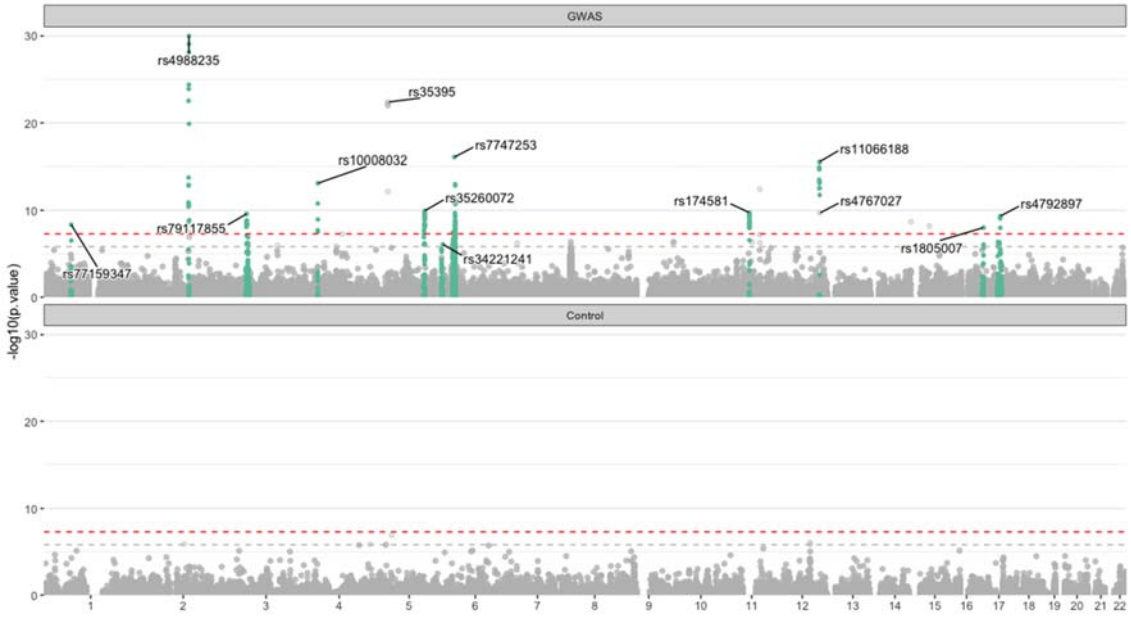


Figure S4a.3. Manhattan plot of the p-values from running *CLUES* on an aDNA time series from all West Eurasian samples in the imputed dataset, for (a) GWAS SNPs from the GWAS Catalog; and (b) Control SNPs, frequency paired with the GWAS SNPs.

Peak 1: LINC02797, AL583808.1

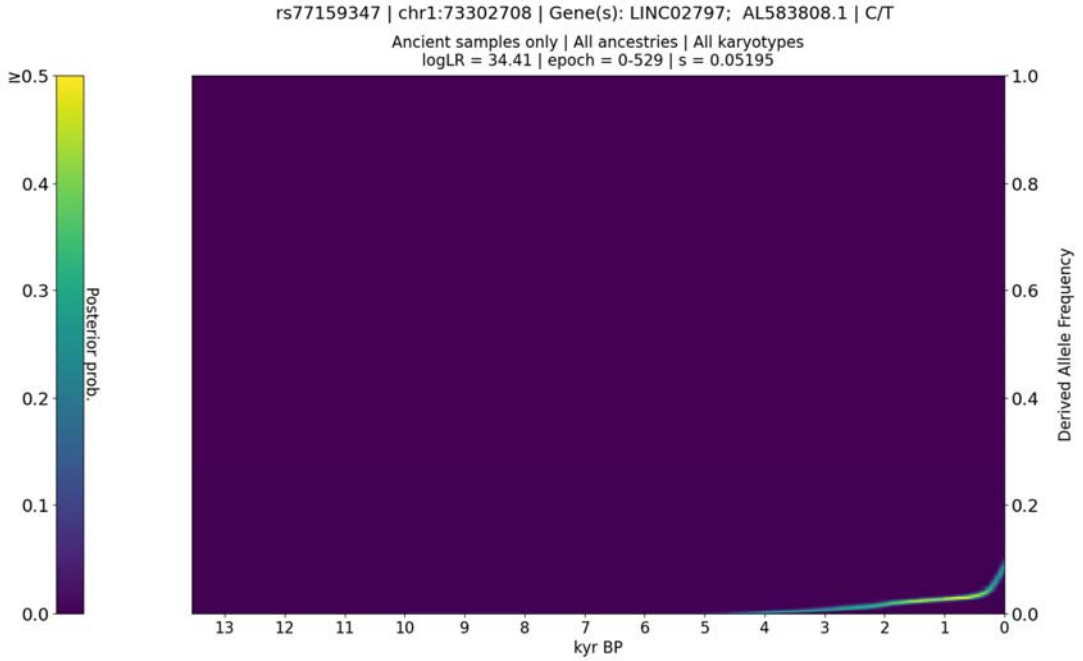


Figure S4a.4. *CLUES* plot of the aDNA time series analysis for all West Eurasian samples in the imputed dataset, showing the posterior probability of the derived allele frequency trajectory for rs77159347, the most significant SNP in the selection peak spanning chr1:72480859-73978570.

The first peak spanned the region chr1:72480859-73978570, with the most significant SNP being rs77159347 (*LINC02797, AL583808.1*; p=4.48e-09; s=0.052), associated with household income (MTAG) [37].

Peak 2: MCM6

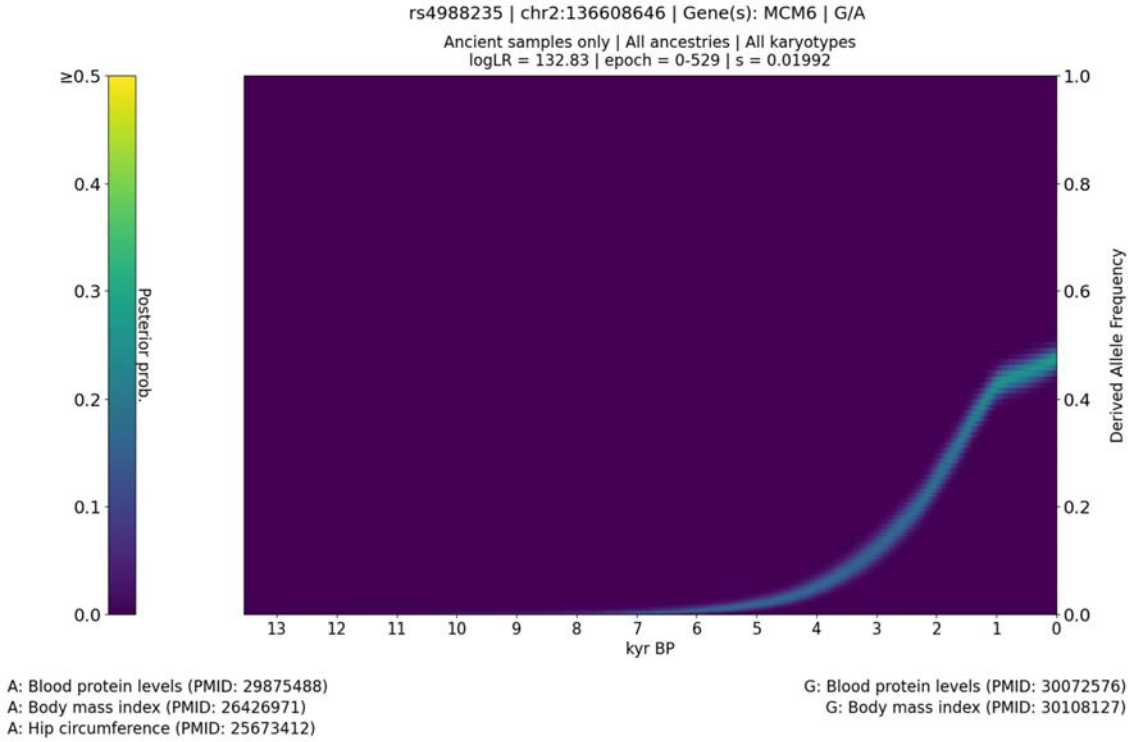
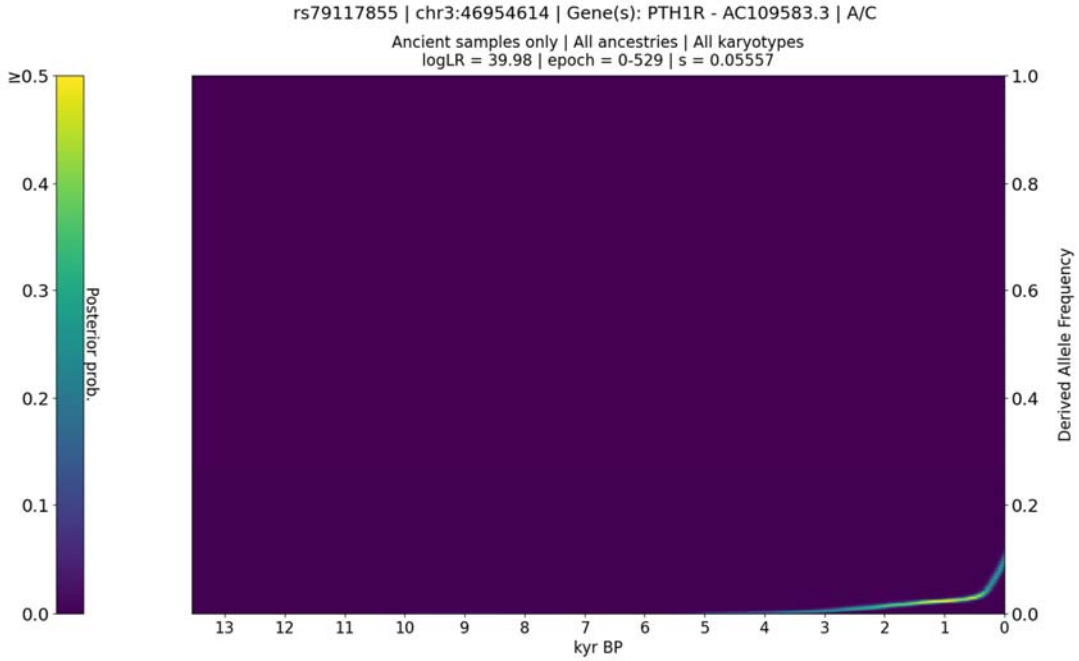


Figure S4a.5. *CLUES* plot of the aDNA time series analysis for all West Eurasian samples in the imputed dataset, showing the posterior probability of the derived allele frequency trajectory for rs4988235, the most significant SNP in the selection peak spanning chr2:135407409-137512400.

The second peak spanned the region chr2:135407409-137512400, with the most significant SNP being rs4988235 (*MCM6*; p=9.86e-31; s=0.0199), associated with lactase persistence; blood protein levels; body mass index; and hip circumference [38–42].

Peak 3: PTH1R - AC109583.3



?: Blood protein levels (PMID: 28915241)

Figure S4a.6. *CLUES* plot of the aDNA time series analysis for all West Eurasian samples in the imputed dataset, showing the posterior probability of the derived allele frequency trajectory for rs79117855, the most significant SNP in the selection peak spanning chr3:44526639-53850000.

The third peak spanned the region chr3:44526639-53850000, with the most significant SNP being rs79117855 (*PTH1R - AC109583.3*; p=2.57e-10; s=0.0556), associated with blood protein levels [43].

Peak 4: KLF3 - RNA5SP158

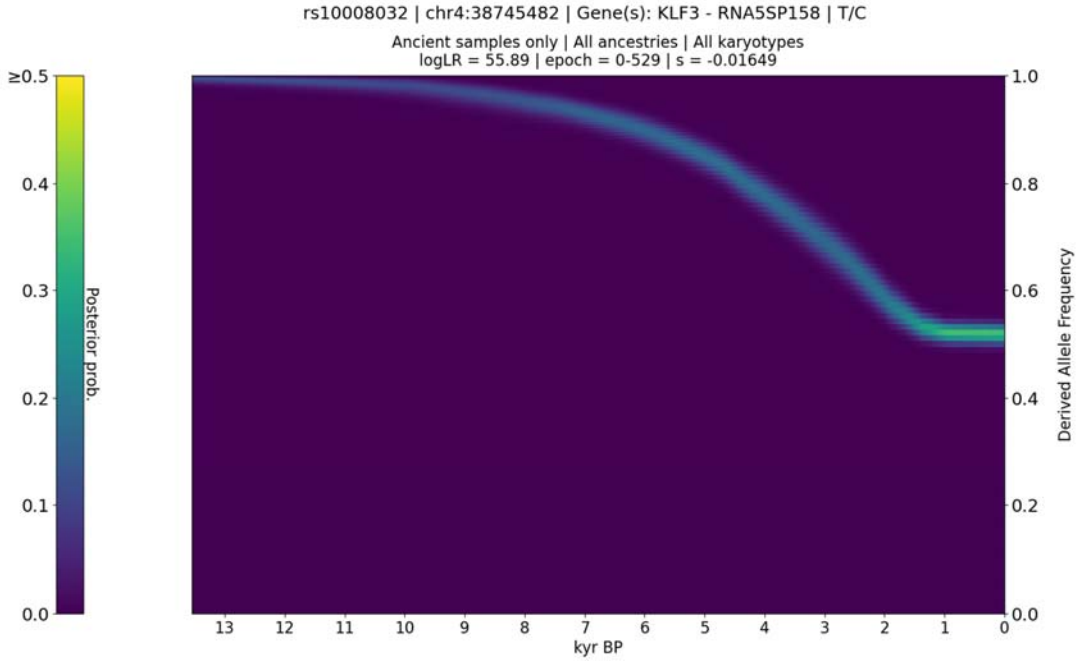
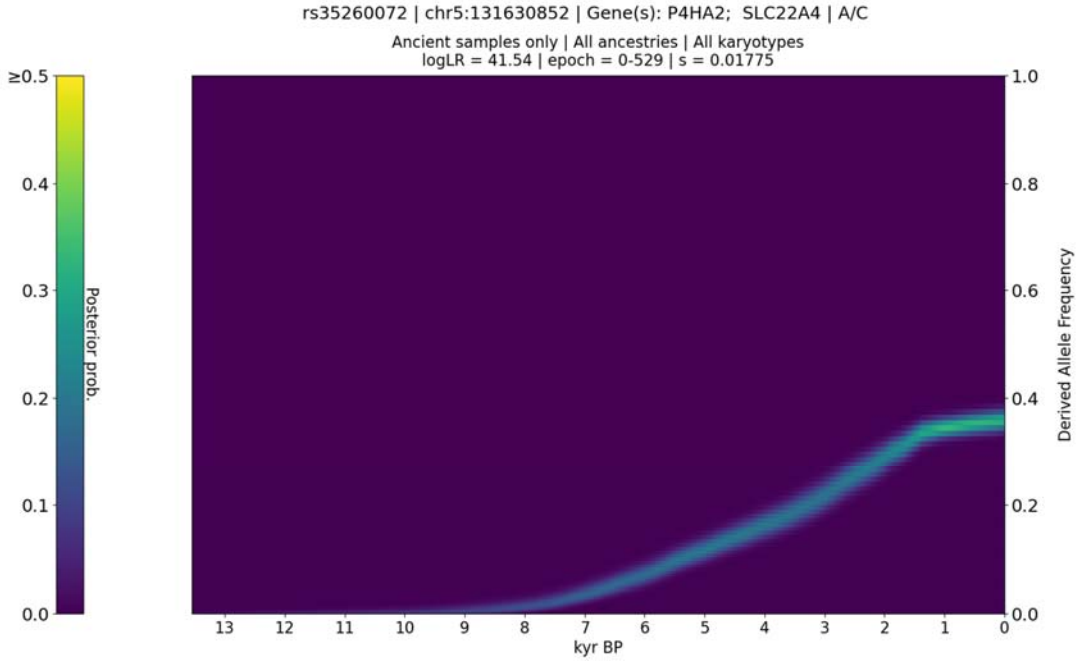


Figure S4a.7. *CLUES* plot of the aDNA time series analysis for all West Eurasian samples in the imputed dataset, showing the posterior probability of the derived allele frequency trajectory for rs10008032, the most significant SNP in the selection peak spanning chr4:38593259-38815500.

The fourth peak spanned the region chr4:38593259-38815500, with the most significant SNP being rs10008032 (*KLF3 - RNA5SP158*; p=7.66e-14; s=-0.0165), associated with allergic disease (asthma, hay fever or eczema) [44].

Peak 5: P4HA2, SLC22A4



C: Itch intensity from mosquito bite (PMID: 28199695)

Figure S4a.8. *CLUES* plot of the aDNA time series analysis for all West Eurasian samples in the imputed dataset, showing the posterior probability of the derived allele frequency trajectory for rs35260072, the most significant SNP in the selection peak spanning chr5:128016159-132349650.

The fifth peak spanned the region chr5:128016159-132349650, with the most significant SNP being rs35260072 (*P4HA2, SLC22A4*; p=1.15e-10; s=0.0177), associated with itch intensity from mosquito bite [34].

Peak 6: FLT4

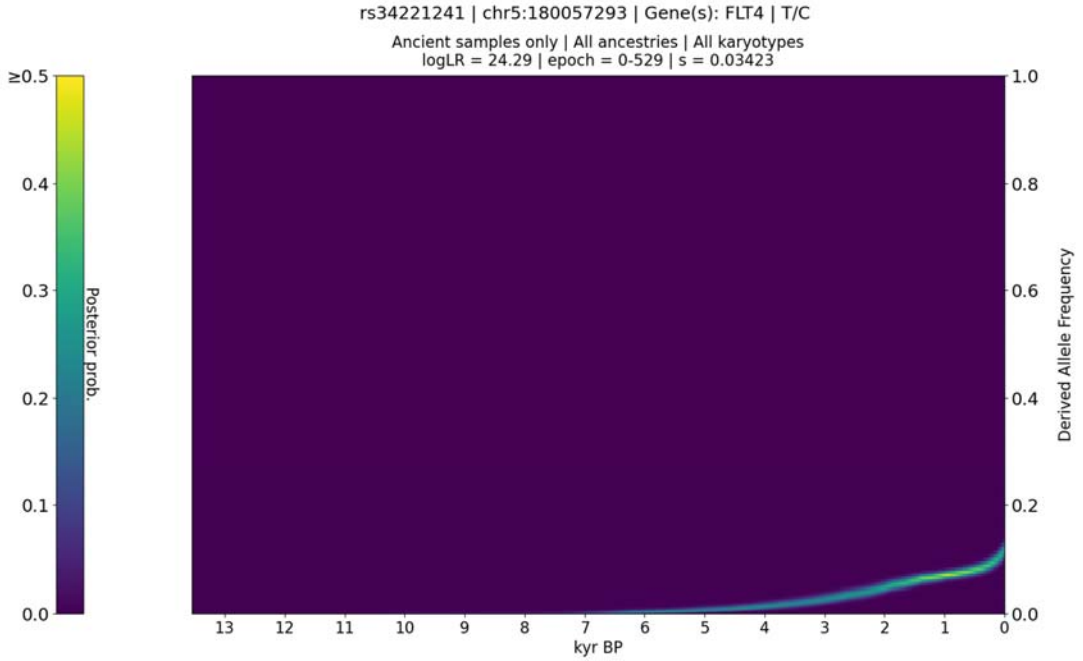
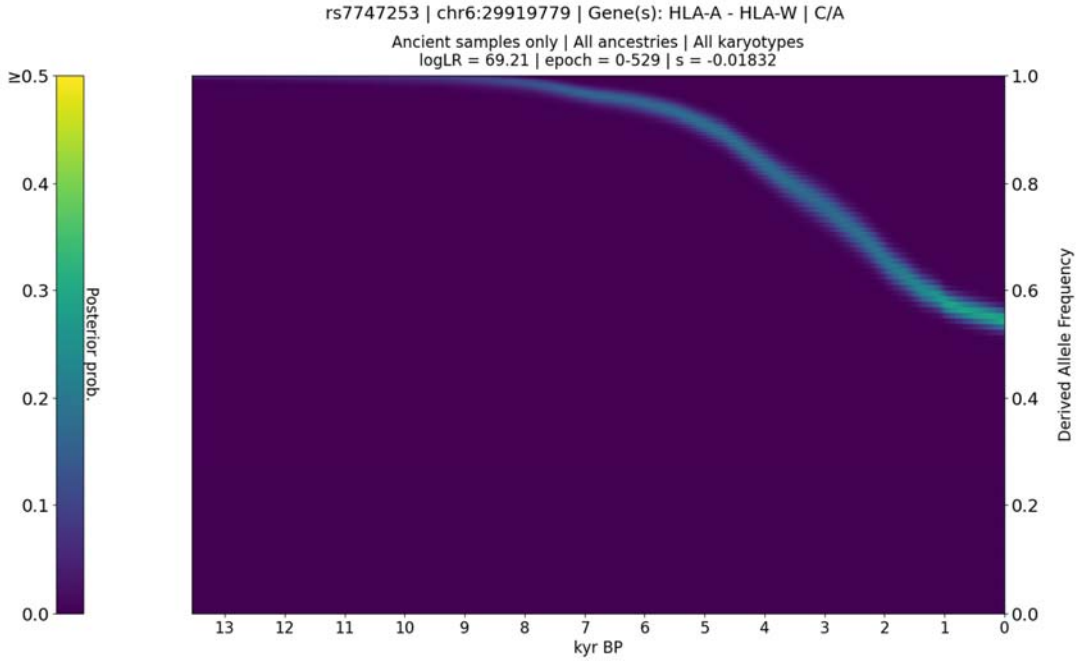


Figure S4a.9. *CLUES* plot of the aDNA time series analysis for all West Eurasian samples in the imputed dataset, showing the posterior probability of the derived allele frequency trajectory for rs34221241, the most significant SNP in the selection peak spanning chr5:176653519-180661980.

The sixth peak spanned the region chr5:176653519-180661980, with the most significant SNP being rs34221241 (*FLT4*; p=8.29e-07; s=0.0342), associated with blood protein levels [40].

Peak 7: HLA-A - HLA-W



C: Heel bone mineral density (PMID: 30598549)

Figure S4a.10. *CLUES* plot of the aDNA time series analysis for all West Eurasian samples in the imputed dataset, showing the posterior probability of the derived allele frequency trajectory for rs7747253, the most significant SNP in the selection peak spanning chr6:25236639-33535470.

The seventh peak spanned the region chr6:25236639-33535470, with the most significant SNP being rs7747253 (*HLA-A - HLA-W*; p=8.86e-17; s=-0.0183), associated with heel bone mineral density [45].

Peak 8: FADS2

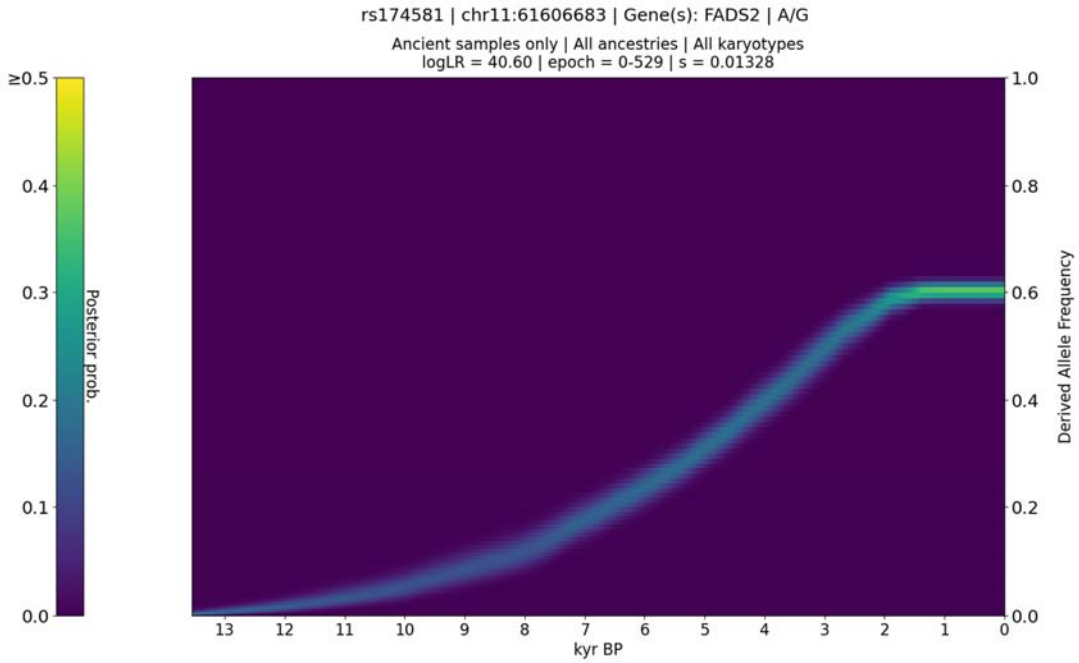


Figure S4a.11. *CLUES* plot of the aDNA time series analysis for all West Eurasian samples in the imputed dataset, showing the posterior probability of the derived allele frequency trajectory for rs174581, the most significant SNP in the selection peak spanning chr11:61543499-61706010.

The eighth peak spanned the region chr11:61543499-61706010, with the most significant SNP being rs174581 (*FADS2*; p=1.87e-10; s=0.0133), associated with male-pattern baldness; and serum metabolite ratios in chronic kidney disease [46,47].

Peak 9: HECTD4

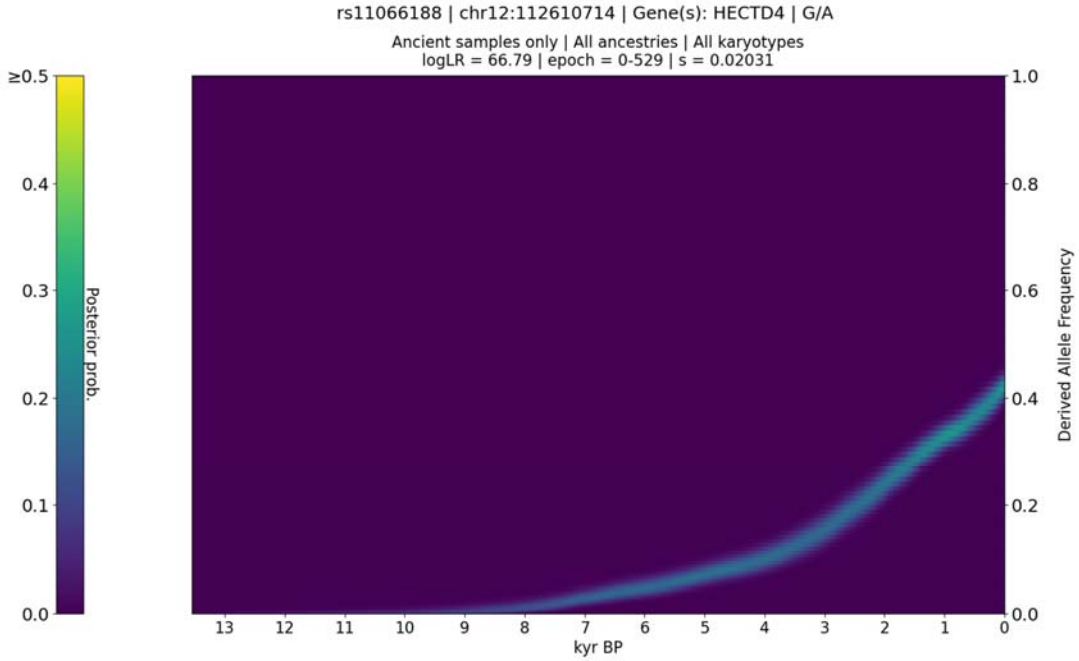


Figure S4a.12. *CLUES* plot of the aDNA time series analysis for all West Eurasian samples in the imputed dataset, showing the posterior probability of the derived allele frequency trajectory for rs11066188, the most significant SNP in the selection peak spanning chr12:111833789-113137570.

The ninth peak spanned the region chr12:111833789-113137570, with the most significant SNP being rs11066188 (*HECTD4*; p=3.02e-16; s=0.0203), associated with celiac disease and rheumatoid arthritis [48].

Peak 10: AC092143.1, MC1R

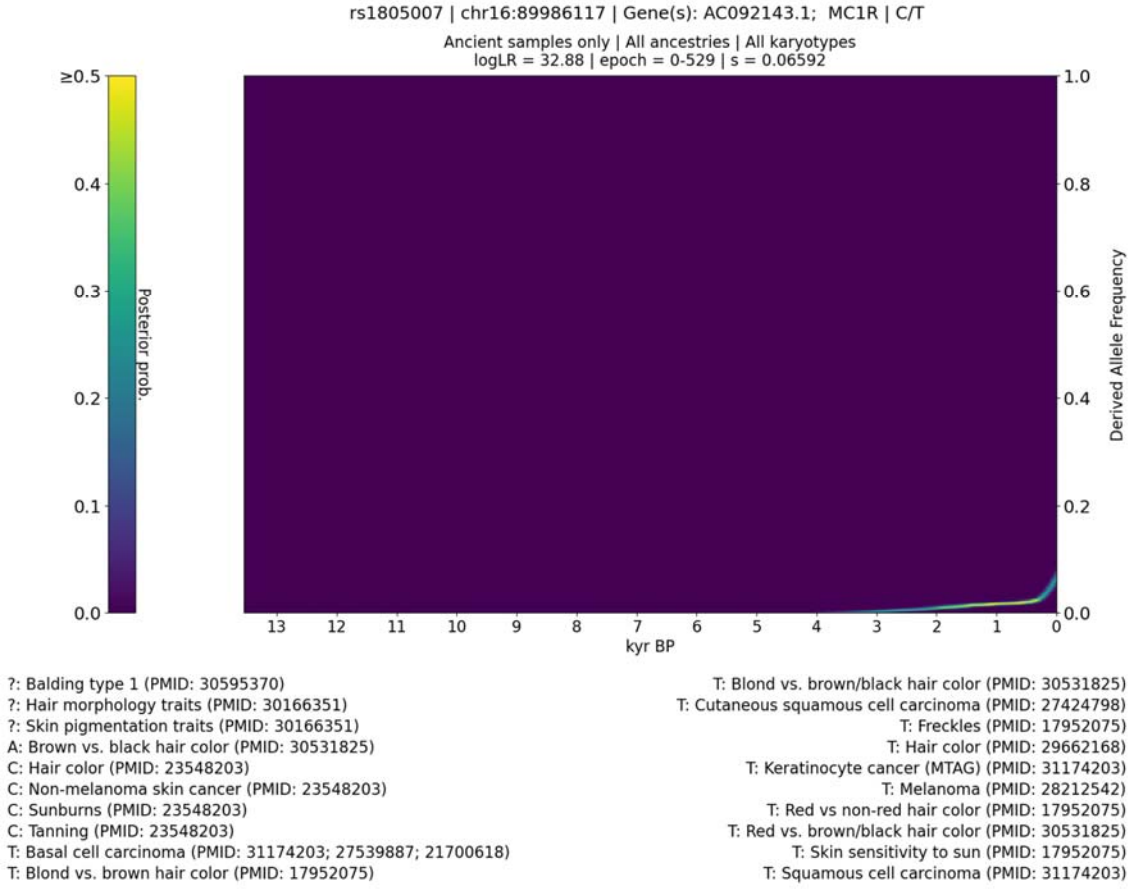
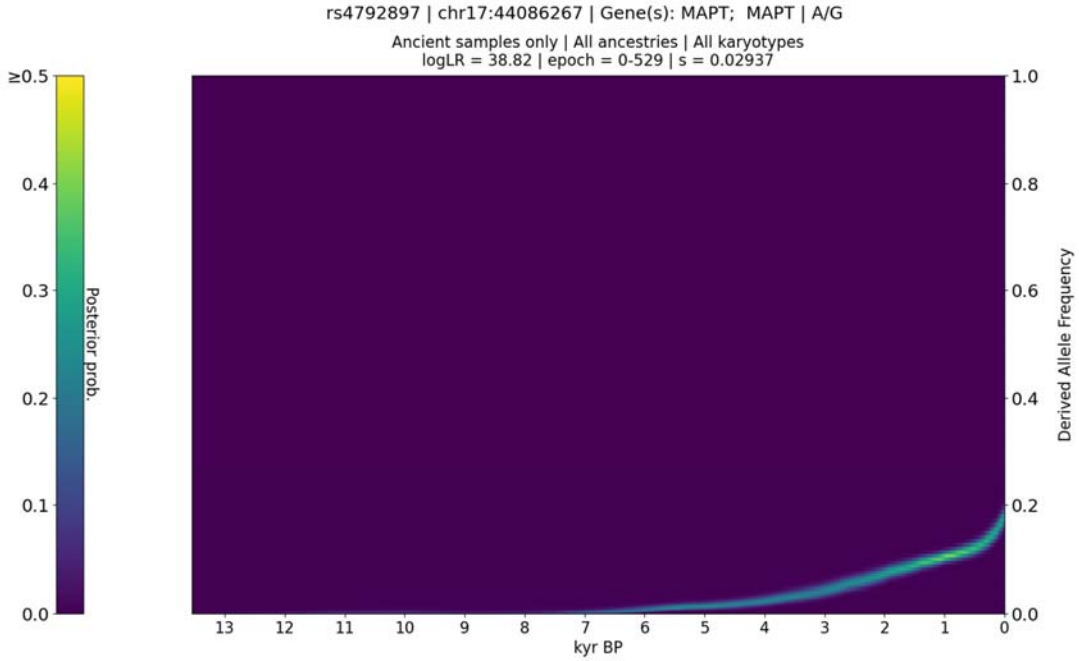


Figure S4a.13. *CLUES* plot of the aDNA time series analysis for all West Eurasian samples in the imputed dataset, showing the posterior probability of the derived allele frequency trajectory for rs1805007, the most significant SNP in the selection peak spanning chr16:86009759-90084560.

The tenth peak spanned the region chr16:86009759-90084560, with the most significant SNP being rs1805007 (*AC092143.1, MC1R*; p=9.83e-09; s=0.0659), associated with balding type 1; basal cell carcinoma; blond vs. brown hair colour; blond vs. brown/black hair colour; brown vs. black hair colour; cutaneous squamous cell carcinoma; freckles; hair colour; hair morphology traits; keratinocyte cancer (MTAG); melanoma; non-melanoma skin cancer; red vs non-red hair colour; red vs. brown/black hair colour; skin pigmentation traits; skin sensitivity to sun; squamous cell carcinoma; sunburns; and tanning [49–59].

Peak 11: MAPT



A: Snoring (PMID: 30804565)

Figure S4a.14. *CLUES* plot of the aDNA time series analysis for all West Eurasian samples in the imputed dataset, showing the posterior probability of the derived allele frequency trajectory for rs4792897, the most significant SNP in the selection peak spanning chr17:36615969-47588760.

The eleventh peak spanned the region chr17:36615969-47588760, with the most significant SNP being rs4792897 (*MAPT*; p=4.65e-10; s=0.0294), associated with snoring [60].

Selection in simulations with Ancestral Paintings

CLUES analysis of all frequency paired simulated SNPs in both the true paths and inferred paths simulations detected no genome-wide significant sweep loci, indicating that the false positive rate of our ancestry stratified selection analysis is low.

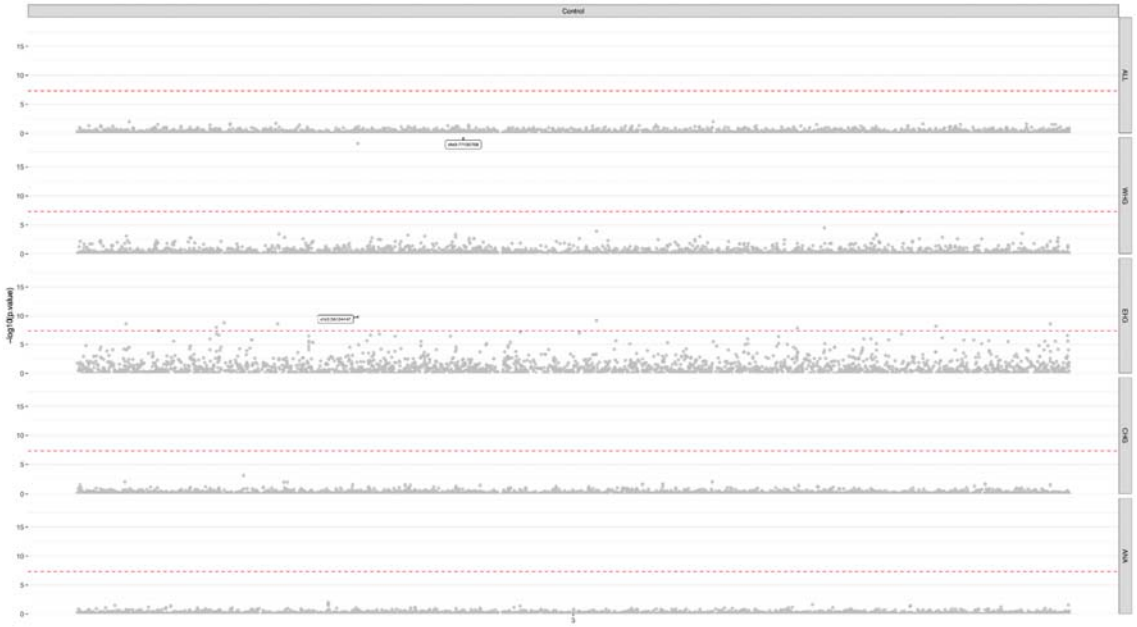


Figure S4a.15. Manhattan plot of the p-values from running *CLUES* on a neutral simulation of chr3, using the true simulated paths of each ancestry painting. The first row shows results for all ancient samples considered in aggregate, and each subsequent row shows the results conditional on one of the four specific ancestral paintings: ANA (Anatolian Farmers), CHG (Caucasus Hunter-gatherers), WHG (Western Hunter-gatherers) and EHG (Eastern Hunter-gatherers).

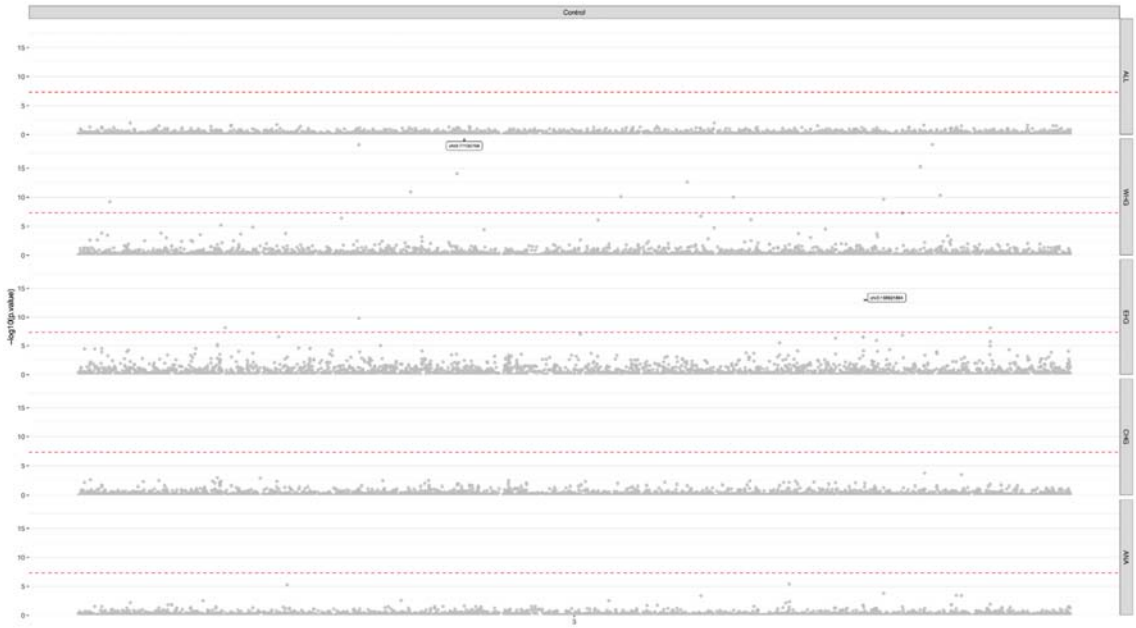


Figure S4a.16. Manhattan plot of the p-values from running *CLUES* on a neutral simulation of chr3, using the inferred paths of each ancestry painting. The first row shows results for all ancient samples considered in aggregate, and each subsequent row shows the results conditional on one of the four specific ancestral paintings: ANA (Anatolian Farmers), CHG (Caucasus Hunter-gatherers), WHG (Western Hunter-gatherers) and EHG (Eastern Hunter-gatherers).

Selection in aDNA with Ancestral Paintings

CLUES analysis of all GWAS (n=33,323) and Control group SNPs (n=33,323) in the aDNA with Ancestral Paintings dataset identified 409 genome-wide significant SNPs (p<5e-8); 346 in the GWAS group and 63 in the Control group. Using a Bonferroni corrected significance threshold, we detected 758 significant SNPs (p < 1.50e-06); 593 in the GWAS group (78.23%) and 165 in the Control group. Within the GWAS group, we identified 21 non-overlapping Bonferroni corrected significant selection peaks across all ancestries (see Fig. S4a.17; Supplementary Table XVI).

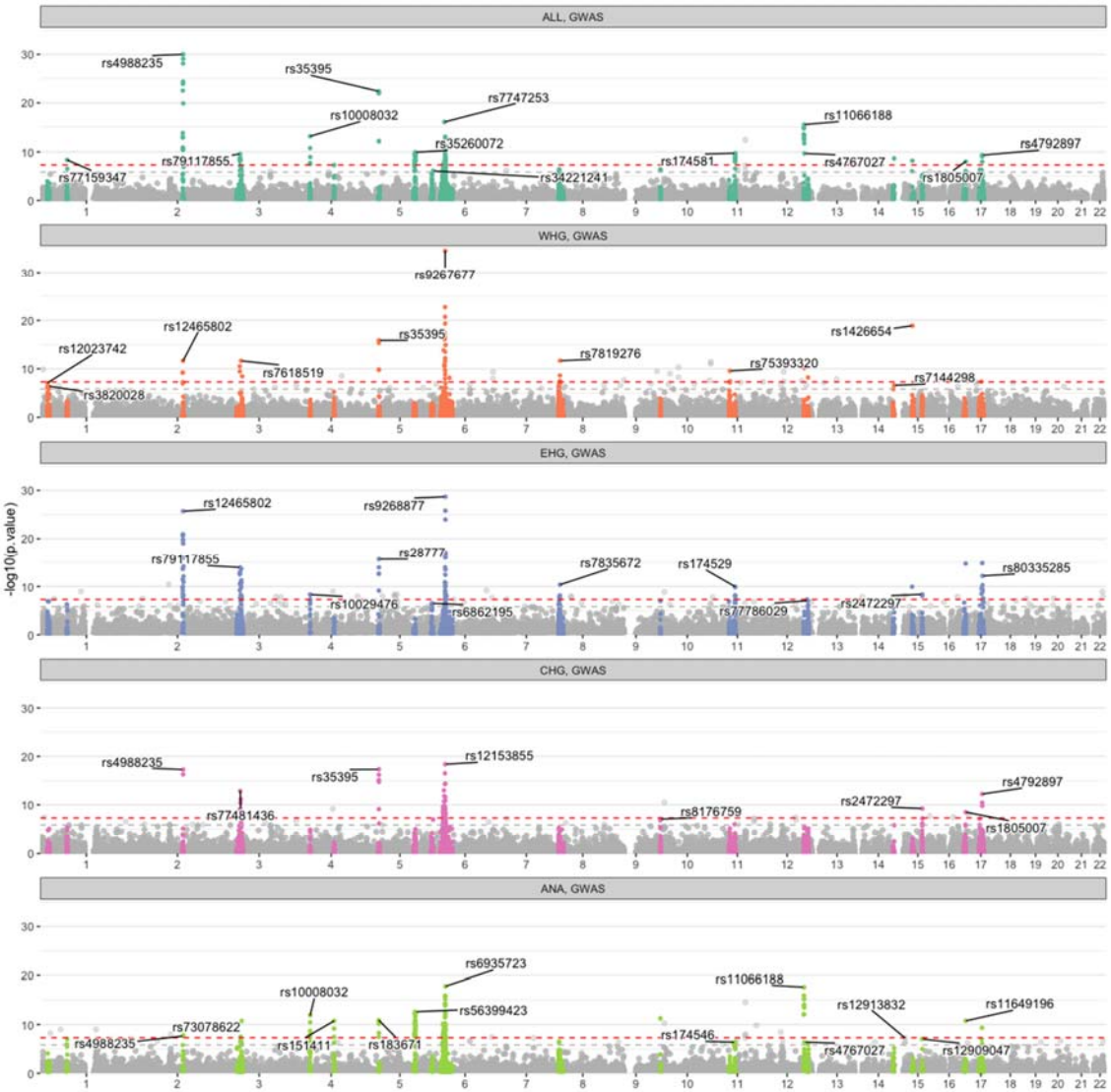


Figure S4a.17. Manhattan plot of the p-values from running *CLUES* on an aDNA time series conditioned on ancestry paintings from all West Eurasian samples in the imputed dataset for GWAS SNPs from the GWAS Catalog. The first row shows results for all ancient samples considered in aggregate, and each subsequent row shows the results conditional on one of

the four specific ancestral paintings: ANA (Anatolian Farmers), CHG (Caucasus Hunter-gatherers), WHG (Western Hunter-gatherers) and EHG (Eastern Hunter-gatherers).

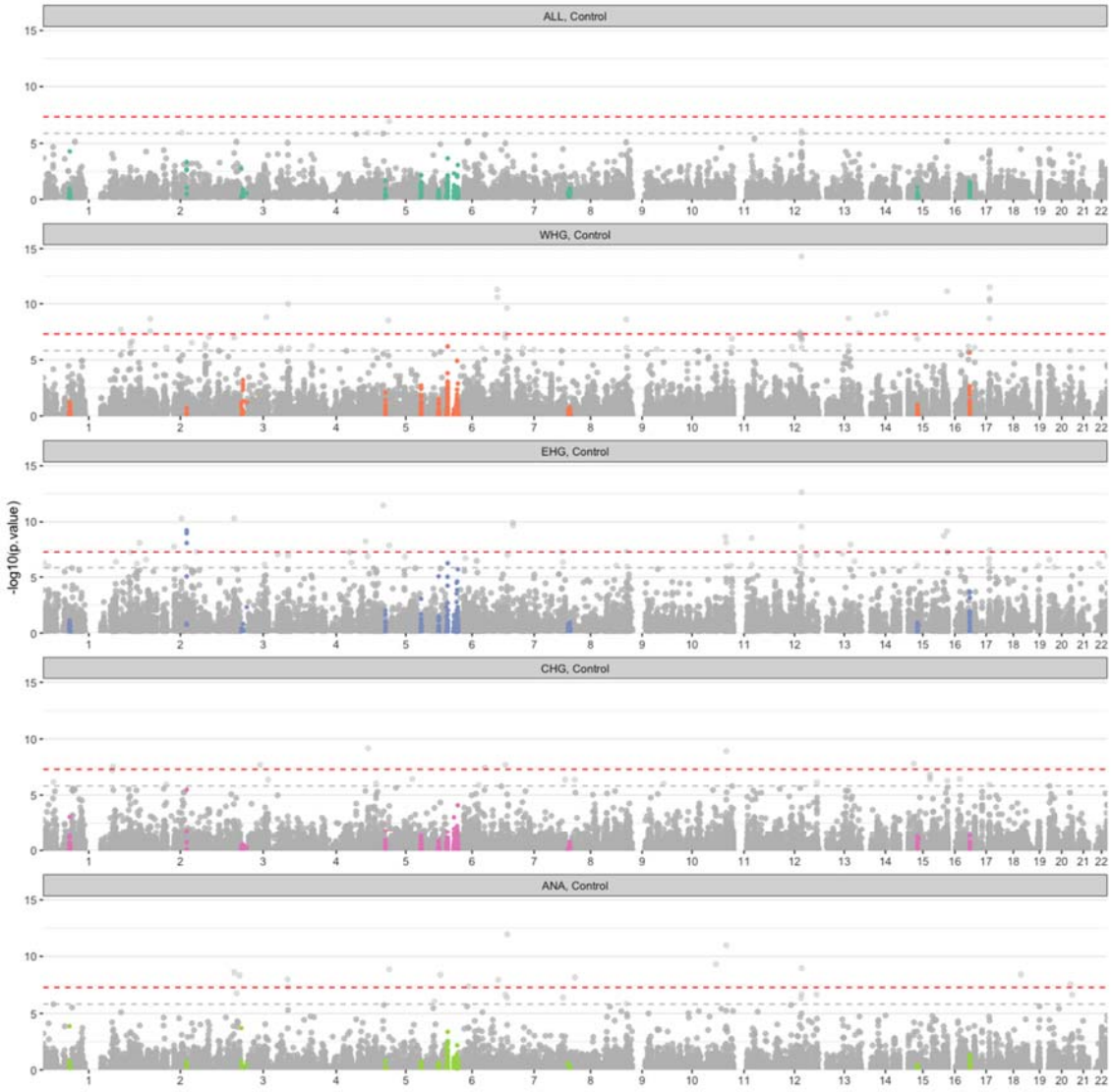


Figure S4a.18. Manhattan plot of the p-values from running *CLUES* on an aDNA time series conditioned on ancestry paintings from all West Eurasian samples in the imputed dataset for Control SNPs, frequency paired with the GWAS SNPs. The first row shows results for all ancient samples considered in aggregate, and each subsequent row shows the results conditional on one of the four specific ancestral paintings: ANA (Anatolian Farmers), CHG (Caucasus Hunter-gatherers), WHG (Western Hunter-gatherers) and EHG (Eastern Hunter-gatherers).

Peak 1: AL591122.1

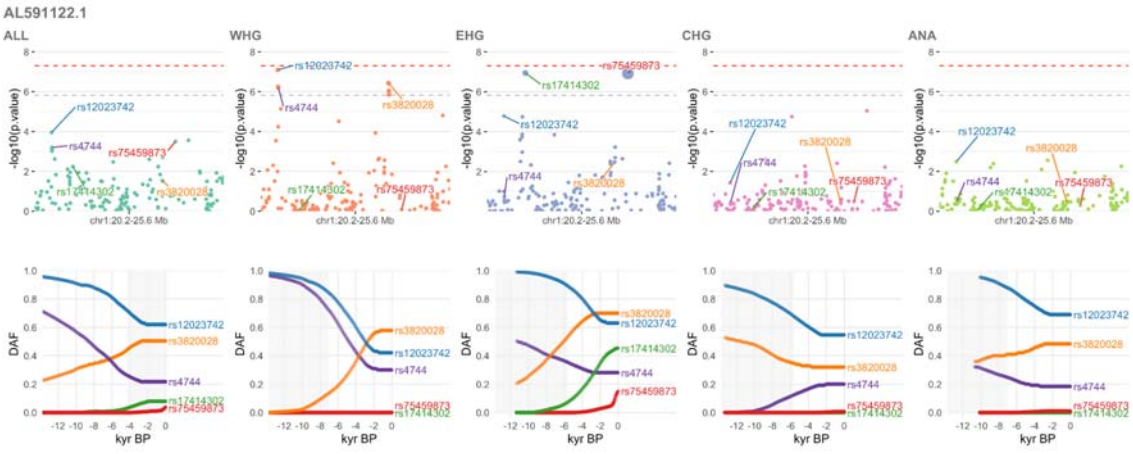


Figure S4a.19. Selection at the *AL591122.1* locus, spanning chr1:20236729-25570080. Results for the pan-ancestry analysis (ALL) plus the four marginal ancestries: Western hunter-gatherers (WHG), Eastern hunter-gatherers (EHG), Caucasus hunter-gatherers (CHG) and Anatolian farmers (ANA). Row one shows zoomed Manhattan plots of the p-values for each ancestry (significant SNPs sized by their selection coefficients), and row two shows allele trajectories for the top SNPs across all ancestries.

This peak spanned the region chr1:20236729-25570080, with significant SNPs including:

- rs12023742 (*PLA2G2E - RN7SL304P*; WHG; p=8.10e-08; s=-0.0136), associated with group IIA secretory phospholipase A2 levels in individuals with elevated hsCRP [61].
- rs17414302 (*PINK1-AS,PINK1*; EHG; p=1.16e-07; s=0.0185), associated with Household income (MTAG); Intelligence (MTAG) [37,62].
- rs75459873 (*MIR378F - H3P1*; EHG; p=1.21e-07; s=0.0437), associated with Psychotic experience (distressing) [63].
- rs3820028 (*E2F2*; WHG; p=3.75e-07; s=0.0144), associated with Heel bone mineral density [58].
- rs4744 (*PLA2G2A*; WHG; p=5.67e-07; s=-0.0128), associated with Blood protein levels [41].

Peak 2: NEGR1

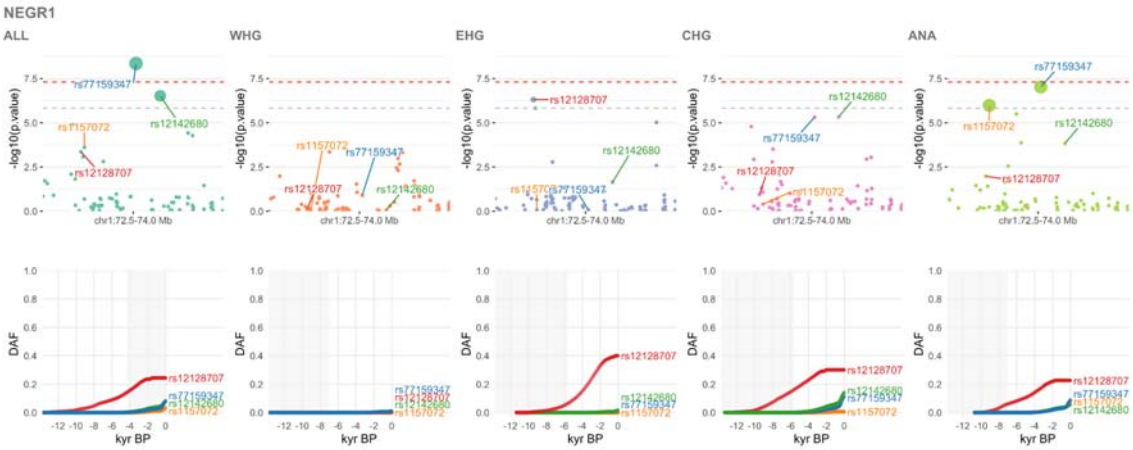


Figure S4a.20. Selection at the *NEGR1* locus, spanning chr1:72480859-73978570. Results for the pan-ancestry analysis (ALL) plus the four marginal ancestries: Western hunter-gatherers (WHG), Eastern hunter-gatherers (EHG), Caucasus hunter-gatherers (CHG) and Anatolian farmers (ANA). Row one shows zoomed Manhattan plots of the p-values for each ancestry (significant SNPs sized by their selection coefficients), and row two shows allele trajectories for the top SNPs across all ancestries.

This peak spanned the region chr1:72480859-73978570, with significant SNPs including:

- rs77159347 (*LINC02797,AL583808.1*; ALL; p=4.48e-09; s=0.052), associated with Household income (MTAG) [37].
- rs12142680 (*KRT8P21 - RN7SKP19*; ALL; p=2.99e-07; s=0.0437), associated with Educational attainment (years of education) [64].
- rs12128707 (*NEGR1*; EHG; p=4.93e-07; s=0.0182), associated with Cognitive performance; Cognitive performance (MTAG); Intelligence; Intelligence (MTAG) [62,65,66].
- rs1157072 (*NEGR1*; ANA; p=1.04e-06; s=0.0498), associated with Household income (MTAG); Intelligence (MTAG) [37,62].

Peak 3: MCM6

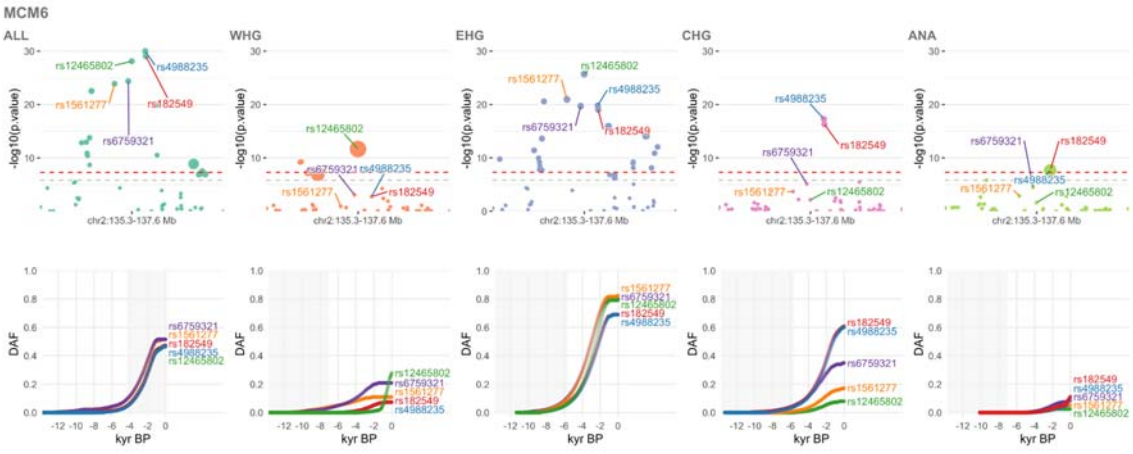


Figure S4a.21. Selection at the *MCM6* locus, spanning chr2:135300859-137564020. Results for the pan-ancestry analysis (ALL) plus the four marginal ancestries: Western hunter-gatherers (WHG), Eastern hunter-gatherers (EHG), Caucasus hunter-gatherers (CHG) and Anatolian farmers (ANA). Row one shows zoomed Manhattan plots of the p-values for each ancestry (significant SNPs sized by their selection coefficients), and row two shows allele trajectories for the top SNPs across all ancestries.

This peak spanned the region chr2:135300859-137564020, with significant SNPs including:

- rs4988235 (*MCM6*; ALL; p=9.86e-31; s=0.0199), associated with Lactase persistence; Blood protein levels; Body mass index; Hip circumference [38–42].
- rs182549 (*MCM6*; ALL; p=8.44e-30; s=0.0198), associated with 1,5-anhydroglucitol levels [67].
- rs12465802 (*R3HDM1*; ALL; p=7.71e-29; s=0.0196), associated with Blood protein levels; Mosquito bite size; Urinary metabolite levels in chronic kidney disease [34,40,68].
- rs6759321 (*R3HDM1*; ALL; p=4.07e-25; s=0.0188), associated with Hand grip strength [69].
- rs1561277 (*ZRANB3*; ALL; p=1.24e-24; s=0.0188), associated with Hip circumference [38].

Peak 4: CACNA2D2

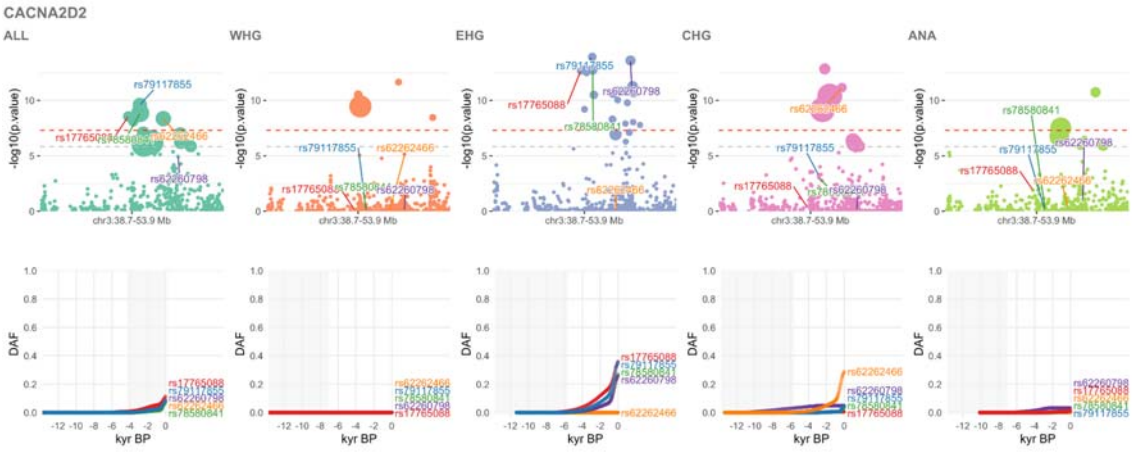


Figure S4a.22. Selection at the *CACNA2D2* locus, spanning chr3:38723219-53850000. Results for the pan-ancestry analysis (ALL) plus the four marginal ancestries: Western hunter-gatherers (WHG), Eastern hunter-gatherers (EHG), Caucasus hunter-gatherers (CHG) and Anatolian farmers (ANA). Row one shows zoomed Manhattan plots of the p-values for each ancestry (significant SNPs sized by their selection coefficients), and row two shows allele trajectories for the top SNPs across all ancestries.

This peak spanned the region chr3:38723219-53850000, with significant SNPs including:

- rs79117855 (*PTH1R - AC109583.3*; EHG; p=1.18e-14; s=0.0305), associated with Blood protein levels [43].
- rs62260798 (*GNAI2*; EHG; p=2.52e-14; s=0.0373), associated with Morning person [70].
- rs78580841 (*CCDC12*; EHG; p=1.68e-13; s=0.037), associated with Chronotype [70].
- rs17765088 (*CCR9,LZTFL1*; EHG; p=2.22e-13; s=0.0257), associated with Macrophage inflammatory protein 1b levels [71].
- rs62262466 (*ARIH2*; CHG; p=6.98e-12; s=0.0351), associated with Morning person [70].

Peak 5: *KLF3-AS1*

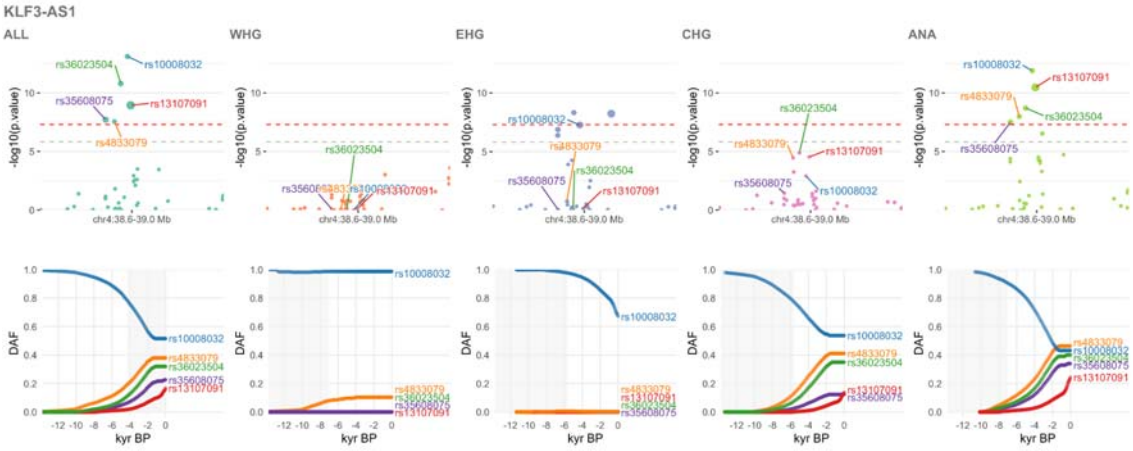


Figure S4a.23. Selection at the *KLF3-AS1* locus, spanning chr4:38593259-38966080. Results for the pan-ancestry analysis (ALL) plus the four marginal ancestries: Western hunter-gatherers (WHG), Eastern hunter-gatherers (EHG), Caucasus hunter-gatherers (CHG) and Anatolian farmers (ANA). Row one shows zoomed Manhattan plots of the p-values for each ancestry (significant SNPs sized by their selection coefficients), and row two shows allele trajectories for the top SNPs across all ancestries.

This peak spanned the region chr4:38593259-38966080, with significant SNPs including:

- rs10008032 (*KLF3 - RNA5SP158*; ALL; p=7.66e-14; s=-0.0165), associated with Allergic disease (asthma, hay fever or eczema) [44].
- rs36023504 (*KLF3*; ALL; p=1.62e-11; s=0.0179), associated with Body mass index; Red cell distribution width [58].
- rs13107091 (*RNA5SP158 - TLR10*; ANA; p=3.41e-11; s=0.0281), associated with Atopic asthma [72].
- rs4833079 (*KLF3-AS1*; ANA; p=1.02e-08; s=0.0158), associated with Body mass index [42].
- rs35608075 (*LINC02278 - KLF3-AS1*; ALL; p=1.95e-08; s=0.0208), associated with Male-pattern baldness [73].

Peak 6: BANK1

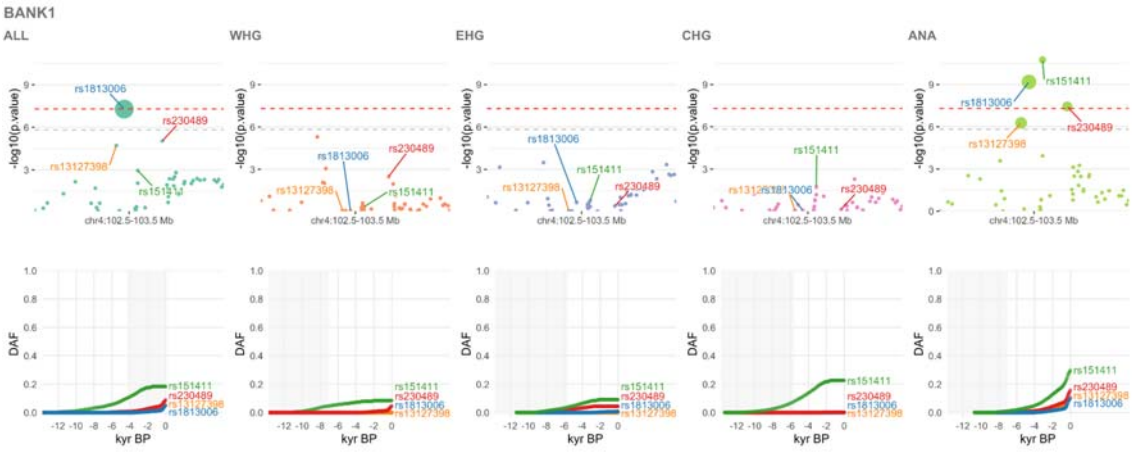


Figure S4a.24. Selection at the *BANK1* locus, spanning chr4:102507109-103525350. Results for the pan-ancestry analysis (ALL) plus the four marginal ancestries: Western hunter-gatherers (WHG), Eastern hunter-gatherers (EHG), Caucasus hunter-gatherers (CHG) and Anatolian farmers (ANA). Row one shows zoomed Manhattan plots of the p-values for each ancestry (significant SNPs sized by their selection coefficients), and row two shows allele trajectories for the top SNPs across all ancestries.

This peak spanned the region chr4:102507109-103525350, with significant SNPs including:

- rs151411 (*BANK1 - SLC39A8*; ANA; p=1.80e-11; s=0.0247), associated with General cognitive ability [74].
- rs1813006 (*BANK1 - SLC39A8*; ANA; p=6.60e-10; s=0.0563), associated with Intelligence (MTAG) [62].
- rs230489 (*AF213884.2 - AF213884.3*; ANA; p=3.63e-08; s=0.036), associated with Brain region volumes; General cognitive ability; Intelligence (MTAG) [62,74,75].
- rs13127398 (*BANK1,AP002075.1*; ANA; p=5.36e-07; s=0.0439), associated with Cardiovascular disease [58].

Peak 7: SLC45A2

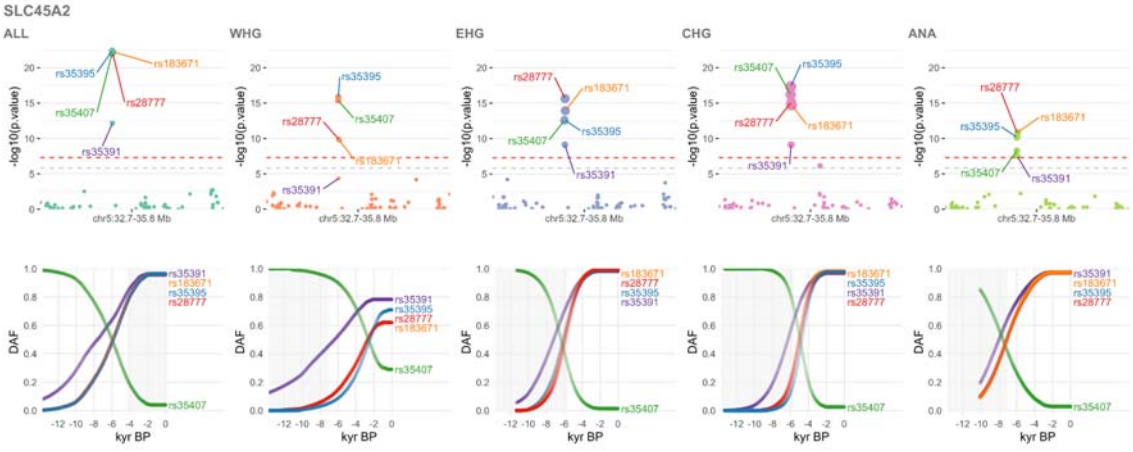


Figure S4a.25. Selection at the *SLC45A2* locus, spanning chr5:32710489-35848560. Results for the pan-ancestry analysis (ALL) plus the four marginal ancestries: Western hunter-gatherers (WHG), Eastern hunter-gatherers (EHG), Caucasus hunter-gatherers (CHG) and Anatolian farmers (ANA). Row one shows zoomed Manhattan plots of the p-values for each ancestry (significant SNPs sized by their selection coefficients), and row two shows allele trajectories for the top SNPs across all ancestries.

This peak spanned the region chr5:32710489-35848560, with significant SNPs including:

- rs35395 (*SLC45A2*; ALL; p=4.13e-23; s=0.022), associated with Skin pigmentation [76].
- rs183671 (*SLC45A2*; ALL; p=5.51e-23; s=0.0221), associated with Hair color; Skin colour saturation; Skin pigmentation traits [56,77,78].
- rs28777 (*SLC45A2*; ALL; p=8.48e-23; s=0.0217), associated with Black vs. blond hair color; Black vs. red hair color; Skin, hair and eye pigmentation (multivariate analysis) [79,80].
- rs35407 (*SLC45A2*; ALL; p=1.06e-22; s=-0.0221), associated with Basal cell carcinoma; Cutaneous squamous cell carcinoma; Keratinocyte cancer (MTAG); Melanoma; Squamous cell carcinoma [52–54,59].
- rs35391 (*SLC45A2*; ALL; p=6.87e-13; s=0.0161), associated with Tanning [81].

Peak 8: SLC22A4

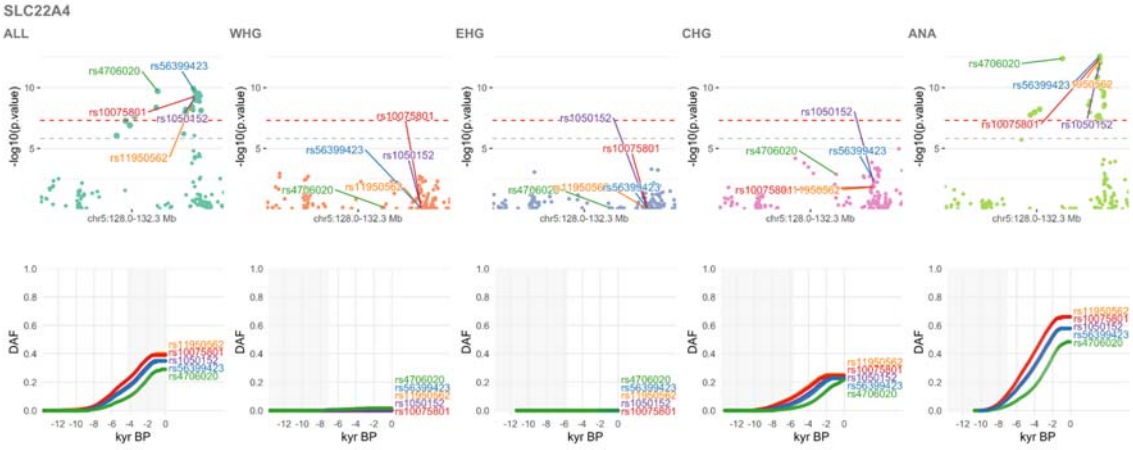


Figure S4a.26. Selection at the *SLC22A4* locus, spanning chr5:128016159-132349650. Results for the pan-ancestry analysis (ALL) plus the four marginal ancestries: Western hunter-gatherers (WHG), Eastern hunter-gatherers (EHG), Caucasus hunter-gatherers (CHG) and Anatolian farmers (ANA). Row one shows zoomed Manhattan plots of the p-values for each ancestry (significant SNPs sized by their selection coefficients), and row two shows allele trajectories for the top SNPs across all ancestries.

This peak spanned the region chr5:128016159-132349650, with significant SNPs including:

- rs56399423 (*MIR3936HG,SLC22A4*; ANA; p=2.48e-13; s=0.017), associated with Inflammatory bowel disease [82].
- rs10075801 (*SLC22A4,MIR3936HG*; ANA; p=3.75e-13; s=0.0164), associated with Granulocyte count; Myeloid white cell count; Neutrophil count; Sum basophil neutrophil counts; Sum neutrophil eosinophil counts; White blood cell count [83].
- rs4706020 (*CDC42SE2*; ANA; p=4.03e-13; s=0.0184), associated with Itch intensity from mosquito bite; Itch intensity from mosquito bite adjusted by bite size [34].
- rs11950562 (*MIR3936HG,SLC22A4*; ANA; p=5.38e-13; s=0.0163), associated with Blood metabolite levels; Mean platelet volume [83,84].
- rs1050152 (*SLC22A4,MIR3936HG*; ANA; p=8.09e-13; s=0.0167), associated with Nasal polyps [85].

Peak 9: GRK6

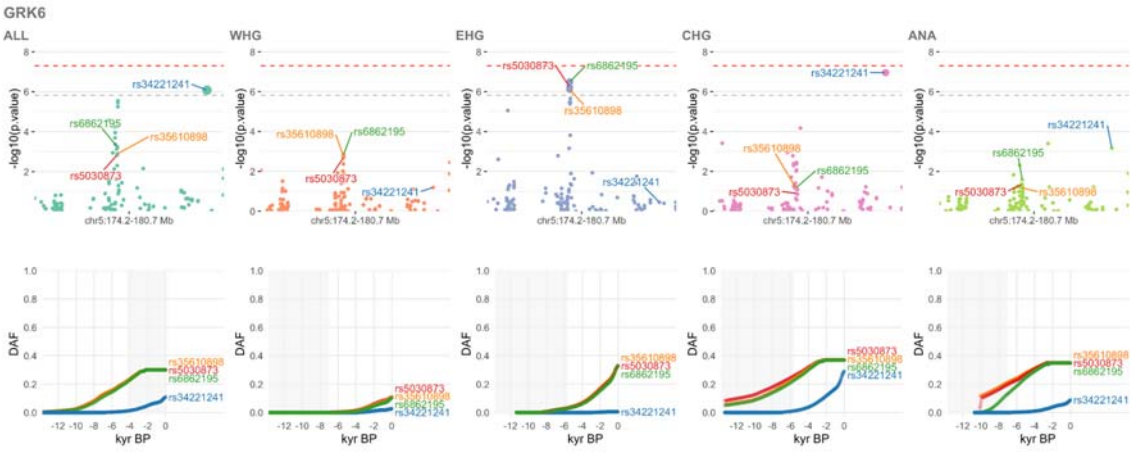


Figure S4a.27. Selection at the *GRK6* locus, spanning chr5:174156169-180661980. Results for the pan-ancestry analysis (ALL) plus the four marginal ancestries: Western hunter-gatherers (WHG), Eastern hunter-gatherers (EHG), Caucasus hunter-gatherers (CHG) and Anatolian farmers (ANA). Row one shows zoomed Manhattan plots of the p-values for each ancestry (significant SNPs sized by their selection coefficients), and row two shows allele trajectories for the top SNPs across all ancestries.

This peak spanned the region chr5:174156169-180661980, with significant SNPs including:

- rs34221241 (*FLT4*; CHG; p=1.11e-07; s=0.0253), associated with Blood protein levels [40].
- rs6862195 (*SLC34A1*; EHG; p=3.18e-07; s=0.0232), associated with Estimated glomerular filtration rate [86].
- rs5030873 (*SLC34A1*; EHG; p=5.85e-07; s=0.0228), associated with Creatinine levels [87].
- rs35610898 (*SLC34A1*; EHG; p=7.84e-07; s=0.0224), associated with Estimated glomerular filtration rate [86].

Peak 10: HLA

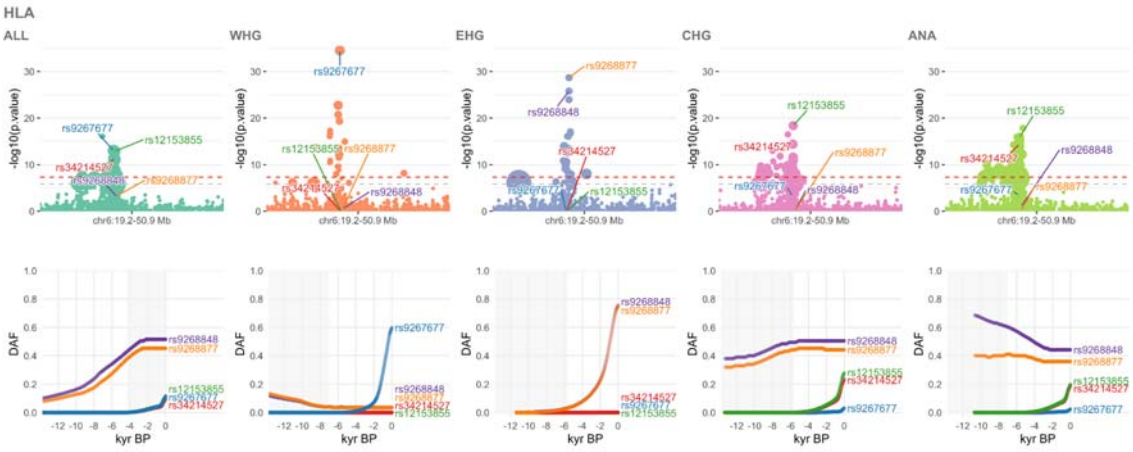


Figure S4a.28. Selection at the *HLA* locus, spanning chr6:19191049-50921600. Results for the pan-ancestry analysis (ALL) plus the four marginal ancestries: Western hunter-gatherers (WHG), Eastern hunter-gatherers (EHG), Caucasus hunter-gatherers (CHG) and Anatolian farmers (ANA). Row one shows zoomed Manhattan plots of the p-values for each ancestry (significant SNPs sized by their selection coefficients), and row two shows allele trajectories for the top SNPs across all ancestries.

This peak spanned the region chr6:19191049-50921600, with significant SNPs including:

- rs9267677 (*C2*; WHG; p=3.03e-35; s=0.0365), associated with Cognitive performance (MTAG); Educational attainment (MTAG); Educational attainment (years of education); Highest math class taken (MTAG) [66].
- rs9268877 (*HLA-DRB9*; EHG; p=2.07e-29; s=0.0249), associated with Poor prognosis in Crohn's disease; Ulcerative colitis [88–90].
- rs9268848 (*HLA-DRB9*; EHG; p=1.56e-26; s=0.0247), associated with Nonatopic asthma; Urate levels [72,91].
- rs12153855 (*TNXB,AL662884.2*; CHG; p=4.19e-19; s=0.0334), associated with Age-related macular degeneration; Atopic dermatitis [92,93].
- rs34214527 (*TNXB*; CHG; p=4.09e-15; s=0.0348), associated with Highest math class taken [66].

Peak 11: MSRA

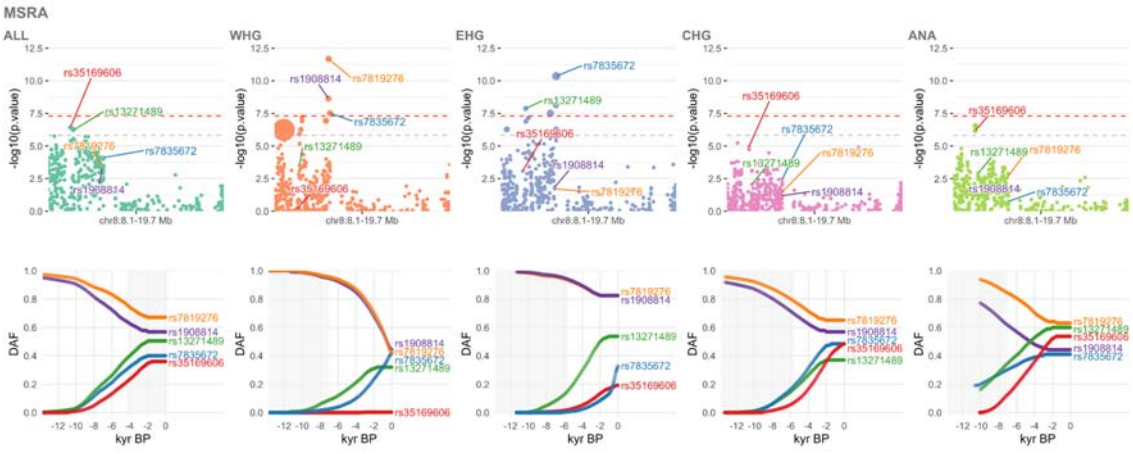


Figure S4a.29. Selection at the *MSRA* locus, spanning chr8:8142579-19746880. Results for the pan-ancestry analysis (ALL) plus the four marginal ancestries: Western hunter-gatherers (WHG), Eastern hunter-gatherers (EHG), Caucasus hunter-gatherers (CHG) and Anatolian farmers (ANA). Row one shows zoomed Manhattan plots of the p-values for each ancestry (significant SNPs sized by their selection coefficients), and row two shows allele trajectories for the top SNPs across all ancestries.

This peak spanned the region chr8:8142579-19746880, with significant SNPs including:

- rs7819276 (*AC025857.1 - OR7E158P*; WHG; p=2.10e-12; s=-0.0207), associated with General factor of neuroticism [94].
- rs7835672 (*AC107918.3*; EHG; p=4.43e-11; s=0.0287), associated with General factor of neuroticism [94].
- rs1908814 (*AC025857.1 - OR7E158P*; WHG; p=2.33e-09; s=-0.0201), associated with General factor of neuroticism; Neuroticism [94,95].
- rs13271489 (*LINC00599 - AC034111.2*; EHG; p=1.33e-08; s=0.016), associated with Diastolic blood pressure x smoking status (current vs non-current) interaction (2df test); Diastolic blood pressure x smoking status (ever vs never) interaction (2df test); Systolic blood pressure x smoking status (current vs non-current) interaction (2df test); Systolic blood pressure x smoking status (ever vs never) interaction (2df test) [96].
- rs35169606 (*TNKS*; ALL; p=3.85e-07; s=0.0137), associated with Lifetime smoking index [97].

Peak 12: ABO

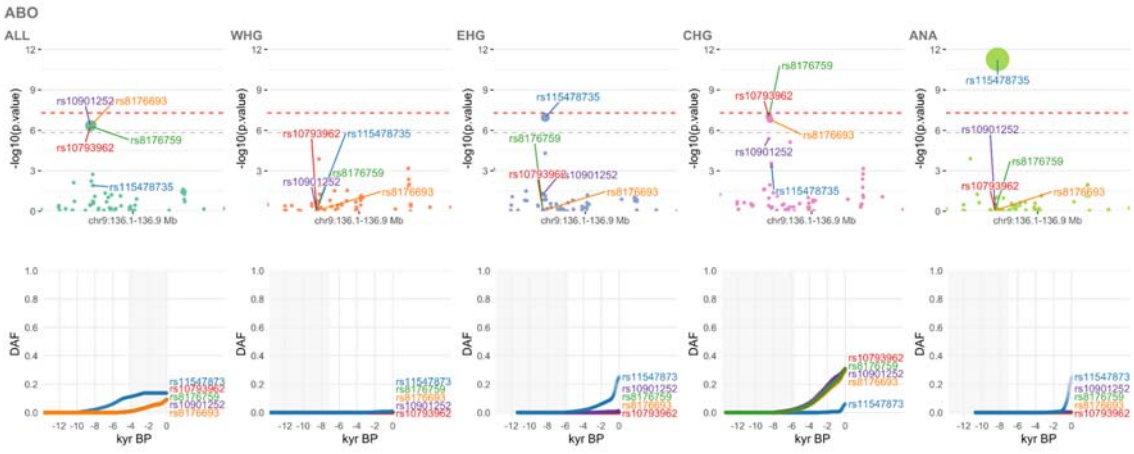


Figure S4a.30. Selection at the *ABO* locus, spanning chr9:136127999-136925660. Results for the pan-ancestry analysis (ALL) plus the four marginal ancestries: Western hunter-gatherers (WHG), Eastern hunter-gatherers (EHG), Caucasus hunter-gatherers (CHG) and Anatolian farmers (ANA). Row one shows zoomed Manhattan plots of the p-values for each ancestry (significant SNPs sized by their selection coefficients), and row two shows allele trajectories for the top SNPs across all ancestries.

This peak spanned the region chr9:136127999-136925660, with significant SNPs including:

- rs115478735 (*ABO*; ANA; p=5.36e-12; s=0.095), associated with Blood protein levels [40].
- rs8176759 (*ABO*; CHG; p=8.90e-08; s=0.0227), associated with Granulocyte percentage of myeloid white cells; Plateletcrit [83].
- rs10793962 (*ABO*; CHG; p=9.78e-08; s=0.0226), associated with Blood protein levels; Intraocular pressure [40,98,99].
- rs8176693 (*ABO*; CHG; p=1.55e-07; s=0.0233), associated with Blood protein levels; Blood protein levels in cardiovascular risk; Endothelial growth factor levels; High serum lipase activity; Serum lipase activity [40,100–102].
- rs10901252 (*ABO*; ALL; p=4.39e-07; s=0.0371), associated with Blood protein levels; Hematocrit; Hemoglobin concentration; vWF levels [43,83,103].

Peak 13: MYBPC3

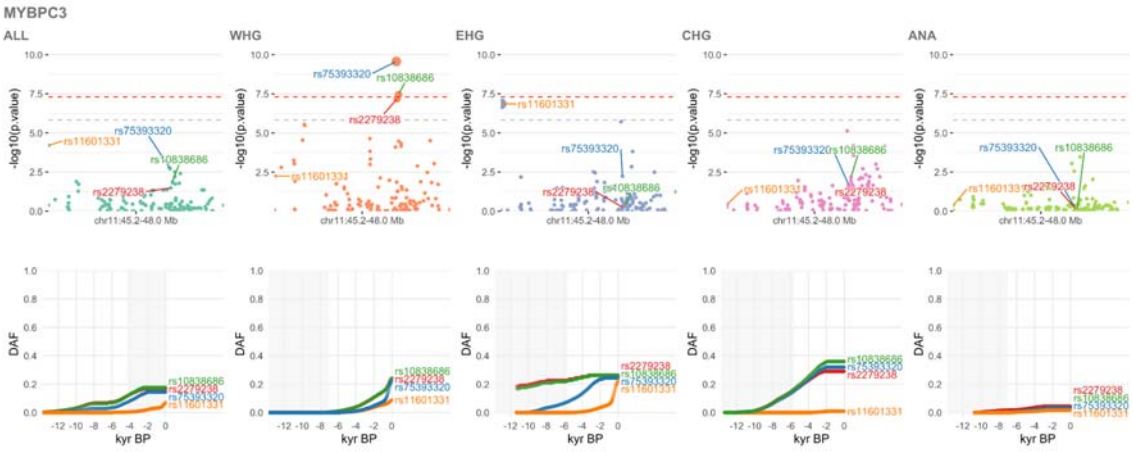


Figure S4a.31. Selection at the *MYBPC3* locus, spanning chr11:45227569-48018360. Results for the pan-ancestry analysis (ALL) plus the four marginal ancestries: Western hunter-gatherers (WHG), Eastern hunter-gatherers (EHG), Caucasus hunter-gatherers (CHG) and Anatolian farmers (ANA). Row one shows zoomed Manhattan plots of the p-values for each ancestry (significant SNPs sized by their selection coefficients), and row two shows allele trajectories for the top SNPs across all ancestries.

This peak spanned the region chr11:45227569-48018360, with significant SNPs including:

- rs75393320 (*ACP2*; WHG; p=2.73e-10; s=0.0351), associated with HDL cholesterol [104].
- rs10838686 (*MADD*; WHG; p=3.56e-08; s=0.0261), associated with High density lipoprotein cholesterol levels [105].
- rs2279238 (*NR1H3*; WHG; p=6.49e-08; s=0.0261), associated with Creatinine levels [106].
- rs11601331 (*TSPAN18*; EHG; p=1.38e-07; s=0.04), associated with Cortical brain region measurements (area, volume and thickness) [107].

Peak 14: FADS2

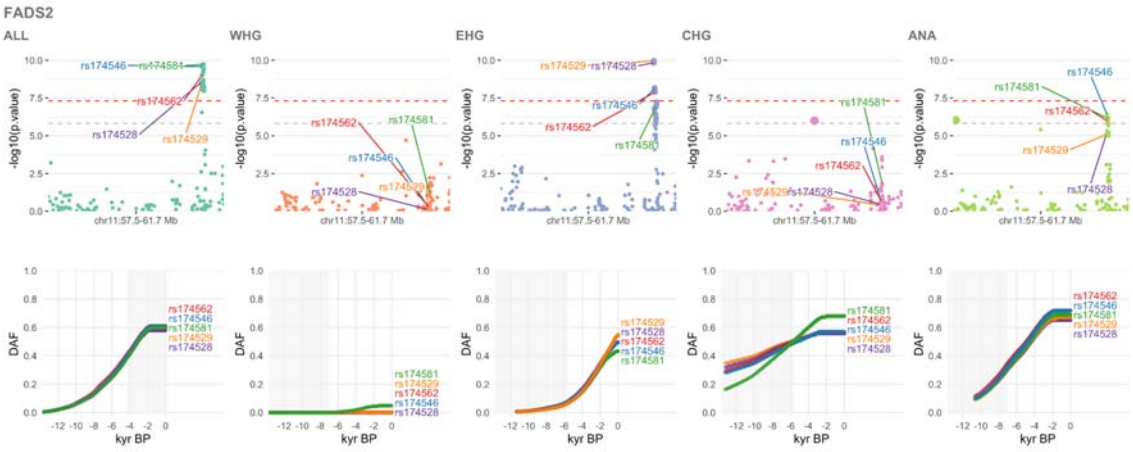


Figure S4a.32. Selection at the *FADS2* locus, spanning chr11:57467039-61706010. Results for the pan-ancestry analysis (ALL) plus the four marginal ancestries: Western hunter-gatherers (WHG), Eastern hunter-gatherers (EHG), Caucasus hunter-gatherers (CHG) and Anatolian farmers (ANA). Row one shows zoomed Manhattan plots of the p-values for each ancestry (significant SNPs sized by their selection coefficients), and row two shows allele trajectories for the top SNPs across all ancestries.

This peak spanned the region chr11:57467039-61706010, with significant SNPs including:

- rs174529 (*TMEM258,MYRF*; EHG; p=1.12e-10; s=0.0198), associated with HDL cholesterol; Heel bone mineral density; High density lipoprotein cholesterol levels; Low density lipoprotein cholesterol levels; Total cholesterol levels; Triglyceride levels; Triglycerides [58,105,108–110].
- rs174528 (*TMEM258,MYRF*; EHG; p=1.48e-10; s=0.0198), associated with Gondoic acid (20:1n-9) levels; Phosphatidylcholine-ether levels; Plasma omega-6 polyunsaturated fatty acid levels (arachidonic acid); Serum metabolite ratios in chronic kidney disease; Trans fatty acid levels; Vaccenic acid (18:1n-7) levels [47,111–114].
- rs174581 (*FADS2*; ALL; p=1.87e-10; s=0.0133), associated with Male-pattern baldness; Serum metabolite ratios in chronic kidney disease [46,47].
- rs174546 (*FADS2,FADS1*; ALL; p=2.65e-10; s=0.0131), associated with C-reactive protein levels or HDL-cholesterol levels (pleiotropy); C-reactive protein levels or triglyceride levels (pleiotropy); Change in serum metabolite levels; Change in serum metabolite levels (CMS); Cholesterol, total; Glycerophospholipid levels; HDL cholesterol; HDL cholesterol levels; High density lipoprotein cholesterol levels; LDL cholesterol; LDL cholesterol levels; Low density lipoprotein cholesterol levels; Plasma omega-6 polyunsaturated fatty acid levels (gamma-linolenic acid); QT interval; Serum metabolite levels; Serum metabolite levels (CMS); Total cholesterol levels; Trans fatty acid levels; Triglyceride levels; Triglycerides [105,110–112,115–121].
- rs174562 (*FADS2,FADS1*; ALL; p=8.63e-10; s=0.0127), associated with Asthma; Serum metabolite ratios in chronic kidney disease [47,122].

Peak 15: HECTD4

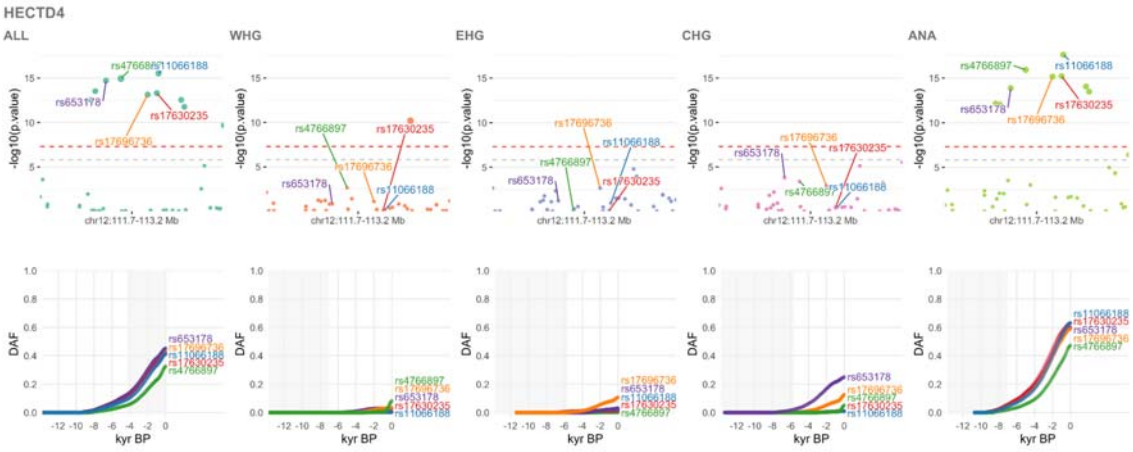


Figure S4a.33. Selection at the *HECTD4* locus, spanning chr12:111706879-113205500. Results for the pan-ancestry analysis (ALL) plus the four marginal ancestries: Western hunter-gatherers (WHG), Eastern hunter-gatherers (EHG), Caucasus hunter-gatherers (CHG) and Anatolian farmers (ANA). Row one shows zoomed Manhattan plots of the p-values for each ancestry (significant SNPs sized by their selection coefficients), and row two shows allele trajectories for the top SNPs across all ancestries.

This peak spanned the region chr12:111706879-113205500, with significant SNPs including:

- rs11066188 (*HECTD4*; ANA; p=2.35e-18; s=0.0202), associated with Celiac disease and Rheumatoid arthritis [48].
- rs4766897 (*ACAD10*; ANA; p=1.17e-16; s=0.0208), associated with Fibrinogen levels [123].
- rs17630235 (*TRAFD1 - HECTD4*; ANA; p=5.98e-16; s=0.0195), associated with Body mass index; Diastolic blood pressure; Parental longevity (combined parental attained age, Martingale residuals); Systolic blood pressure; Tonsillectomy [39,42,124–126].
- rs17696736 (*NAA25*; ANA; p=6.77e-16; s=0.0196), associated with Coronary artery disease; Diastolic blood pressure x alcohol consumption interaction (2df test); Estimated glomerular filtration rate; Mean arterial pressure; Mean arterial pressure x alcohol consumption interaction (2df test); Parental longevity (combined parental attained age, Martingale residuals); Systolic blood pressure x alcohol consumption interaction (2df test); Type 1 diabetes; Urate levels [91,126–133].
- rs653178 (*ATXN2*; ALL; p=1.92e-15; s=0.0194), associated with Allergic disease (asthma, hay fever or eczema); Asthma; Blood pressure; Celiac disease; Celiac disease or Rheumatoid arthritis; Chronic kidney disease; Diastolic blood pressure; Eczema; Eosinophil counts; Eosinophil percentage of granulocytes; Eosinophil percentage of white cells; Hay fever and/or eczema; Inflammatory bowel disease; LDL cholesterol; Mean arterial pressure; Monocyte count; Myocardial infarction; Neutrophil percentage of granulocytes; Sarcoidosis; Sum eosinophil basophil counts; Systemic lupus erythematosus; Thyroid peroxidase antibody positivity; Tonsillectomy; Total cholesterol levels; Type 1 diabetes; Urate levels [58,82,83,104,125,130,134–148].

Peak 16: RFLNA

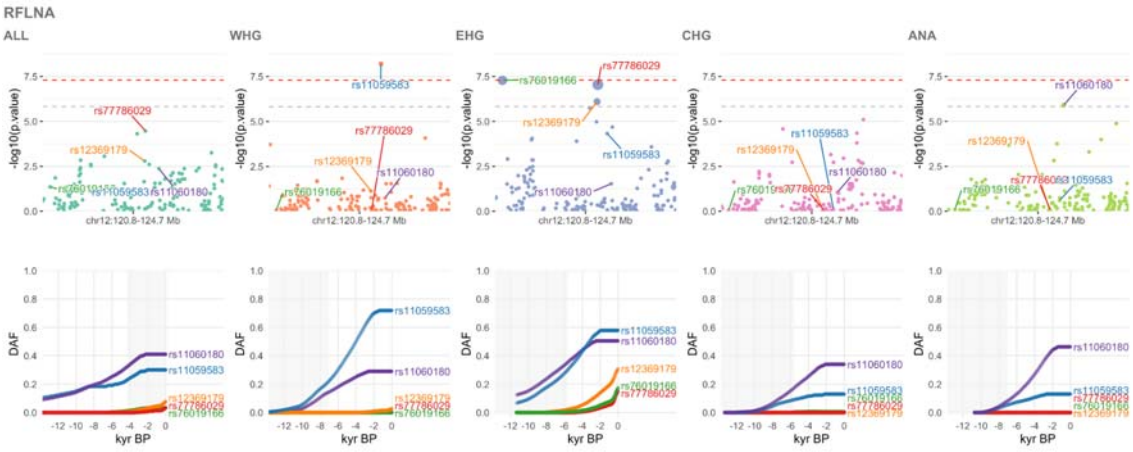


Figure S4a.34. Selection at the *RFLNA* locus, spanning chr12:120813919-124667680. Results for the pan-ancestry analysis (ALL) plus the four marginal ancestries: Western hunter-gatherers (WHG), Eastern hunter-gatherers (EHG), Caucasus hunter-gatherers (CHG) and Anatolian farmers (ANA). Row one shows zoomed Manhattan plots of the p-values for each ancestry (significant SNPs sized by their selection coefficients), and row two shows allele trajectories for the top SNPs across all ancestries.

This peak spanned the region chr12:120813919-124667680, with significant SNPs including:

- rs11059583 (*HCAR3 - HCAR1*; WHG; p=6.21e-09; s=0.0144), associated with Waist-to-hip ratio adjusted for BMI [72].
- rs76019166 (*PXN*; EHG; p=5.20e-08; s=0.0339), associated with Appendicular lean mass [149].
- rs77786029 (*ZCCHC8*; EHG; p=9.02e-08; s=0.0399), associated with Red cell distribution width [58].
- rs12369179 (*ZCCHC8*; EHG; p=7.96e-07; s=0.0239), associated with Body mass index; Waist-to-hip ratio adjusted for BMI [72,150].
- rs11060180 (*CCDC62*; ANA; p=1.27e-06; s=0.0146), associated with Parkinson's disease [151–153].

Peak 17: MARK3

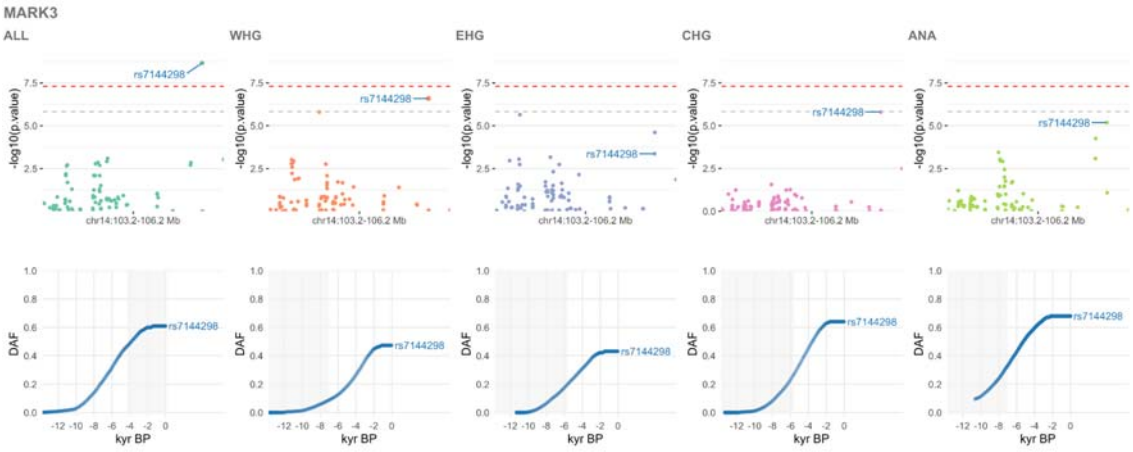


Figure S4a.35. Selection at the *MARK3* locus, spanning chr14:103239629-106248400. Results for the pan-ancestry analysis (ALL) plus the four marginal ancestries: Western hunter-gatherers (WHG), Eastern hunter-gatherers (EHG), Caucasus hunter-gatherers (CHG) and Anatolian farmers (ANA). Row one shows zoomed Manhattan plots of the p-values for each ancestry (significant SNPs sized by their selection coefficients), and row two shows allele trajectories for the top SNPs across all ancestries.

This peak spanned the region chr14:103239629-106248400, with significant SNPs including:

- rs7144298 (*IGHG3 - AL122127.1*; ALL; p=2.19e-09; s=0.0122), associated with Blood protein levels [40].

Peak 18: SEMA6D

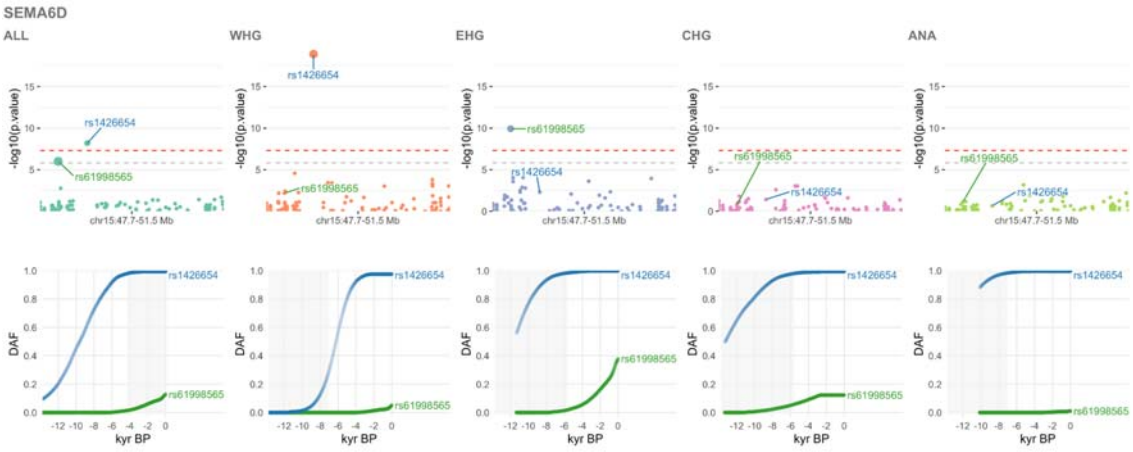


Figure S4a.36. Selection at the *SEMA6D* locus, spanning chr15:47665179-51534060. Results for the pan-ancestry analysis (ALL) plus the four marginal ancestries: Western hunter-gatherers (WHG), Eastern hunter-gatherers (EHG), Caucasus hunter-gatherers (CHG) and Anatolian farmers (ANA). Row one shows zoomed Manhattan plots of the p-values for each ancestry (significant SNPs sized by their selection coefficients), and row two shows allele trajectories for the top SNPs across all ancestries.

This peak spanned the region chr15:47665179-51534060, with significant SNPs including:

- rs1426654 (*SLC24A5*; WHG; p=1.25e-19; s=0.0305), associated with Eye colour; Eye colour (brightness); Eye colour (saturation); Hair colour; Iris colour (b* coordinate); Skin pigmentation; Skin reflectance (Melanin index); Skin, hair and eye pigmentation (multivariate analysis) [76,78,80,154,155].
- rs61998565 (*SEMA6D,AC023905.1*; EHG; p=1.19e-10; s=0.023), associated with Heel bone mineral density [45,58,109].

Peak 19: CSK

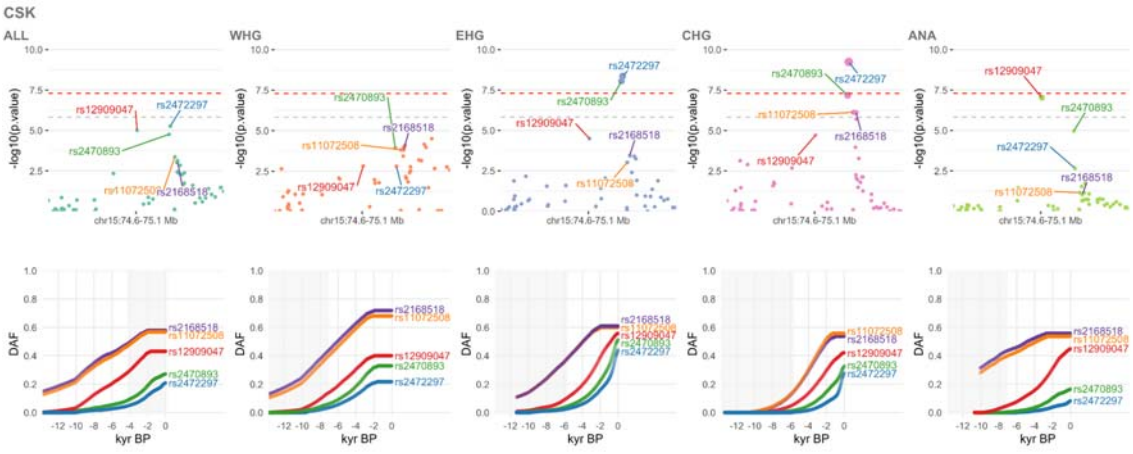


Figure S4a.37. Selection at the *CSK* locus, spanning chr15:74607429-75101530. Results for the pan-ancestry analysis (ALL) plus the four marginal ancestries: Western hunter-gatherers (WHG), Eastern hunter-gatherers (EHG), Caucasus hunter-gatherers (CHG) and Anatolian farmers (ANA). Row one shows zoomed Manhattan plots of the p-values for each ancestry (significant SNPs sized by their selection coefficients), and row two shows allele trajectories for the top SNPs across all ancestries.

This peak spanned the region chr15:74607429-75101530, with significant SNPs including:

- rs2472297 (*CYP1A1 - CYP1A2*; CHG; p=5.80e-10; s=0.0304), associated with Alcohol consumption (drinks per week); Alcohol consumption (drinks per week) (MTAG); Bitter beverage consumption; Bitter non-alcoholic beverage consumption; Caffeine metabolism (plasma 1,3,7-trimethylxanthine (caffeine) level); Caffeine metabolism (plasma 1,7-dimethylxanthine (paraxanthine) to 1,3,7-trimethylxanthine (caffeine) ratio); Coffee consumption; Coffee consumption (cups per day); Estimated glomerular filtration rate; Estimated glomerular filtration rate in non-diabetics; Plasma clozapine levels in treatment-resistant schizophrenia; Predicted visceral adipose tissue; Tea consumption; Urate levels; Urinary albumin excretion; Urinary albumin-to-creatinine ratio; Urinary potassium excretion; Urinary sodium excretion [86,91,133,156–165].
- rs2470893 (*CYP1A1 - CYP1A2*; EHG; p=8.93e-09; s=0.0217), associated with Blood urea nitrogen levels; Caffeine consumption; Caffeine metabolism (plasma 1,3,7-trimethylxanthine (caffeine) level); Caffeine metabolism (plasma 1,7-dimethylxanthine (paraxanthine) to 1,3,7-trimethylxanthine (caffeine) ratio); Coffee consumption; Microalbuminuria; Platelet distribution width; Urinary albumin excretion (no hypertensive medication); Urinary albumin-to-creatinine ratio [83,133,157–159,166–168].
- rs12909047 (*AC012435.2,AC012435.3,UBL7-AS1*; ANA; p=9.39e-08; s=0.019), associated with Caffeine metabolism (plasma 1,3,7-trimethylxanthine (caffeine) level); Caffeine metabolism (plasma 1,3-dimethylxanthine (theophylline) level) [158].
- rs11072508 (*CYP1A2 - CSK*; CHG; p=7.45e-07; s=0.0163), associated with Cardiovascular disease; Medication use (agents acting on the renin-angiotensin system) [58,169].
- rs2168518 (*MIR4513,CSK*; CHG; p=7.73e-07; s=0.0163), associated with Medication use (calcium channel blockers) [169].

Peak 20: DPEP1

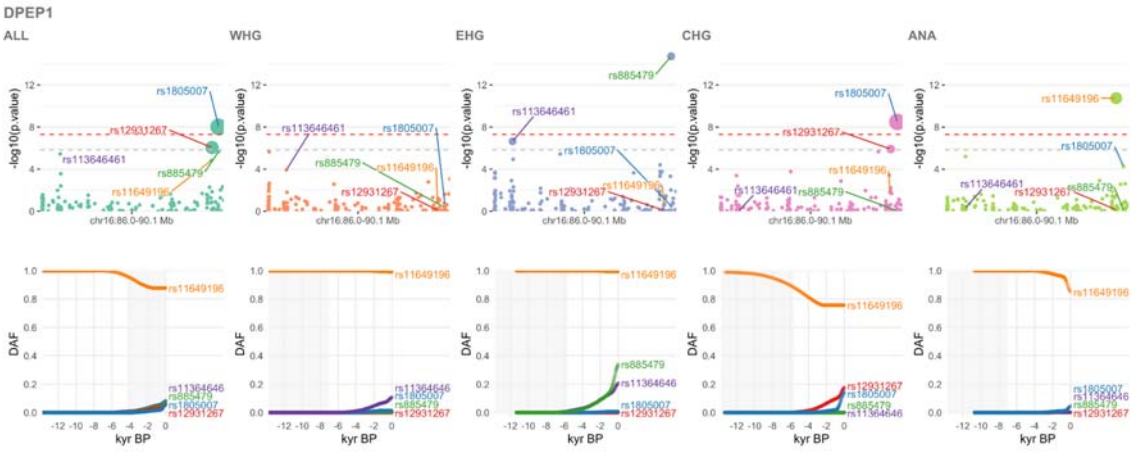


Figure S4a.38. Selection at the *DPEP1* locus, spanning chr16:85972609-90084560. Results for the pan-ancestry analysis (ALL) plus the four marginal ancestries: Western hunter-gatherers (WHG), Eastern hunter-gatherers (EHG), Caucasus hunter-gatherers (CHG) and Anatolian farmers (ANA). Row one shows zoomed Manhattan plots of the p-values for each ancestry (significant SNPs sized by their selection coefficients), and row two shows allele trajectories for the top SNPs across all ancestries.

This peak spanned the region chr16:85972609-90084560, with significant SNPs including:

- rs885479 (*AC092143.1,MC1R*; EHG; p=1.92e-15; s=0.0284), associated with Blond vs. brown/black hair colour [57].
- rs11649196 (*ZNF276*; ANA; p=1.79e-11; s=-0.0457), associated with Keratinocyte cancer (MTAG) [59].
- rs1805007 (*AC092143.1,MC1R*; CHG; p=3.22e-09; s=0.0648), associated with Balding type 1; Basal cell carcinoma; Blond vs. brown hair colour; Blond vs. brown/black hair colour; Brown vs. black hair colour; Cutaneous squamous cell carcinoma; Freckles; Hair colour; Hair morphology traits; Keratinocyte cancer (MTAG); Melanoma; Non-melanoma skin cancer; Red vs non-red hair colour; Red vs. brown/black hair colour; Skin pigmentation traits; Skin sensitivity to sun; Squamous cell carcinoma; Sunburns; Tanning [49–59].
- rs113646461 (*AC092723.5 - AC092723.4*; EHG; p=2.28e-07; s=0.0273), associated with Monocyte percentage of white cells [83].
- rs12931267 (*FANCA*; ALL; p=8.87e-07; s=0.0506), associated with Freckling; Hair colour; Hair morphology traits; Skin pigmentation traits; Skin sensitivity to sun [56,77,170].

Peak 21: MAPT

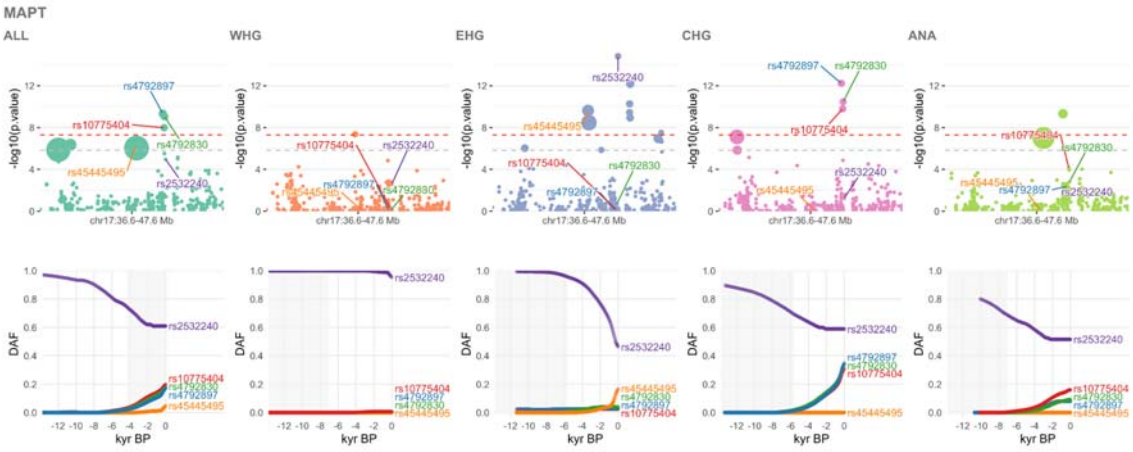


Figure S4a.39. Selection at the *MAPT* locus, spanning chr17:36615969-47588760. Results for the pan-ancestry analysis (ALL) plus the four marginal ancestries: Western hunter-gatherers (WHG), Eastern hunter-gatherers (EHG), Caucasus hunter-gatherers (CHG) and Anatolian farmers (ANA). Row one shows zoomed Manhattan plots of the p-values for each ancestry (significant SNPs sized by their selection coefficients), and row two shows allele trajectories for the top SNPs across all ancestries.

This peak spanned the region chr17:36615969-47588760, with significant SNPs including:

- rs2532240 (*KANSL1*; EHG; p=1.51e-15; s=-0.0217), associated with White matter microstructure (axial diusivities) [171].
- rs4792897 (*MAPT*; CHG; p=5.94e-13; s=0.0248), associated with Snoring [60].
- rs10775404 (*KANSL1*; CHG; p=1.60e-10; s=0.0248), associated with Reaction time [74].
- rs45445495 (*SLC4A1*; EHG; p=2.34e-10; s=0.0448), associated with Appendicular lean mass [149].

Discussion

Using our ancient genomic panel, we sought to identify phenotype-associated variants that have evidence for directional selection over the last 13,000 years. To estimate allele frequency trajectories and selection coefficients of trait-associated variants through time, we used the software *CLUES* [1] which can perform inference of allele frequency trajectories using marginal trees sampled from a reconstruction of an ancestral recombination graph (ARG) [2] for a set of genomic sequences, in combination with genotype likelihoods from serially sampled ancient DNA (aDNA).

Our results show that the incorporation of ancient DNA considerably improves our power to detect variants under selection, compared to a method that only uses the ARG inferred from present-day data alone. Using genomes from the 1,000 Genomes Project project (populations GBR, FIN and TSI), we inferred allele trajectories and selection coefficients for 35,592 phenotype-associated variants, ascertained from the GWAS Catalog [4], along with an equal number of putatively neutral "control" variants. Our analysis identified no genome-wide significant selective sweeps ($p < 5e-8$) using present-day data alone. However, the trait-associated variants were significantly enriched for evidence of selection when compared to the control group (Wilcoxon signed-rank test, $p < 2.2e-16$).

Pan-ancestry selection

In contrast to patterns observed in present-day genomes, our selection analysis based on a time-series of ancient DNA genotype probabilities identified 11 genome-wide significant ($p < 5e-8$) selective sweeps in the GWAS variants, and none in the control group; consistent with widespread selection acting on trait-associated variants. This analysis confirms many of the previously reported selection loci in West Eurasians, identified from present-day and ancient DNA [16,172–174], and reveals novel selective sweeps, while refining the temporal dynamics of the selected alleles.

The strongest overall signal of selection in the pan-ancestry analysis is at the *LCT / MCM6* locus (rs4988235; p=9.86e-31; s=0.020), the derived allele of which is casual for lactase persistence [35,36]. The inferred trajectory indicates that this allele began rising in frequency c. 6,000 years ago, and has continued to rise in frequency up to the present (Supplementary Figure S4a.5).

We find a strong signal of selection at the *FADS1* (rs174546; p=2.65e-10; s=0.013) and *FADS2* (rs174581; p=1.87e-10; s=0.013) locus, associated with fatty acid metabolism

[116,118,121,175–177]. The trajectories for these variants indicate a rise in frequency, beginning around 13,000 years ago, and continuing up until c. 2,000 years ago, after which their frequencies plateaued (Supplementary Figure S4a.11). In contrast to earlier findings [16], we do not detect a significant signal of selection at the *DHCR7* and *NADSYN1* locus, associated with vitamin D levels (most significant SNP rs4423214; p=5.54e-03; s=-0.006).

We detect an 8 megabase (Mb) wide selection sweep signal in chromosome 6 (chr6:25.2-33.5 Mb), spanning the human leukocyte antigen (HLA) region. The selection trajectories of the variants within this locus support multiple independent sweeps, occurring at different times and with differing intensities. The strongest signal of selection at this locus in the pan-ancestry analysis is at an intergenic variant, located between *HLA-A* and *HLA-W* (rs7747253; p=8.86e-17; s=-0.018), associated with heel bone mineral density [45], the derived allele of which rapidly reduced in frequency, beginning c. 8,000 years ago (Supplementary Figure S4a.10). In contrast, the signal of selection at *C2* (rs9267677; p= 9.82e-14; s= 0.04463), also found within this sweep, and associated with educational attainment [66], shows a gradual increase in frequency beginning c. 4,000 years ago, before rising more rapidly c. 1,000 years ago; highlighting the complex temporal dynamics of selection at the HLA locus.

We also identify selection signals at the *SLC22A4* (rs35260072; p=1.15e-10; s=0.018) and *RAPGEF6* (rs11950815; p=1.82e-12; s=0.021) loci, associated with asthma [148] and itch intensity from mosquito bites [34], and find that these alleles have been steadily rising in frequency, beginning c. 8,000 years ago (Supplementary Figure S4a.8). However, we find that the frequency of rs1050152 plateaued c. 1,500 years ago, contrary to previous reports suggesting a recent rise in frequency [16]. Similarly, we detect selection at the *HECTD4* (rs11066188; p=3.02e-16; s=0.020) and ATXN2 (rs653178; p=1.92e-15; s=0.019) locus, associated with celiac disease and rheumatoid arthritis [48], which has been rising in frequency for c. 9,000 years (Supplementary Figure S4a.12), also contrary to previous reports of a more recent rise in frequency [16].

We detect strong selection at the *SLC45A2* (rs35395; p=4.13e-23; s=0.022) locus, associated with skin pigmentation [76,178], and find that the selected allele began rising in frequency c. 13,000 years ago, after which it plateaued at high frequency c. 2,000 years ago. This is similar to the selection trajectory at the independent *GRM5* (rs7119749; p=8.54e-09; s=0.011) locus, which plateaued at medium-high frequency at approximately the same time. We also detect strong selection at the *SLC24A5* (rs1426654; p=6.45e-09; s=0.019) locus, and find that the selected allele, also associated with skin pigmentation [76,154], began rising in frequency even earlier than *SLC45A2*, and reached near fixation c. 3,500 years ago.

We further detect strong selection in an 11 Mb sweep in chromosome 17 (chr17:36.6-47.5), spanning the 17q21.31 locus, a 900-kb inversion polymorphism [179,180]. The strongest signal of selection in this sweep is at *MAPT (*rs4792897; p=4.65e-10; s=0.03), associated with snoring [60], the trajectory for which indicates a steady increase in frequency, beginning c. 7,000 years ago (Supplementary Figure S4a.14).

Ancestry stratified selection trajectories

To account for population structure in our samples, we also applied a novel chromosome painting technique, based on inference of a sample's nearest neighbours in the marginal trees of an ARG that contains individuals classified into different ancient populations (Supplementary Note 3i). This method allows us to accurately assign ancestral population labels to haplotypes found in both ancient and present-day individuals. By conditioning our selection analyses on these haplotype backgrounds, we can infer the selection trajectories of GWAS risk alleles in a manner that is approximately invariant to change in the admixture proportions through time. These ancestry specific allele trajectories reveal many novel aspects about the dynamic interplay between selection and admixture in West Eurasia throughout the Holocene. We find that the allele trajectories of directionally selected sites become much more apparent once we perform this ancestry partitioning. We often find variants with strong allele frequency changes in one ancestral population but not another, and analysing all ancient individuals without accounting for their ancestry composition leads to a decrease in our ability to identify selected variants, and a blurring of the temporal signal of allele frequency changes.

When conditioned on one of our four marginal ancestries—Western hunter-gatherers (WHG), Eastern hunter-gatherers (EHG), Caucasus hunter-gatherers (CHG) and Anatolian farmers (ANA)—we find 21 genome-wide significant selection peaks (substantially more than in our pan-ancestry analysis), indicating that admixture between ancestral populations has masked evidence of selection at many loci. Furthermore, we find that some strong signals of selection identified in the pan-ancestry analysis are driven by sweeps in the marginal ancestries which substantially differ in their most significant SNPs, suggesting that multiple selected alleles may be common within genome-wide significant sweep loci.

For example, in the pan-ancestry analysis, the sweep at *MCM6* is led by rs4988235, consistent with the widespread interpretation that selection has acted upon the lactase persistence phenotype. However, in the ancestry stratified analysis, this selection signal is primarily driven by sweeps in the EHG and CHG ancestral backgrounds, which differ in their

most significant SNPs (Supplementary Figure S4a.21). The strongest sweep signal in all marginal ancestries at this locus is in the EHG background; however, conditional on that background, rs1246580 (*R3HDM1*) is the most significant SNP (associated with blood protein levels [40] and mosquito bite size [34]). In CHG, rs4988235 is the most significant SNP, but there is no evidence for selection at rs1246580 in this background. Conversely, in WHG, we find no evidence for selection at rs4988235, and instead find evidence for strong selection at rs1246580 occurring in the last c. 2,000 years. Despite the highly studied nature of this locus, a satisfactory explanation for the observed strength of selection has remained elusive [181–184].

In comparison, the sweep at the *SLC45A2* locus shows a much simpler pattern, in which all marginal ancestries show broad agreement at this locus (Supplementary Figure S4a.25). Where they differ lies primarily in the timing of their frequency rises. The ANA ancestry background shows the earliest evidence for selection at *SLC45A2*, followed by EHG and WHG, beginning around c. 10,000 years ago, and CHG c. 2,000 years later. In all ancestry backgrounds except WHG, the selected haplotypes reach near fixation by c. 3,000 years ago, whilst the WHG haplotype background contains the majority of the ancestral alleles still segregating in present-day Europeans.

At the *FADS2* locus, the strong signal of selection in the pan-ancestry analysis is driven primarily by a sweep occurring on the EHG haplotypic background. Interestingly, we find no evidence for selection at this locus in the WHG background, and most of the frequency rise in the EHG background occurs after their admixture with CHG, where the selected alleles were already at close to present-day frequencies.

References

1. Stern, A. J., Wilton, P. R. & Nielsen, R. An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. *PLoS Genet.* **15**, e1008384 (2019).

2. Speidel, L., Forest, M., Shi, S. & Myers, S. R. A method for genome-wide genealogy estimation for thousands of samples. *Nat. Genet.* **51**, 1321–1329 (2019).

3. Köster, J. & Rahmann, S. Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).

4. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).

5. Jupp, S., Burdett, T., Leroy, C. & Parkinson, H. E. A new Ontology Lookup Service at EMBL-EBI. in *SWAT4LS* 118–119 (ceur-ws.org, 2015).

6. Yates, A. *et al.* The Ensembl REST API: Ensembl Data for Any Language. *Bioinformatics* **31**, 143–145 (2015).

7. Aken, B. L. *et al.* The Ensembl gene annotation system. *Database* **2016**, (2016).

8. Kelleher, J., Etheridge, A. M. & McVean, G. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLoS Comput. Biol.* **12**, e1004842 (2016).

9. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

10. Vilella, A. J. *et al.* EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* **19**, 327–335 (2009).

11. Rasmussen, M. D., Hubisz, M. J., Gronau, I. & Siepel, A. Genome-wide inference of ancestral recombination graphs. *PLoS Genet.* **10**, e1004342 (2014).

12. Hein, J., Schierup, M. & Wiuf, C. *Gene Genealogies, Variation and Evolution: A primer in coalescent theory*. (Oxford University Press, USA, 2004).

13. Stern, A. J., Speidel, L., Zaitlen, N. A. & Nielsen, R. Disentangling selection on genetically correlated polygenic traits via whole-genome genealogies. *Am. J. Hum. Genet.* (2021) doi:10.1016/j.ajhg.2020.12.005.

14. 1000 Genomes Project Consortium *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).

15. Haller, T., Tasa, T. & Metspalu, A. Manhattan Harvester and Cropper: a system for GWAS peak detection. *BMC Bioinformatics* **20**, 22 (2019).

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

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

18. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

19. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

20. Anaconda Software Distribution. *Anaconda Documentation* (2020).

21. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).

22. McKinney, W. & Others. Data structures for statistical computing in python. in *Proceedings of the 9th Python in Science Conference* vol. 445 51–56 (Austin, TX, 2010).

23. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

24. Van Rossum, G. & Drake, F. L. *Python 3 Reference Manual: (Python Documentation Manual Part 2)*. (CreateSpace Independent Publishing Platform, 2009).

25. R Core Team. R: A Language and Environment for Statistical Computing. (2019).

26. Haider, S. *et al.* A bedr way of genomic interval processing. *Source Code Biol. Med.* **11**, 14 (2016).

27. Wickham, H., François, R., Henry, L. & Müller, K. dplyr: A Grammar of Data Manipulation. (2019).

28. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer, 2016).

29. Petukhov, V., van den Brand, T. & Biederstedt, E. ggrastr: Raster Layers for 'ggplot2'. (2020).

30. Slowikowski, K. ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'. (2020).

31. Wilke, C. O. ggridges: Ridgeline Plots in ‘ggplot2’. (2018).

32. Wickham, H. stringr: Simple, Consistent Wrappers for Common String Operations. (2019).

33. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

34. Jones, A. V. *et al.* GWAS of self-reported mosquito bite size, itch intensity and attractiveness to mosquitoes implicates immune-related predisposition loci. *Hum. Mol. Genet.* **26**, 1391–1406 (2017).

35. Enattah, N. S. *et al.* Identification of a variant associated with adult-type hypolactasia. *Nat. Genet.* **30**, 233–237 (2002).

36. Bersaglieri, T. *et al.* Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* **74**, 1111–1120 (2004).

37. Hill, W. D. *et al.* Genome-wide analysis identifies molecular systems and 149 genetic loci associated with income. *Nat. Commun.* **10**, 5741 (2019).

38. Shungin, D. *et al.* New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187–196 (2015).

39. Winkler, T. W. *et al.* The Influence of Age and Sex on Genetic Associations with Adult Body Size and Shape: A Large-Scale Genome-Wide Interaction Study. *PLoS Genet.* **11**, e1005378 (2015).

40. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).

41. Emilsson, V. *et al.* Co-regulatory networks of human serum proteins link genetics to disease. *Science* **361**, 769–773 (2018).

42. Hoffmann, T. J. *et al.* A Large Multiethnic Genome-Wide Association Study of Adult Body Mass Index Identifies Novel Loci. *Genetics* **210**, 499–515 (2018).

43. Ahsan, M. *et al.* The relative contribution of DNA methylation and genetic variants on protein biomarkers for human diseases. *PLoS Genet.* **13**, e1007005 (2017).

44. Zhu, Z. *et al.* A genome-wide cross-trait analysis from UK Biobank highlights

the shared genetic architecture of asthma and allergic diseases. *Nat. Genet.* **50**, 857–864 (2018).

45. Morris, J. A. *et al.* An atlas of genetic influences on osteoporosis in humans and mice. *Nat. Genet.* **51**, 258–266 (2018).

46. Hagenaars, S. P. *et al.* Genetic prediction of male pattern baldness. *PLoS Genet.* **13**, e1006594 (2017).

47. Li, Y. *et al.* Genome-Wide Association Studies of Metabolites in Patients with CKD Identify Multiple Loci and Illuminate Tubular Transport Mechanisms. *J. Am. Soc. Nephrol.* **29**, 1513–1524 (2018).

48. Gutierrez-Achury, J. *et al.* Functional implications of disease-specific variants in loci jointly associated with coeliac disease and rheumatoid arthritis. *Hum. Mol. Genet.* **25**, 180–190 (2016).

49. Sulem, P. *et al.* Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat. Genet.* **39**, 1443–1452 (2007).

50. Nan, H. *et al.* Genome-wide association study identifies novel alleles associated with risk of cutaneous basal cell carcinoma and squamous cell carcinoma. *Hum. Mol. Genet.* **20**, 3718–3724 (2011).

51. Zhang, M. *et al.* Genome-wide association studies identify several new loci associated with pigmentation traits and skin cancer risk in European Americans. *Hum. Mol. Genet.* **22**, 2948–2959 (2013).

52. Chahal, H. S. *et al.* Genome-wide association study identifies novel susceptibility loci for cutaneous squamous cell carcinoma. *Nat. Commun.* **7**, 12048 (2016).

53. Chahal, H. S. *et al.* Genome-wide association study identifies 14 novel risk alleles associated with basal cell carcinoma. *Nat. Commun.* **7**, 12510 (2016).

54. Ransohoff, K. J. *et al.* Two-stage genome-wide association study identifies a novel susceptibility locus associated with melanoma. *Oncotarget* **8**, 17586–17592 (2017).

55. Hysi, P. G. *et al.* Genome-wide association meta-analysis of individuals of European ancestry identifies new loci explaining a substantial fraction of hair color variation and heritability. *Nat. Genet.* **50**, 652–656 (2018).

56. Galván-Femenía, I. *et al.* Multitrait genome association analysis identifies new susceptibility genes for human anthropometric variation in the GCAT cohort. *J. Med. Genet.* **55**, 765–778 (2018).

57. Morgan, M. D. *et al.* Genome-wide study of hair colour in UK Biobank explains most of the SNP heritability. *Nat. Commun.* **9**, 5271 (2018).

58. Kichaev, G. *et al.* Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *Am. J. Hum. Genet.* **104**, 65–75 (2019).

59. Liyanage, U. E. *et al.* Combined analysis of keratinocyte cancers identifies novel genome-wide loci. *Hum. Mol. Genet.* **28**, 3148–3160 (2019).

60. Jansen, P. R. *et al.* Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. *Nat. Genet.* **51**, 394–403 (2019).

61. Akinkuolie, A. O. *et al.* Group IIA Secretory Phospholipase A2, Vascular Inflammation, and Incident Cardiovascular Disease. *Arterioscler. Thromb. Vasc. Biol.* **39**, 1182–1190 (2019).

62. Hill, W. D. *et al.* A combined analysis of genetically correlated traits identifies 187 loci and a role for neurogenesis and myelination in intelligence. *Mol. Psychiatry* **24**, 169–181 (2019).

63. Legge, S. E. *et al.* Association of Genetic Liability to Psychotic Experiences With Neuropsychotic Disorders and Traits. *JAMA Psychiatry* **76**, 1256–1265 (2019).

64. Okbay, A. *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533**, 539–542 (2016).

65. Savage, J. E. *et al.* Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat. Genet.* **50**, 912–919 (2018).

66. Lee, J. J. *et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121 (2018).

67. Li, M. *et al.* Genome-wide association study of 1,5-anhydroglucitol identifies novel genetic loci linked to glucose metabolism. *Sci. Rep.* **7**, 2812 (2017).

68. Schlosser, P. *et al.* Genetic studies of urinary metabolites illuminate mechanisms of detoxification and excretion in humans. *Nat. Genet.* **52**, 167–176 (2020).

69. Tikkanen, E. *et al.* Biological Insights Into Muscular Strength: Genetic Findings in the UK Biobank. *Sci. Rep.* **8**, 6451 (2018).

70. Jones, S. E. *et al.* Genome-wide association analyses of chronotype in 697,828 individuals provides insights into circadian rhythms. *Nat. Commun.* **10**, 343 (2019).

71. Ahola-Olli, A. V. *et al.* Genome-wide Association Study Identifies 27 Loci Influencing Concentrations of Circulating Cytokines and Growth Factors. *Am. J. Hum. Genet.* **100**, 40–50 (2017).

72. Zhu, Z. *et al.* Shared genetic and experimental links between obesity-related traits and asthma subtypes in UK Biobank. *J. Allergy Clin. Immunol.* **145**, 537–549 (2020).

73. Yap, C. X. *et al.* Dissection of genetic variation and evidence for pleiotropy in male pattern baldness. *Nat. Commun.* **9**, 5407 (2018).

74. Davies, G. *et al.* Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function. *Nat. Commun.* **9**, 2098 (2018).

75. Zhao, B. *et al.* Genome-wide association analysis of 19,629 individuals identifies variants influencing regional brain volumes and refines their genetic co-architecture with cognitive and mental health traits. *Nat. Genet.* **51**, 1637–1644 (2019).

76. Lona-Durazo, F. *et al.* Meta-analysis of GWA studies provides new insights on the genetic architecture of skin pigmentation in recently admixed populations. *BMC*

Genet. **20**, 59 (2019).

77. Liu, F. *et al.* Genetics of skin color variation in Europeans: genome-wide association studies with functional follow-up. *Hum. Genet.* **134**, 823–835 (2015).

78. Adhikari, K. *et al.* A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features. *Nat. Commun.* **7**, 10815 (2016).

79. Han, J. *et al.* A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. *PLoS Genet.* **4**, e1000074 (2008).

80. Adhikari, K. *et al.* A GWAS in Latin Americans highlights the convergent evolution of lighter skin pigmentation in Eurasia. *Nat. Commun.* **10**, 358 (2019).

81. Nan, H. *et al.* Genome-wide association study of tanning phenotype in a population of European ancestry. *J. Invest. Dermatol.* **129**, 2250–2257 (2009).

82. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).

83. Astle, W. J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167**, 1415–1429.e19 (2016).

84. Shin, S.-Y. *et al.* An atlas of genetic influences on human blood metabolites. *Nat. Genet.* **46**, 543–550 (2014).

85. Kristjansson, R. P. *et al.* A loss-of-function variant in ALOX15 protects against nasal polyps and chronic rhinosinusitis. *Nat. Genet.* **51**, 267–276 (2019).

86. Hellwege, J. N. *et al.* Mapping eGFR loci to the renal transcriptome and phenome in the VA Million Veteran Program. *Nat. Commun.* **10**, 3842 (2019).

87. Graham, S. E. *et al.* Sex-specific and pleiotropic effects underlying kidney function identified from GWAS meta-analysis. *Nat. Commun.* **10**, 1847 (2019).

88. Franke, A. *et al.* Sequence variants in IL10, ARPC2 and multiple other loci contribute to ulcerative colitis susceptibility. *Nat. Genet.* **40**, 1319–1323 (2008).

89. UK IBD Genetics Consortium *et al.* Genome-wide association study of

ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. *Nat. Genet.* **41**, 1330–1334 (2009).

90. Lee, H.-S. *et al.* An Intergenic Variant rs9268877 Between HLA-DRA and HLA-DRB Contributes to the Clinical Course and Long-term Outcome of Ulcerative Colitis. *J. Crohns. Colitis* **12**, 1113–1121 (2018).

91. Tin, A. *et al.* Target genes, variants, tissues and transcriptional pathways influencing human serum urate levels. *Nat. Genet.* **51**, 1459–1474 (2019).

92. Cipriani, V. *et al.* Genome-wide association study of age-related macular degeneration identifies associated variants in the TNXB-FKBPL-NOTCH4 region of chromosome 6p21.3. *Hum. Mol. Genet.* **21**, 4138–4150 (2012).

93. Weidinger, S. *et al.* A genome-wide association study of atopic dermatitis identifies loci with overlapping effects on asthma and psoriasis. *Hum. Mol. Genet.* **22**, 4841–4856 (2013).

94. Hill, W. D. *et al.* Genetic contributions to two special factors of neuroticism are associated with affluence, higher intelligence, better health, and longer life. *Mol. Psychiatry* **25**, 3034–3052 (2020).

95. Luciano, M. *et al.* Association analysis in over 329,000 individuals identifies 116 independent variants influencing neuroticism. *Nat. Genet.* **50**, 6–11 (2018).

96. Sung, Y. J. *et al.* A Large-Scale Multi-ancestry Genome-wide Study Accounting for Smoking Behavior Identifies Multiple Significant Loci for Blood Pressure. *Am. J. Hum. Genet.* **102**, 375–400 (2018).

97. Wootton, R. E. *et al.* Evidence for causal effects of lifetime smoking on risk for depression and schizophrenia: a Mendelian randomisation study. *Psychol. Med.* **50**, 2435–2443 (2020).

98. Gao, X. R., Huang, H., Nannini, D. R., Fan, F. & Kim, H. Genome-wide association analyses identify new loci influencing intraocular pressure. *Hum. Mol. Genet.* **27**, 2205–2213 (2018).

99. Khawaja, A. P. *et al.* Genome-wide analyses identify 68 new loci associated

with intraocular pressure and improve risk prediction for primary open-angle glaucoma. *Nat. Genet.* **50**, 778–782 (2018).

100. Weiss, F. U. *et al.* Fucosyltransferase 2 (FUT2) non-secretor status and blood group B are associated with elevated serum lipase activity in asymptomatic subjects, and an increased risk for chronic pancreatitis: a genetic association study. *Gut* **64**, 646–656 (2015).

101. Lieb, W. *et al.* Genome-wide association study for endothelial growth factors. *Circ. Cardiovasc. Genet.* **8**, 389–397 (2015).

102. Folkersen, L. *et al.* Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease. *PLoS Genet.* **13**, e1006706 (2017).

103. Sabater-Lleal, M. *et al.* Genome-Wide Association Transethnic Meta-Analyses Identifies Novel Associations Regulating Coagulation Factor VIII and von Willebrand Factor Plasma Levels. *Circulation* **139**, 620–635 (2019).

104. Klarin, D. *et al.* Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. *Nat. Genet.* **50**, 1514–1523 (2018).

105. Hoffmann, T. J. *et al.* A large electronic-health-record-based genome-wide study of serum lipids. *Nat. Genet.* **50**, 401–413 (2018).

106. Kanai, M. *et al.* Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat. Genet.* **50**, 390–400 (2018).

107. Bi, X. *et al.* Common genetic variants have associations with human cortical brain regions and risk of schizophrenia. *Genet. Epidemiol.* **43**, 548–558 (2019).

108. Surakka, I. *et al.* The impact of low-frequency and rare variants on lipid levels. *Nat. Genet.* **47**, 589–597 (2015).

109. Kim, S. K. Identification of 613 new loci associated with heel bone mineral density and a polygenic risk score for bone mineral density, osteoporosis and fracture. *PLoS One* **13**, e0200785 (2018).

110. Wojcik, G. L. *et al.* Genetic analyses of diverse populations improves

discovery for complex traits. *Nature* **570**, 514–518 (2019).

111. Mozaffarian, D. *et al.* Genetic loci associated with circulating phospholipid trans fatty acids: a meta-analysis of genome-wide association studies from the CHARGE Consortium. *Am. J. Clin. Nutr.* **101**, 398–406 (2015).

112. Dorajoo, R. *et al.* A genome-wide association study of n-3 and n-6 plasma fatty acids in a Singaporean Chinese population. *Genes Nutr.* **10**, 53 (2015).

113. Hu, Y. *et al.* Discovery and fine-mapping of loci associated with MUFAs through trans-ethnic meta-analysis in Chinese and European populations. *J. Lipid Res.* **58**, 974–981 (2017).

114. Tabassum, R. *et al.* Genetic architecture of human plasma lipidome and its link to cardiovascular disease. *Nat. Commun.* **10**, 4329 (2019).

115. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).

116. Willer, C. J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).

117. Draisma, H. H. M. *et al.* Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels. *Nat. Commun.* **6**, 7208 (2015).

118. Ligthart, S. *et al.* Bivariate genome-wide association study identifies novel pleiotropic loci for lipids and inflammation. *BMC Genomics* **17**, 443 (2016).

119. van Setten, J. *et al.* Genome-wide association meta-analysis of 30,000 samples identifies seven novel loci for quantitative ECG traits. *Eur. J. Hum. Genet.* **27**, 952–962 (2019).

120. Bentley, A. R. *et al.* Multi-ancestry genome-wide gene-smoking interaction study of 387,272 individuals identifies new loci associated with serum lipids. *Nat. Genet.* **51**, 636–648 (2019).

121. Gallois, A. *et al.* A comprehensive study of metabolite genetics reveals strong pleiotropy and heterogeneity across time and context. *Nat. Commun.* **10**, 4788 (2019).

122. Olafsdottir, T. A. *et al.* Eighty-eight variants highlight the role of T cell regulation and airway remodeling in asthma pathogenesis. *Nat. Commun.* **11**, 393 (2020).

123. de Vries, P. S. *et al.* Comparison of HapMap and 1000 Genomes Reference Panels in a Large-Scale Genome-Wide Association Study. *PLoS One* **12**, e0167742 (2017).

124. Hoffmann, T. J. *et al.* Genome-wide association analyses using electronic health records identify new loci influencing blood pressure variation. *Nat. Genet.* **49**, 54–64 (2017).

125. Tian, C. *et al.* Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. *Nat. Commun.* **8**, 599 (2017).

126. Pilling, L. C. *et al.* Human longevity: 25 genetic loci associated in 389,166 UK biobank participants. *Aging* **9**, 2504–2520 (2017).

127. Todd, J. A. *et al.* Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat. Genet.* **39**, 857–864 (2007).

128. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).

129. Cooper, J. D. *et al.* Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nat. Genet.* **40**, 1399–1401 (2008).

130. Liu, C. *et al.* Meta-analysis identifies common and rare variants influencing blood pressure and overlapping with metabolic trait loci. *Nat. Genet.* **48**, 1162–1170 (2016).

131. Feitosa, M. F. *et al.* Novel genetic associations for blood pressure identified via gene-alcohol interaction in up to 570K individuals across multiple ancestries. *PLoS One* **13**, e0198166 (2018).

132. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341

(2018).

133. Wuttke, M. *et al.* A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nat. Genet.* **51**, 957–972 (2019).

134. Newton-Cheh, C. *et al.* Genome-wide association study identifies eight loci associated with blood pressure. *Nat. Genet.* **41**, 666–676 (2009).

135. Dubois, P. C. A. *et al.* Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.* **42**, 295–302 (2010).

136. Köttgen, A. *et al.* New loci associated with kidney function and chronic kidney disease. *Nat. Genet.* **42**, 376–384 (2010).

137. Zhernakova, A. *et al.* Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. *PLoS Genet.* **7**, e1002004 (2011).

138. Wain, L. V. *et al.* Genome-wide association study identifies six new loci influencing pulse pressure and mean arterial pressure. *Nat. Genet.* **43**, 1005–1011 (2011).

139. Köttgen, A. *et al.* Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. *Nat. Genet.* **45**, 145–154 (2013).

140. Medici, M. *et al.* Identification of novel genetic Loci associated with thyroid peroxidase antibodies and clinical thyroid disease. *PLoS Genet.* **10**, e1004123 (2014).

141. Onengut-Gumuscu, S. *et al.* Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat. Genet.* **47**, 381–386 (2015).

142. Fischer, A. *et al.* Identification of Immune-Relevant Factors Conferring Sarcoidosis Genetic Risk. *Am. J. Respir. Crit. Care Med.* **192**, 727–736 (2015).

143. Nikpay, M. *et al.* A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* **47**, 1121–1130 (2015).

144. Kato, N. *et al.* Trans-ancestry genome-wide association study identifies 12

genetic loci influencing blood pressure and implicates a role for DNA methylation. *Nat. Genet.* **47**, 1282–1293 (2015).

145. de Lange, K. M. *et al.* Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* **49**, 256–261 (2017).

146. Langefeld, C. D. *et al.* Transancestral mapping and genetic load in systemic lupus erythematosus. *Nat. Commun.* **8**, 16021 (2017).

147. Johansson, Å., Rask-Andersen, M., Karlsson, T. & Ek, W. E. Genome-wide association analysis of 350 000 Caucasians from the UK Biobank identifies novel loci for asthma, hay fever and eczema. *Hum. Mol. Genet.* **28**, 4022–4041 (2019).

148. Han, Y. *et al.* Genome-wide analysis highlights contribution of immune system pathways to the genetic architecture of asthma. *Nat. Commun.* **11**, 1776 (2020).

149. Hernandez Cordero, A. I. *et al.* Genome-wide Associations Reveal Human-Mouse Genetic Convergence and Modifiers of Myogenesis, CPNE1 and STC2. *Am. J. Hum. Genet.* **105**, 1222–1236 (2019).

150. Pulit, S. L. *et al.* Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. *Hum. Mol. Genet.* **28**, 166–174 (2019).

151. Nalls, M. A. *et al.* Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat. Genet.* **46**, 989–993 (2014).

152. Pickrell, J. K. *et al.* Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* **48**, 709–717 (2016).

153. Chang, D. *et al.* A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. *Nat. Genet.* **49**, 1511–1516 (2017).

154. Martin, A. R. *et al.* An Unexpectedly Complex Architecture for Skin Pigmentation in Africans. *Cell* **171**, 1340–1353.e14 (2017).

155. Jonnalagadda, M. *et al.* A Genome-Wide Association Study of Skin and Iris Pigmentation among Individuals of South Asian Ancestry. *Genome Biol. Evol.* **11**,

1066–1076 (2019).

156. Sulem, P. *et al.* Sequence variants at CYP1A1-CYP1A2 and AHR associate with coffee consumption. *Hum. Mol. Genet.* **20**, 2071–2077 (2011).

157. Coffee and Caffeine Genetics Consortium *et al.* Genome-wide meta-analysis identifies six novel loci associated with habitual coffee consumption. *Mol. Psychiatry* **20**, 647–656 (2015).

158. Cornelis, M. C. *et al.* Genome-wide association study of caffeine metabolites provides new insights to caffeine metabolism and dietary caffeine-consumption behavior. *Hum. Mol. Genet.* **25**, 5472–5482 (2016).

159. Haas, M. E. *et al.* Genetic Association of Albuminuria with Cardiometabolic Disease and Blood Pressure. *Am. J. Hum. Genet.* **103**, 461–473 (2018).

160. Liu, M. *et al.* Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat. Genet.* **51**, 237–244 (2019).

161. Zanetti, D. *et al.* Identification of 22 novel loci associated with urinary biomarkers of albumin, sodium, and potassium excretion. *Kidney Int.* **95**, 1197–1208 (2019).

162. Pardiñas, A. F. *et al.* Pharmacogenomic Variants and Drug Interactions Identified Through the Genetic Analysis of Clozapine Metabolism. *Am. J. Psychiatry* **176**, 477–486 (2019).

163. Zhong, V. W. *et al.* A genome-wide association study of bitter and sweet beverage consumption. *Hum. Mol. Genet.* **28**, 2449–2457 (2019).

164. Pazoki, R. *et al.* GWAS for urinary sodium and potassium excretion highlights pathways shared with cardiovascular traits. *Nat. Commun.* **10**, 3653 (2019).

165. Karlsson, T. *et al.* Contribution of genetics to visceral adiposity and its relation to cardiovascular and metabolic disease. *Nat. Med.* **25**, 1390–1395 (2019).

166. Cornelis, M. C. *et al.* Genome-wide meta-analysis identifies regions on 7p21 (AHR) and 15q24 (CYP1A2) as determinants of habitual caffeine consumption. *PLoS*

3844 *Genet.* **7**, e1002033 (2011).
3845 167. Amin, N. *et al.* Genome-wide association analysis of coffee drinking suggests
3846 association with CYP1A1/CYP1A2 and NRCAM. *Mol. Psychiatry* **17**, 1116–1129 (2012).
3847 168. Teumer, A. *et al.* Genome-wide association meta-analyses and fine-mapping
3848 elucidate pathways influencing albuminuria. *Nat. Commun.* **10**, 4130 (2019).
3849 169. Wu, Y. *et al.* Genome-wide association study of medication-use and
3850 associated disease in the UK Biobank. *Nat. Commun.* **10**, 1891 (2019).
3851 170. Eriksson, N. *et al.* Web-based, participant-driven studies yield novel genetic
3852 associations for common traits. *PLoS Genet.* **6**, e1000993 (2010).
3853 171. Zhao, B. *et al.* Large-scale GWAS reveals genetic architecture of brain white
3854 matter microstructure and genetic overlap with cognitive and mental health traits (n =
3855 17,706). *Mol. Psychiatry* (2019) doi:10.1038/s41380-019-0569-z.
3856 172. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent
3857 positive selection in the human genome. *PLoS Biol.* **4**, e72 (2006).
3858 173. Sabeti, P. C. *et al.* Genome-wide detection and characterization of positive
3859 selection in human populations. *Nature* **449**, 913–918 (2007).
3860 174. Pickrell, J. K. *et al.* Signals of recent positive selection in a worldwide sample
3861 of human populations. *Genome Res.* **19**, 826–837 (2009).
3862 175. Buckley, M. T. *et al.* Selection in Europeans on Fatty Acid Desaturases
3863 Associated with Dietary Changes. *Mol. Biol. Evol.* **34**, 1307–1318 (2017).
3864 176. Ye, K., Gao, F., Wang, D., Bar-Yosef, O. & Keinan, A. Dietary adaptation of
3865 FADS genes in Europe varied across time and geography. *Nat Ecol Evol* **1**, 167 (2017).
3866 177. Mathieson, S. & Mathieson, I. FADS1 and the Timing of Human Adaptation to
3867 Agriculture. *Mol. Biol. Evol.* **35**, 2957–2970 (2018).
3868 178. Ju, D. & Mathieson, I. The evolution of skin pigmentation-associated variation
3869 in West Eurasia. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
3870 179. Stefansson, H. *et al.* A common inversion under selection in Europeans. *Nat.*
3871 *Genet.* **37**, 129–137 (2005).

180. Boettger, L. M., Handsaker, R. E., Zody, M. C. & McCarroll, S. A. Structural haplotypes and recent evolution of the human 17q21.31 region. *Nat. Genet.* **44**, 881–885 (2012).

181. Itan, Y., Powell, A., Beaumont, M. A., Burger, J. & Thomas, M. G. The origins of lactase persistence in Europe. *PLoS Comput. Biol.* **5**, e1000491 (2009).

182. Antelope, C. X., Marnetto, D., Casey, F. & Huerta-Sanchez, E. Leveraging Multiple Populations across Time Helps Define Accurate Models of Human Evolution: A Reanalysis of the Lactase Persistence Adaptation. *Hum. Biol.* **89**, 81–97 (2017).

183. Segurel, L. *et al.* Why and when was lactase persistence selected for? Insights from Central Asian herders and ancient DNA. *PLoS Biol.* **18**, e3000742 (2020).

184. Burger, J. *et al.* Low Prevalence of Lactase Persistence in Bronze Age Europe Indicates Ongoing Strong Selection over the Last 3,000 Years. *Curr. Biol.* **30**, 4307–4315.e13 (2020).

4b) Detangling Direct and Indirect impacts of sample age from the Mesolithic-Neolithic data on genotype imputation

Rasmus Amund Henriksen[1], Rasmus Nielsen[1,2], Lei Zhao[1], Thorfinn Sand Korneliussen[1]

[1]Lundbeck Foundation GeoGenetics Centre, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark
[2]Department of Integrative Biology, University of California, Berkeley

Many factors can influence the genotype imputation/calling at specific sites, especially when present-day reference panels are used to impute missing genotypes for ancient samples. The unique features of aDNA relative to modern DNA include 1) relatively shallower read depths [1], 2) shorter read lengths [2], and 3) different error profiles [3]. Mapping biases that depend on read length and error rates are of particular concern for aDNA and may cause spurious signals of selection. Naturally, read depth may also affect genotype calling. In order to filter SNPs that might be affected by such biases, and to generally correct for and quantify the biases, we develop a causal inference method for distinguishing direct effects of sample age on allele frequency from other (indirect) effects mediated by age-dependent errors, read depth, and read length.

As shown in Figure 4b.1, to distinguish the true selection signal of age on allele frequency, from the signal caused by mapping biases, and other biases, in the ancient samples, we create a model and a workflow which can decompose the influence of sample age (A_i, age of individual *i*) on genotype (G_{ij}, Genotype of individual *i* at position *j*), into its indirect and direct effects. The imputed genotype for each individual is converted into allele frequencies, representing the homozygous for the reference allele, heterozygous and homozygous for the alternative allele as 0, 0.5 and 1 respectively. The indirect effects are mediated through the three unique features of aDNA as previously described, while the direct impact reflects the true change in allele frequency over the time range from the ancient samples to the modern ones. The three factors are the mean read depth across all sites per individual, the mean read length for each individual at each site, and the third is an overall error estimate for each individual. These three factors are all identified using ANGSD [4] from the aligned BAM files for all individuals used in the study. The depth and length are calculated based solely on the imputed positions (*j*) extracted from the imputed vcf files, whereas the overall error is calculated as described in Orlando et al. 2013 from the entire BAM file containing all the invariant sites.

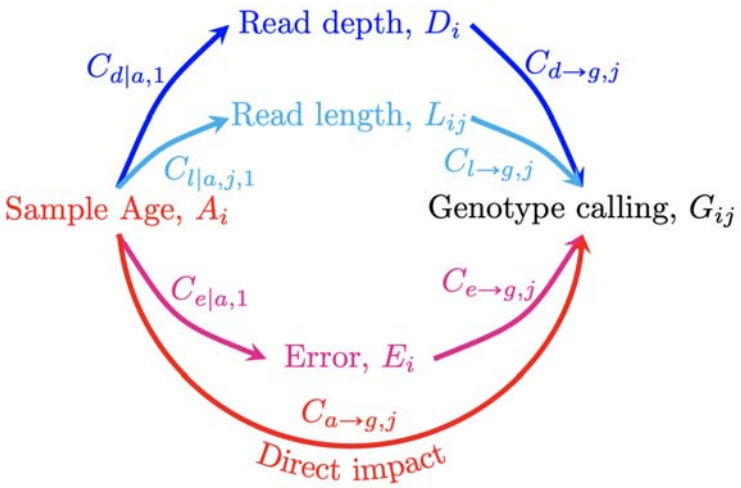


Figure S4b.1. Illustrative figure of the factors that influence genotype calling procedure. The denotations with symbol "|" can be obtained by regression, while the denotations with symbol "→" are derived after PCA.

The initial step is to conduct three linear regressions, one for each of the three factors, mean depth (equation 1), mean length (equation 2) and error estimate (equation 3) with respect to one explanatory variable, i.e., sample age (i.e., A_i). This step is to investigate the influences of sample age on each of the three factors.

$$D_i \triangleq \overline{D_{ij}} \sim C_{d|a,1}A_i + C_{d|a,0} + \epsilon_{d|a,i} \quad (1)$$

$$L_{ij} \triangleq \overline{L_{ijk}} \sim C_{l|a,j,1}A_i + C_{l|a,j,0} + \epsilon_{l|a,ij} \quad (2)$$

$$E_i \sim C_{e|a,1}A_i + C_{e|a,0} + \epsilon_{e|a,i}. \quad (3)$$

Where $\triangleq$ means definition, $\overline{D_{ij}}$ means the average depth with regard to *j* for a specific individual *i*, and $\overline{L_{ijk}}$ means the average read length, for all reads *k* stretching over site *j* of individual *i*. $E_i$means the overall error rate of individual *i*.

We assume that the sample age influences these factors, as such it is necessary to eliminate this influence as done in equation 4-6.

$$\Delta_i \triangleq D_i - C_{d|a,1}A_i \quad (4)$$

$$\Lambda_{ij} \triangleq L_{ij} - C_{l|a,j,1}A_i \quad (5)$$

$$\Sigma_i \triangleq E_i - C_{e|a,1}A_i \quad (6)$$

A second round of regression is necessary to detect the direct impact of the three factors as well as the direct impact of age on the genotype. The explanatory variables can be the

sample age (A_i) and the three remainders (or the linear combinations of the three remainders). To avoid potential correlations between the three remainders, we perform a Principal Component Analysis. The relationship between the PCA scores and the remainders can be represented as in equation 7 and 8.

$$\left(\frac{\Lambda_{ij}-\Lambda_j}{\mathrm{sd}(\Lambda_{.j})},\frac{\Sigma_i-\Sigma}{\mathrm{sd}(\Sigma_.)},\frac{\Delta_i-\Delta}{\mathrm{sd}(\Delta_.)}\right)_{n\times 3}=(\alpha_{ij}\,\beta_{ij},\gamma_{ij})_{n\times 3}\begin{pmatrix}\overrightarrow{\omega}'_{j,1}\\ \overrightarrow{\omega}'_{j,2}\\ \overrightarrow{\omega}'_{j,3}\end{pmatrix} \quad (7)$$

$$(\alpha_{ij}\,\beta_{ij},\gamma_{ij})_{n\times 3}=\left(\frac{\Lambda_{ij}-\Lambda_j}{\mathrm{sd}(\Lambda)},\frac{\Sigma_i-\Sigma}{\mathrm{sd}(\Sigma_.)},\frac{\Delta_i-\Delta}{\mathrm{sd}(\Delta_.)}\right)_{n\times 3}(\overrightarrow{\omega}_{j,1},\overrightarrow{\omega}_{j,2},\overrightarrow{\omega}_{j,3}) \quad (8)$$

Where$\Lambda_j$and sd($\Lambda_{.j}$) are the mean and the standard deviation of Λ across individuals *i* at fixed position *j*. $\Sigma$and sd($\Sigma_.$) are the mean and the standard deviation of Σ across individuals *i*. $\Delta$and sd($\Delta_.$) are the mean and the standard deviation of Δacross individuals *i*. $\overrightarrow{\omega}_{j,2}$, $\overrightarrow{\omega}_{j,1}$, $\overrightarrow{\omega}_{j,3}$ are the three principle directions (column eigenvectors) and the $\overrightarrow{\omega}'_{j,2}$, $\overrightarrow{\omega}'_{j,1}$,$\overrightarrow{\omega}'_{j,3}$ are the corresponding transposed vectors. $\alpha_{ij},\beta_{ij},\gamma_{ij}$ are the principal component scores of $\left(\frac{\Lambda_{ij}-\Lambda_j}{sd(\Lambda_{.j})},\frac{\Sigma_i-\Sigma}{sd(\Sigma_.)},\frac{\Delta_i-\Delta}{sd(\Delta_.)}\right)$, with *n* being the number of individuals carrying site *j*.

Once all the explanatory variables are independent, the coefficient $C_{g|a,j,1}$represent the total impact of age which can be obtained by conducting a final round of regression (equation 9).

$$G_{ij}\sim C_{g|\alpha,j,1}\alpha_{ij}+C_{g|\beta,j,1}\beta_{ij}+C_{g|\gamma,j,1}\gamma_{ij}+C_{g|a,j,1}A_i+C_{0,j}+\epsilon_{ij} \quad (9)$$

The sum of the first three terms of equation 9 can be obtained by multiplying the coefficients $(C_{g|\alpha,j,1},C_{g|\beta,j,1},C_{g|\gamma,j,1})'$ with equation 8 (shown in equation 10). And the effective regression coefficients for $(\Delta_{ij},\Lambda_i,\Sigma_i)$ will be observed when calculating the linear slopes of the remainders in Equation 10. Such effective coefficients can be viewed as measurements of the direct impact of the corresponding factors, i.e. length ($C_{l\to g,j}$), error ($C_{e\to g,j}$) and depth ($C_{d\to g,j}$), on the genotype calling (equation 11 - 13).

$$(\alpha_{ij}\,\beta_{ij},\gamma_{ij})_{n\times 3}\begin{pmatrix}C_{g|\alpha,j,1}\\ C_{g|\beta,j,1}\\ C_{g|\gamma,j,1}\end{pmatrix}=\left(\frac{\Lambda_{ij}-\Lambda_j}{\mathrm{sd}(\Lambda_{.j})},\frac{\Sigma_i-\Sigma}{\mathrm{sd}(\Sigma_.)},\frac{\Delta_i-\Delta}{\mathrm{sd}(\Delta_.)}\right)_{n\times 3}(\overrightarrow{\omega}_{j,1},\overrightarrow{\omega}_{j,2},\overrightarrow{\omega}_{j,3})_{3\times 3}\begin{pmatrix}C_{g|\alpha,j,1}\\ C_{g|\beta,j,1}\\ C_{g|\gamma,j,1}\end{pmatrix} \quad (10)$$

$$\vec{\omega}_{j,1} = \begin{pmatrix} \omega_{j,1,1} \\ \omega_{j,2,1} \\ \omega_{j,3,1} \end{pmatrix}, \vec{\omega}_{j,2} = \begin{pmatrix} \omega_{j,1,2} \\ \omega_{j,2,2} \\ \omega_{j,3,2} \end{pmatrix}, \vec{\omega}_{j,3} = \begin{pmatrix} \omega_{j,1,3} \\ \omega_{j,2,3} \\ \omega_{j,3,3} \end{pmatrix}$$

$$\mathrm{C}_{l\to g,j} = \frac{1}{\mathrm{sd}(\Lambda_{.j})}\left(\omega_{j,1,1}C_{g|\alpha,j,1} + \omega_{j,1,2}C_{g|\beta,j,1} + \omega_{j,1,3}C_{g|\gamma,j,1}\right) \quad (11)$$

$$\mathrm{C}_{e\to g,j} = \frac{1}{\mathrm{sd}(\Sigma_{.})}\left(\omega_{j,2,1}C_{g|\alpha,j,1} + \omega_{j,2,2}C_{g|\beta,j,1} + \omega_{j,2,3}C_{g|\gamma,j,1}\right) \quad (12)$$

$$\mathrm{C}_{d\to g,j} = \frac{1}{\mathrm{sd}(\Delta_{.})}\left(\omega_{j,3,1}C_{g|\alpha,j,1} + \omega_{j,3,2}C_{g|\beta,j,1} + \omega_{j,3,3}C_{g|\gamma,j,1}\right) \quad (13)$$

Any influence of age on the genotype calling imposed through one of the three factors (depth, length, error) is an indirect impact. The measurements of such indirect impacts are calculated by multiplying equations 11-13 with each of the factors corresponding coefficients obtained from our first round of regression (equation 4-6) for each site *j*.

$$\mathrm{C}_{a\to l\to g,j} = C_{l|a,j,1} \cdot \mathrm{C}_{l\to g,j} \quad (14)$$

$$\mathrm{C}_{a\to e\to g,j} = C_{e|a,1} \cdot \mathrm{C}_{e\to g,j} \quad (15)$$

$$\mathrm{C}_{a\to d\to g,j} = C_{d|a,1} \cdot \mathrm{C}_{d\to g,j} \quad (16)$$

Finally, the direct impact of age on the genotype calling at site *j* (equation 17) can be obtained by subtracting all indirect impacts of age (equation 14-16) from the total impact of age $C_{g|a,j,1}$, obtained from the second round of regression (equation 9).

$$C_{a\to g,j} = C_{g|a,j,1} - \mathrm{C}_{a\to l\to g,j} - \mathrm{C}_{a\to e\to g,j} - \mathrm{C}_{a\to d\to g,j} \quad (17)$$

Then to apply all these measurements, we calculate a ratio of the sum of all possible indirect effects and the direct effect for each site (i.e., R_j, equation 18). This ratio can be used to quantify the relative contribution of indirect and direct effects on changes in allele frequency. If the ratio is negative or close to zero, an observed change in allele frequency cannot solely be attributed to biases, and selection at such sites was likely to occur during the course of evolution. While a relatively large positive ratio indicates that the potential selection is indistinguishable with the effects of mapping biases and/or other biases, thus we will filter out such sites while detecting selection signals.

$$R_j = \frac{\mathrm{C}_{a\to l\to g,j} + \mathrm{C}_{a\to e\to g,j} + \mathrm{C}_{a\to d\to g,j}}{C_{a\to g,j}} \quad (18)$$

R_j can also be converted into a fraction representing the proportion of indirect effects on age relative to the total effect (equation 19), which can equivalently be used for filtering:

$$F_j = \frac{R_j}{1 + R_j} = \frac{1}{1 + \frac{1}{R_j}} \qquad (19)$$

To filter out sites in selection analyses that may be affected by biases, we use a fixed threshold of $0.5 < F_j \leq 1$.

References

1. Shapiro, B. & Hofreiter, M. A paleogenomic perspective on evolution and gene function: new insights from ancient DNA. *Science* **343**, 1236573 (2014).

2. Paabo, S. Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. *Proceedings of the National Academy of Sciences* vol. 86 1939–1943 (1989).

3. Orlando, L. *et al.* Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* **499**, 74–78 (2013).

4. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15**, 356 (2014).

4c) Over-dispersion in polygenic scores across ancient populations

Alba Refoyo Martínez[1], Fernando Racimo[1]

[1]Lundbeck Foundation GeoGenetics Centre, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark

Introduction

We aimed to test whether there was evidence for over-dispersion in polygenic scores across different ancient populations. This could indicate if there was evidence for strong differentiation in the genetics of traits across ancient populations, beyond what can be explained by genetic drift alone.

Methods

We used summary statistics from GWASs performed on the UK Biobank cohort [1] by the Neale lab (Round 2: http://www.nealelab.is/uk-biobank/), which include both quantitative and case-control traits. From the case-control case traits, we filtered out those where the N-case/N ratio is lower than 1/100 and those that did not have any associated variants at the standard genome-wide significance threshold (P = 5e-8). We then filtered out variants with minor allele frequency lower than 5%, variants with INFO score < 50%, variants with a genotype probability lower than 0.8 in more than 10% of the individuals, triallelic variants and variants flagged as “low-confident” after imputation (Supplementary Note S2).

We divided the genome into 1,703 non-overlapping and approximately independent linkage disequilibrium (LD) blocks [2]. From each block, we retrieved the variant with the lowest association P-value. We then selected only the variants with a P-value lower than the genome-wide significance threshold, 5e-8, for downstream analyses.

Polygenic scores were calculated by summing over the allele frequency of the filtered trait-associated variants and weighting them by the effect size obtained from the summary statistics from the UK Biobank (round 2). The allele frequency was retrieved from the imputed ancient West-Eurasian individuals. Individuals were clustered into groups of closely-related individuals using MDS (see Supplementary Note S3). In our overdispersion test, we did not include genomes that were not successfully classified as belonging to any of the clusters, and removed clusters that contained less than four individuals. This resulted in 1,119 individuals clustered into 41 groups.

We evaluated our choice of the summary statistics by comparing the Neale Lab scores to scores obtained using a different GWAS on the same phenotype and on the same cohort - the GWAS ATLAS [3,4] (https://atlas.ctglab.nl/). Despite both association tests being performed on the same cohort, the two studies applied different filters on the data, which resulted in differences in the effect size estimates and in the significance of the SNP associations (Figure **S4c.1**, panel C), which in turn lead to discrepancies on the polygenic scores, particularly for pre-Neolithic populations, like the Western hunter-gatherers, and for Eastern

Eurasian individuals, like the ancient Siberians, which are more distantly related to the present-day British individuals included in the GWAS cohorts (Figure **S4c.1**, panel B). We followed the Sohail et al. [5] procedure to detect residual population stratification along the axes of population variation by looking for strong correlations between axes of population structure and the magnitude of effect size estimates. We found that the GWAS ATLAS had substantially more uncorrected population stratification than the Neale Lab GWAS (Figure **S4c.1**, panel A). Therefore, all results presented below were based on the scores obtained using effect sizes from the Neale Lab (round 2) estimates. We also observed that polygenic scores for pre-Neolithic individuals were particularly sensitive to the choice of cohort and to the SNP filtering scheme, so we urge caution in the interpretation of those values.

The Q_X statistic was introduced by [6] to look for overdispersion in polygenic scores across populations that cannot be explained simply by genetic drift, assuming the polygenic scores were not biassed by population stratification in the GWAS cohort from which effect size estimates were obtained. Polygenic scores are also assumed to follow a multivariate normal distribution under a null model of genetic drift, and this statistic serves to look for departures from this model. To compute the Q_X statistic, an empirical genome-wide covariance matrix is needed, which we constructed using a subset of SNPs with a trait-association p-value larger than 5e-8, and then we sampled every 20th “non-associated” SNPs across the genome.

To further test the significance of the overdispersion and account for possible deviations from the assumptions the Q_X statistic makes, we also computed P-values using two randomization schemes simulating a neutral scenario. The first method was based on the randomization of the signs of the effect size estimates of trait-associated SNPs while the other was based on sampling variants across the genome with frequencies matching those of trait-associated variants in the GBR panel from the 1000 Genomes Project.

To address mapping biases we performed a one-tailed wilcoxon rank-sum test for each trait. We evaluated if the candidate associated SNPs had higher values of the artefactual effect estimates than the non-associated SNPs used as our neutral baseline (Supplementary Note S4b). None of the tests were significant (min-P value = 0.19; max-P Value = 0.99; Mean=0.56).

Results

Polygenic scores across ancient populations

We computed polygenic scores of trait-associated SNPs across the 41 ancient population groups. We used the effect size estimates from the UK Biobank Neale lab GWAS [1] and the allele frequencies from the 41 population clusters previously described (Supplementary Note S3). We filtered out those traits that had less than 10 genome-wide significantly associated SNPs, restricting our analysis to a total of 320 polygenic traits. We then used the Q_X statistic on these traits to test for overdispersion across clusters.

We applied Q_X statistic to each of the 320 traits. Figure **S4c.2** shows p-values for Q_X statistic for the standard genome-wide set of trait-associated SNPs ($P < 5e-8$). From these, 119 resulted in a nominally significant Q_X statistic and only 39 remained significantly over-dispersed after controlling for multiple testing via a Bonferroni correction (Figure **S4c.3**). We grouped these 39 traits into ten broad categories in agreement with different sub-categories levels from the Data Showcase (https://biobank.ctsu.ox.ac.uk/showcase/): “body measurements”, “impedance measures”, “spirometry”, “sun exposure: hair and skin pigmentation”, “assay results”, “diabetes, cholesterol or blood pressure”, “diet”, “medical conditions”, “mental health” and “sex-specific factors” (Figure **S4c.4**).

We also computed p-values using two randomization schemes to account for possible violations of the assumption that the Q_X statistic is chi-squared distributed: one based on randomising of the effect sizes before calculating polygenic scores and the other based on frequency-matched variants that were not significantly associated with the trait under study [6,7].

Individual scores across time and space

We also computed polygenic scores individually for each ancient genome. Figure **S4c.5** shows the ancient genomes projected into the first two principal components, colouring each genome with its corresponding polygenic score for height. Maps for each of the clusters are also available to better interpret how scores change across space and time (Figures **S4c.6** - **S4c.8**).

Differentiation among ancient populations and GBR

Finally, we were interested in determining how differentiated our ancient populations were to each other, and to a present-day British panel: the 1000 Genomes GBR panel [8]. This would

enable us to better understand how portable the polygenic scores created using the UK Biobank were to each ancient population, by drawing analogies with pairwise comparisons with known estimates of score portability between present-day populations [9,10]. In Figure **S4c.9.B** , we show genome-wide pairwise Fst estimates [11] computed between GBR and all the other present-day population panels from the 1000 Genomes Project using vcftools [12]. In turn, in Figure **S4c.9.A**, we show genome-wide pairwise Fst estimates computed between each of our ancient population clusters, as well as between each ancient population cluster and GBR. GBR shows lower Fst values when compared to IBD Global Cluster "Eurasia_5000BP_200BP" which represents modern Eurasian diversity after the major Bronze Age migrations events. Within the "Eurasia_5000BP_200BP" cluster, GBR is more genetically distant to the Steppe populations. GBR shows low values in the order of 0.01 when compared to "EuropeWCAsia_25000BP_300BP". The "EuropeWCAsia_25000BP_300BP" cluster shows west-asian populations that are thought to have brought Neolithic farming practises into Europe, as well as individuals with ancestry associated with Caucasus hunter-gatherers. This cluster shows strong differentiation when compared to "Europe_15000_4000BP". The cluster "Europe_15000_4000BP" (which includes both eastern and western hunter-gatherer populations that share more ancestry with Siberian samples) clustered together with "Asia_45000BP_2000BP". The "Asia_45000BP_2000BP" cluster represents Siberian hunter-gatherers sampled during the Neolithic and Bronze Age.

Discussion

After testing for genetic polygenic score differentiation among ancient populations, we found 39 traits with significant overdispersion in scores after controlling for multiple testing (Table **S4c.10**). Most significant traits are related to pigmentation, body size differences, disorders related to diet and sugar levels and mental health conditions. A few of these differences among ancient populations have been reported before [13–15][15] but our analysis provides a much more fine-scale account of these differences across time and space, due to the relatively higher availability of population genomic data, and across different phenotypes.

The top ten most significantly overdispersed traits are related with either hair or skin pigmentation. These phenotypes (Data-field: 1737, 2267, 1717, 1727, C44, C_SKIN and C3_SKIN) are very closely linked to sun exposure, from which the last three represent associations with malignant neoplasms of skin. Polygenic scores for these three traits tend to be higher in Bronze Age Europeans and Neolithic Farmers compared to northeastern asian and hunter-gathers groups. Skin colour polygenic scores are higher in hunter-gatherers groups (Panel C of Figure **S4c.5**). Similarly, Europeans tend to have higher polygenic scores

for blonde and light brown hair colour, while East Asian have higher scores for dark brown and black hair (Panel A and B of Figure **S4c.5**).

Intriguingly, we find that Eastern hunter-gatherers had much higher polygenic scores for height than Western hunter-gatherers, indicating that even in Mesolithic Europe, before the arrival of Neolithic farmers, there was already strong genetic differentiation at variants associated with height between local populations. We find that the Yamnaya steppe people and the Caucasus hunter-gatherers / Iranian Neolithic people generally have high genetic scores for height, while Levant Neolithic peoples, early European farmers and western hunter-gatherers have low genetic height scores (Figure **S4c.6**).

Lung capacity is very influenced by height, BMI and ethnicity. One of the most common clinical practices to measure lung capacity is forced expiratory volume in the first second (FEV-1), which happens to be one of the significant traits (Data-field: 20153) [16] (Figure **S4c.4**).

Impedance measures are used for estimating body composition, especially body fat and muscle mass [17]. The difference in height between the samples could also partially explain the distribution of polygenic scores in the impedance traits category. For instance, taller populations require a higher energy intake which is proportional to their basal metabolic rate (BMR) [18]. As we can see in Panel D of Figure **S4c.5**, farmers, who have a very similar distribution of height polygenic scores as Eastern hunter-gatherers, have lower BMR polygenic scores and also a more sedentary life [19]. The body max index (BMI) is also strongly influenced by the percentage of body fat and muscle mass. Therefore, ethnic differences in BMI-associated health risks may be caused by differences in impedance measures [17]. Height could be a confounding factor when interpreting the significance of some closely related traits.

There are a few traits directly or indirectly connected with human diets and local environment adaptation. Some of these traits are linked to sugar levels, cholesterol and blood pressure, which can increase the risk of developing a cardiovascular disease or diabetes. For instance, low levels of potassium in urine is a risk factor for cardiovascular disease and diabetes [20–22]. Abnormal low levels of mean corpuscular haemoglobin concentration have been associated with local adaptations to high altitude environments [23].

Three of the significant traits (UK Biobank codes “1940”, “2030” and “1980”) correspond to human feelings and mood instability which can have an effect on mental health. Both worrier

and nervous feelings are associated with anxiety disorders while irritability is one of the critical systems to clinically assess depression and/or bipolar affective disorders [24,25].

It is important to keep in mind that all these polygenic scores rely on effect size estimates obtained from the UK Biobank and, in particular, those individuals identifying as “white British” within that panel. This might lead to biases in the portability of these scores to distantly related ancient populations. For example, ancient Western Eurasian hunter-gatherers are as differentiated from present-day British individuals as present-day British individuals are to some of the South East Asian populations (ITU, STU, BEB, GIH, PJL) and American (MXL). This means that the expected portability of these scores in Western hunter-gatherers should be comparable to that observed in South Asians when computing polygenic scores using European effect size estimates [9,10].

We also note that the value of polygenic scores and the magnitude of the Q_X statistic may both be affected by population stratification in the GWAS panel from which effect size estimates were obtained. This seems to be less of a problem in UK Biobank GWAS than in other GWAS based on meta-analyses [5,26,27] but we should nevertheless be cautious about conclusions drawn purely from these effect size estimates.

Figures

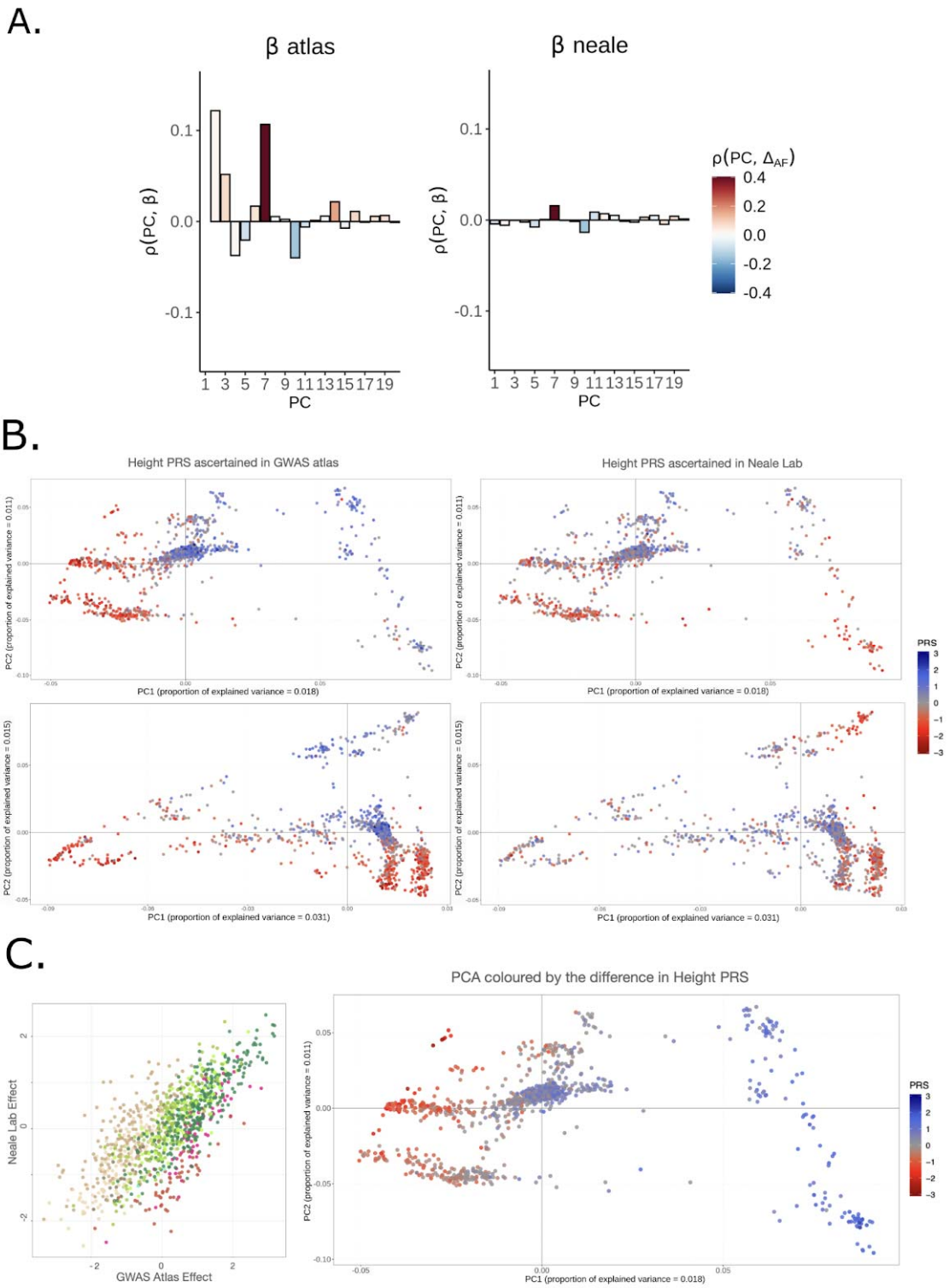


Figure S4c.1. A. Pearson correlations between 20 PC loadings and height effect size estimates from the GWAS atlas, compared to the same correlation using effect size estimates from the Neale Lab GWAS, both summary statistics performed on the UKBiobank. The correlations were computed using SNPs that are present in both the GWAS atlas and Neale

Lab GWAS summary statistics, and in the 1000 Genomes Project. The barplots are coloured by the correlation between each loading and the allele frequency difference between GBR and TSI. **B.** Height PRS scores ascertained in GWAS atlas (left panel) and in Neale Lab (right panel). **C.** Right panel: SNP-associated effect size in GWAS atlas against their effect size in the Neale Lab. Left panel: Height PRS scores are coloured by the difference between the two ascertainments (GWAS atlas scores versus Neale Lab scores.)

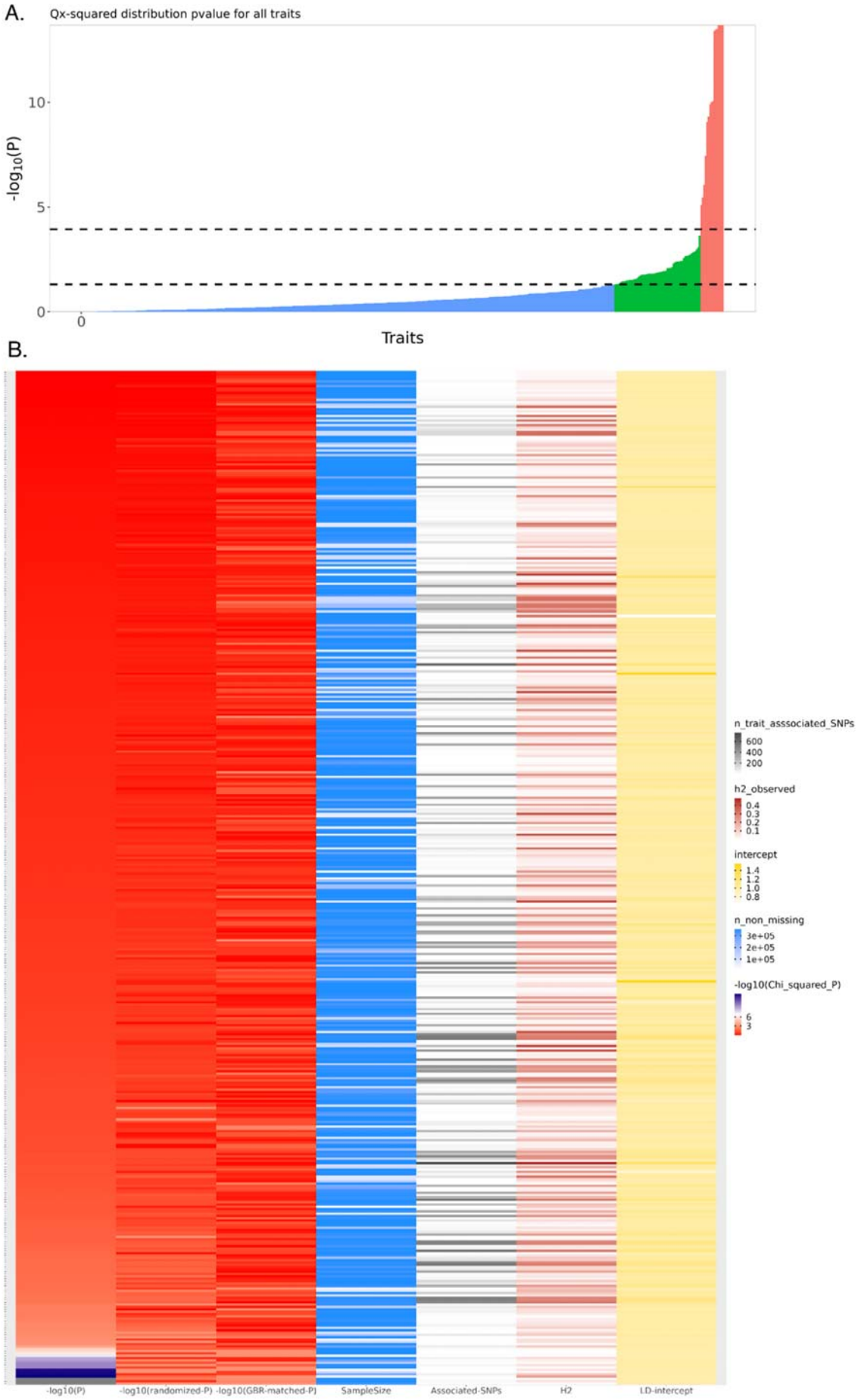

A.
Qx-squared distribution pvalue for all traits
10
5
0
-log10(P)
0
Traits
B.
n_trait_asssociated_SNPs
600
400
200
h2_observed
0.4
0.3
0.2
0.1
intercept
1.4
1.2
1.0
0.8
n_non_missing
3e+05
2e+05
1e+05
-log10(Chi_squared_P)
6
3
-log10(P)
-log10(randomized-P)
-log10(GBR-matched-P)
SampleSize
Associated-SNPs
H2
LD-intercept

Figure S4c.2. Qx p-value for 320 traits. A. Trait-associated SNPs are selected using the 5e-8 cutoff. 119 traits are significant (p-value < 0.05) and only 39 are significant after Bonferroni correction (p-value < 0.05/320). B. Heatmap includes the log values for QX p-value, randomised p-value and GBR-matched p-value, sample size, number of associated SNPs used to compute Qx statistic, heritability coefficient and LD-intercept value.

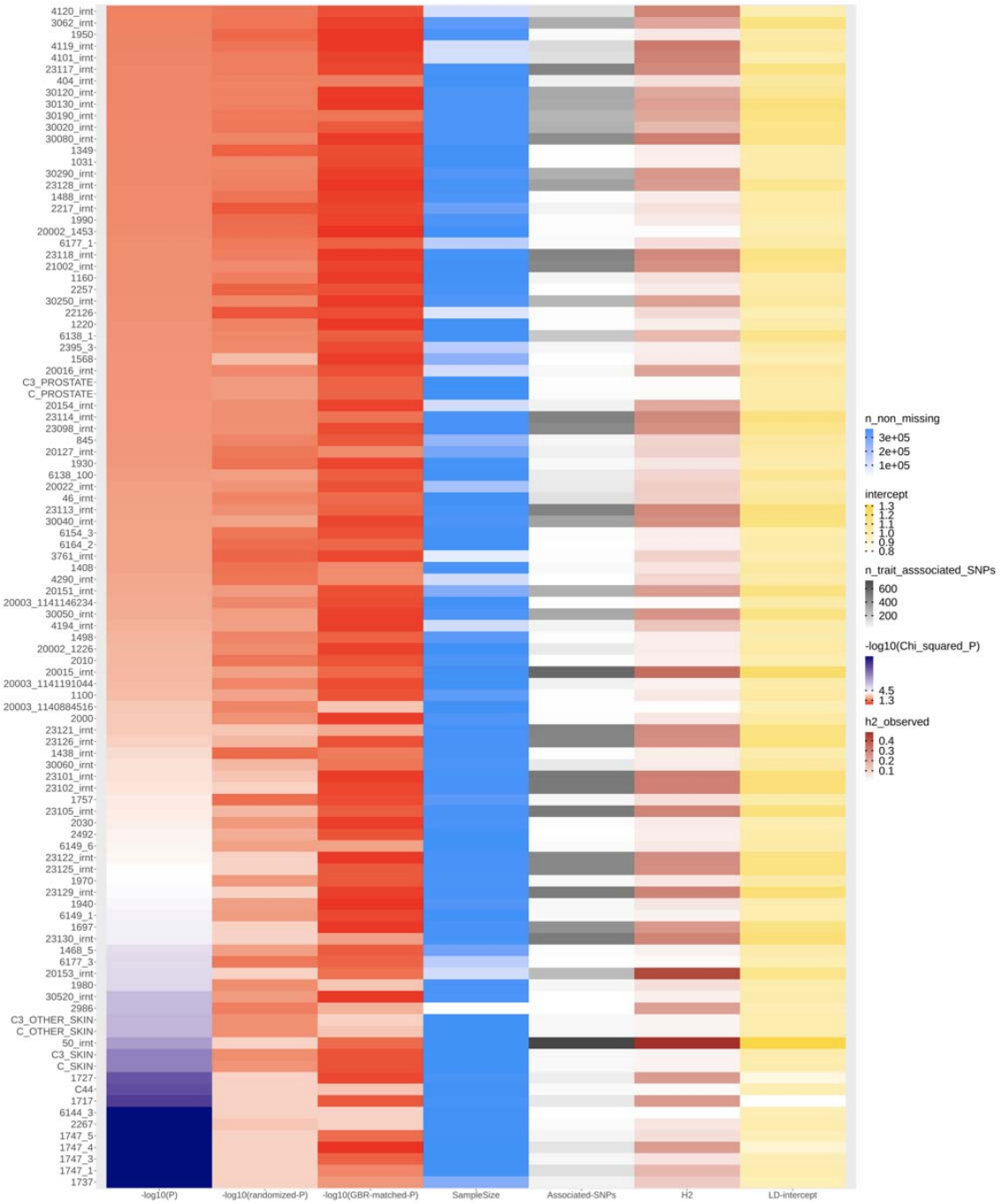


Figure S4c.3. Qx p-value for 119 significant traits. Traits are labelled with the corresponding Uk Biobank data coding system. Heatmap includes the log values for QX p-value, randomised p-value and GBR-matched p-value, sample size, number of associated SNPs used to compute Qx statistic, heritability coefficient and LD-intercept value.


Potassium in urine
Mean corpuscular haemoglobin concentration
Assay results
Standing height
Comparative height size at age 10
Body measurement
Started insulin within 1 year diagnosis of diabetes
Medication for cholesterol, blood pressure or diabetes: Insulin
Diabetes, cholesterol or blood pressure
Never eat eggs, dairy, wheat, sugar: Wheat products
Cereal type: Other (e.g. Cornflakes, Frosties)
Diet
Whole body water mass
Whole body fat-free mass
Trunk predicted mass
Trunk fat-free mass
Basal metabolic rate
Arm predicted mass (right)
Arm predicted mass (left)
Arm fat-free mass (right)
Arm fat-free mass (left)
Impedance measures
Treatment/medication code: thyroxine product
Taking other prescription medications
Mouth/teeth dental problems: Mouth ulcers
Medical conditions
Worrier / anxious feelings
Irritability
Guilty feelings
Mental health
Relative age of first facial hair
Sex-specific factors
Forced expiratory volume in 1-sec (FEV1), predicted
Spirometry
Use of sun/uv protection
Skin colour
Other malignant neoplasms of skin
Malignant neoplasm of skin
Hair colour (natural, before greying): Light brown
Hair colour (natural, before greying): Dark brown
Hair colour (natural, before greying): Blonde
Hair colour (natural, before greying): Black
Facial ageing
Ease of skin tanning
Diagnoses - main ICD10: C44 Other malignant neoplasms of skin
Childhood sunburn occasions
Sun exposure
-2
0
2
Polygenic scores
Europe: Iron, Viking & Medieval Ages
Bronze Age Europe
Steppe
EuropeNE
Faroe Modern
Caucasus Hunter-Gatherer & Iran Neolithic
Iron Age & Medieval Southern Europe
Neolithic Farmers
Late Neolithic
Eastern Hunter-Gatherer
Western Hunter-Gatherer

Figure S4c.4. Polygenic scores for 39 significant traits after Bonferroni correction grouped by traits category (trait-associated SNPs with a p-value < 5e-08). Polygenic scores for each of the different populations are shown. They are coloured by broader population groups.

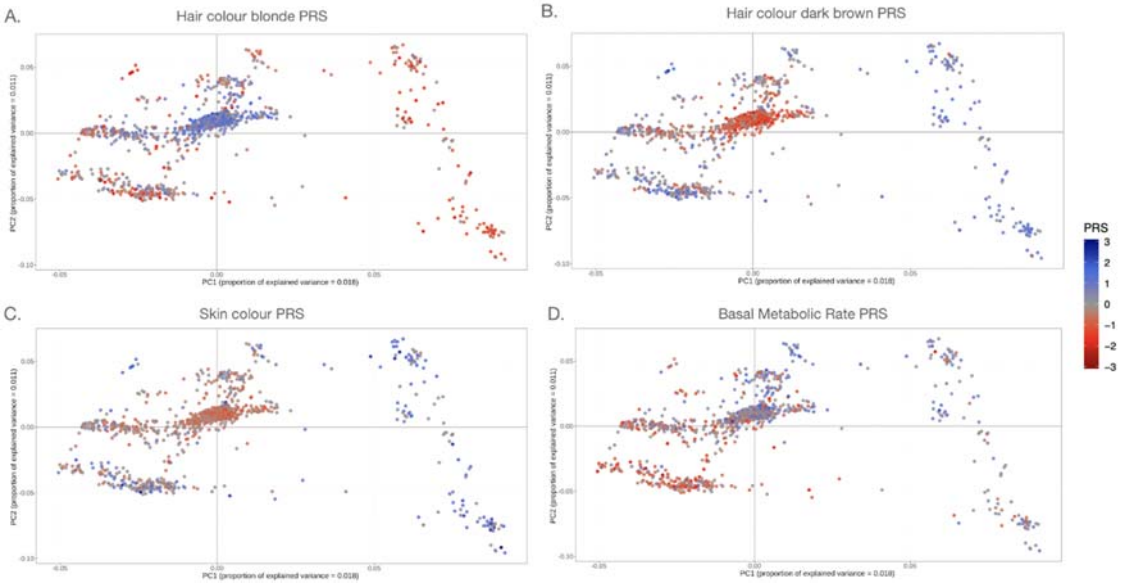


Figure S4c.5. Principal component analysis on West Eurasian samples coloured by individual polygenic scores. A-C. Polygenic scores of sun exposure traits. D. Polygenic scores of Impedance measures trait.

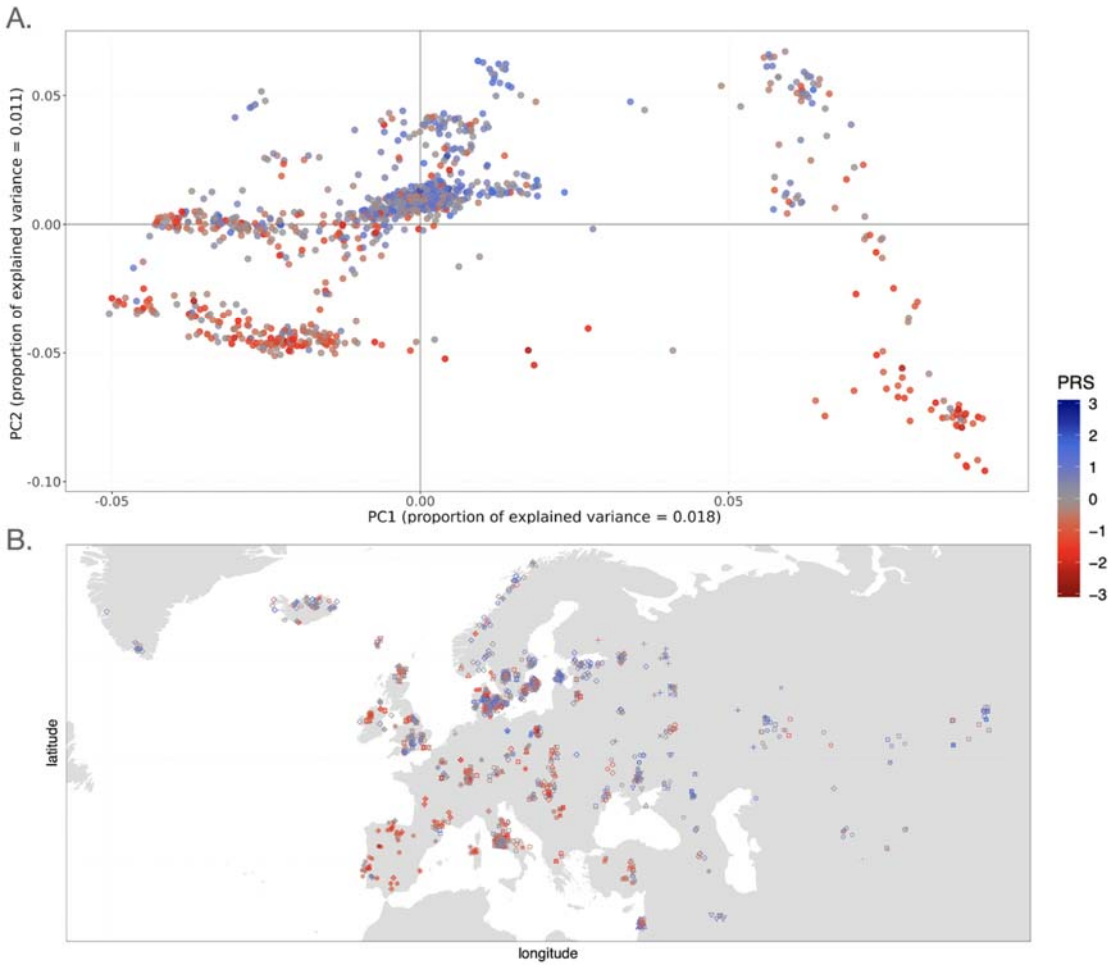


Figure S4c.6. Principal component analysis on West Eurasian samples coloured by individual polygenic scores.

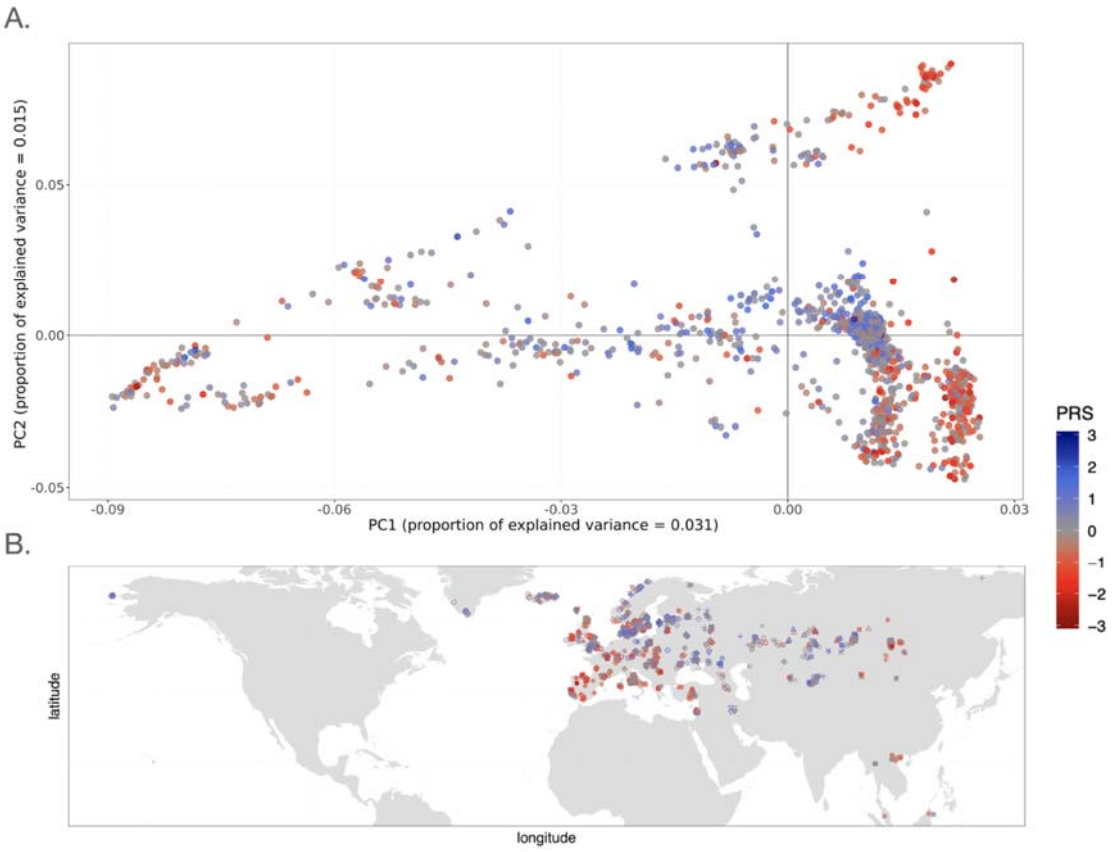


Figure S4c.7. A. **Principal component analysis on 1,332 Eurasian samples coloured by individual polygenic scores.** B. World map World map of all Eurasian samples coloured by polygenic scores for height.

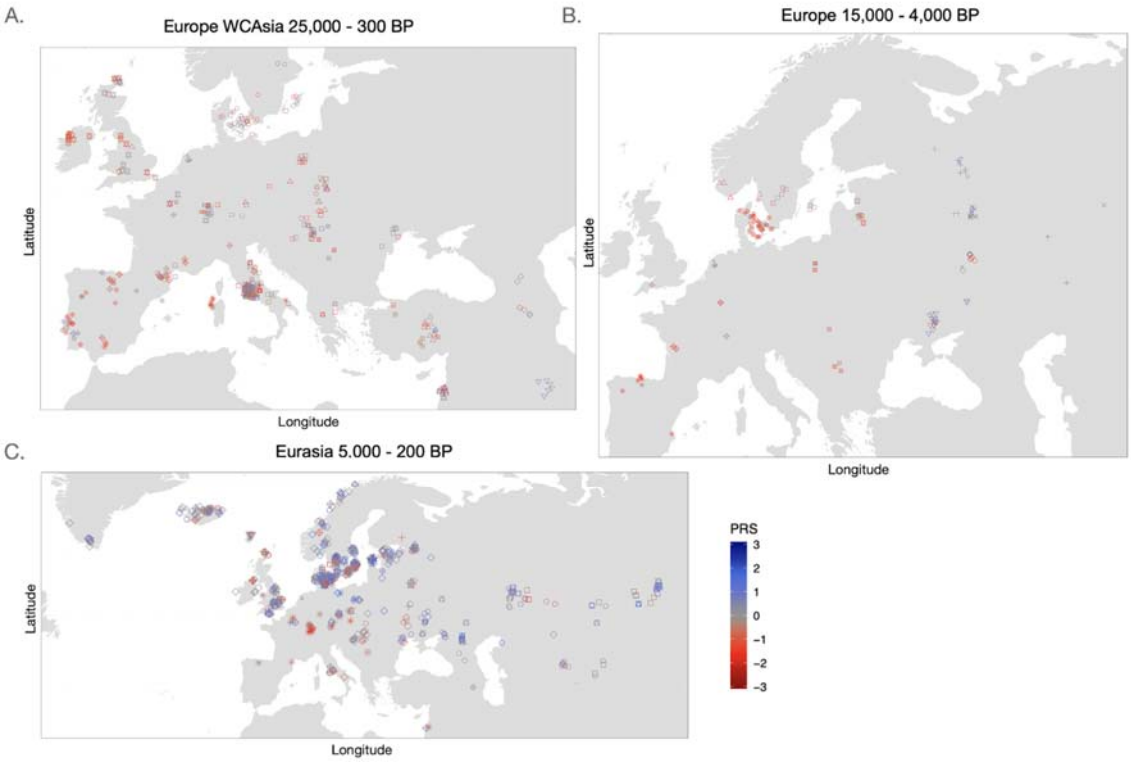


Figure S4c.8. Map of Western Eurasian polygenic scores for height.

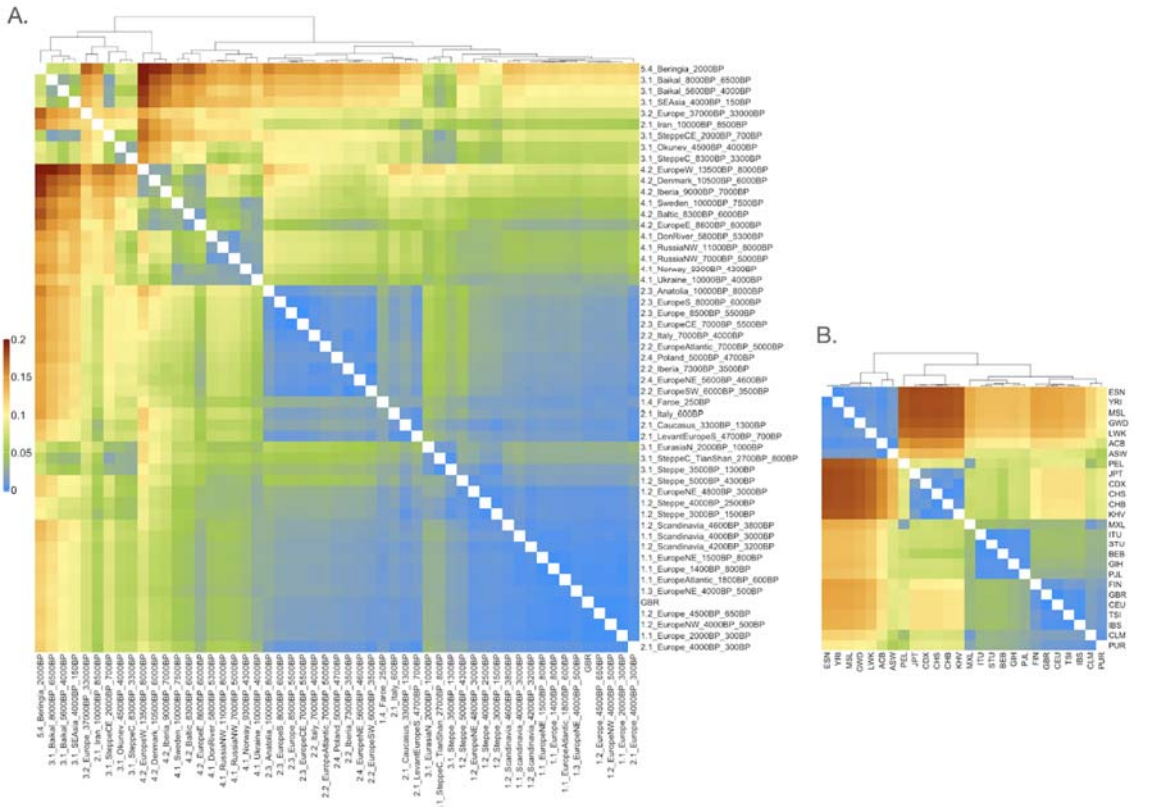


Figure S4c.9. Genetic differentiation between A. ancient clusters and B. modern population in 1000 Genomes Project data set. Pairwise Fst

Population Code	Population Description	Super Population Code
CHB	Han Chinese in Beijing, China	EAS
JPT	Japanese in Tokyo, Japan	EAS
CHS	Southern Han Chinese	EAS
CDX	Chinese Dai in Xishuangbanna, China	EAS
KHV	Kinh in Ho Chi Minh City, Vietnam	EAS
CEU	Utah Residents (CEPH) with Northern and Western European Ancestry	EUR
TSI	Toscani in Italia	EUR
FIN	Finnish in Finland	EUR
GBR	British in England and Scotland	EUR
IBS	Iberian Population in Spain	EUR
YRI	Yoruba in Ibadan, Nigeria	AFR
LWK	Luhya in Webuye, Kenya	AFR
GWD	Gambian in Western Divisions in the Gambia	AFR
MSL	Mende in Sierra Leone	AFR
ESN	Esan in Nigeria	AFR
ASW	Americans of African Ancestry in SW USA	AFR
ACB	African Caribbeans in Barbados	AFR
MXL	Mexican Ancestry from Los Angeles USA	AMR
PUR	Puerto Ricans from Puerto Rico	AMR
CLM	Colombians from Medellin, Colombia	AMR
PEL	Peruvians from Lima, Peru	AMR
GIH	Gujarati Indian from Houston, Texas	SAS
PJL	Punjabi from Lahore, Pakistan	SAS
BEB	Bengali from Bangladesh	SAS
STU	Sri Lankan Tamil from the UK	SAS
ITU	Indian Telugu from the UK	SAS

Figure S4c.10. Full population descriptions of 1000 Genomes Project panels used in our analysis.

References

1. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562, 203–209 (2018).

2. Berisa, T. & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* 32, 283–285 (2016).

3. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. doi:10.1101/500090.

4. Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in UK Biobank. *Nat. Genet.* 50, 1593–1599 (2018).

5. Sohail, M. *et al.* Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *Elife* 8, (2019).

6. Berg, J. J. & Coop, G. A population genetic signal of polygenic adaptation. *PLoS Genet.* 10, e1004412 (2014).

7. Racimo, F., Berg, J. J. & Pickrell, J. K. Detecting Polygenic Adaptation in Admixture Graphs. *Genetics* 208, 1565–1584 (2018).

8. Consortium, T. 1000 G. P. & The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* vol. 526 68–74 (2015).

9. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* 51, 584–591 (2019).

10. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am. J. Hum. Genet.* 100, 635–649 (2017).

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

12. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158 (2011).

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

14. Cox, S. L., Ruff, C. B., Maier, R. M. & Mathieson, I. Genetic contributions to variation in human stature in prehistoric Europe. *Proc. Natl. Acad. Sci. U. S. A.* 116, 21484–21492 (2019).

15. Ju, D. & Mathieson, I. The evolution of skin pigmentation associated variation in West Eurasia. doi:10.1101/2020.05.08.085274.

16. Gao, C. *et al.* Reference values for lung function screening in 10- to 81-year-old, healthy, never-smoking residents of Southeast China. *Medicine* 97, (2018).

17. Jensen, B. *et al.* Ethnic differences in fat and muscle mass and their implication for interpretation of bioelectrical impedance vector analysis. *Appl. Physiol. Nutr. Metab.* 44, 619–626 (2019).

18. Bosy-Westphal, A., Plachta-Danielzik, S., Dörhöfer, R.-P. & Müller, M. J. Short stature and obesity: positive association in adults but inverse association in children and adolescents. *Br. J. Nutr.* 102, 453–461 (2009).

19. Sjödin, A. M. *et al.* The influence of physical activity on BMR. *Med. Sci. Sports Exercise* 28, 85 (1996).

20. [No title]. https://www.ahajournals.org/doi/10.1161/HYPERTENSIONAHA.119.14028.

21. Zanetti, D. *et al.* Urinary Albumin, Sodium, and Potassium and Cardiovascular Outcomes in the UK Biobank: Observational and Mendelian Randomization Analyses. *Hypertension* 75, 714–722 (2020).

22. O'Donnell, M. *et al.* Urinary sodium and potassium excretion, mortality, and cardiovascular events. *N. Engl. J. Med.* 371, 612–623 (2014).

23. Marciniak, S. & Perry, G. H. Harnessing ancient genomes to study the history of human adaptation. *Nature Reviews Genetics* vol. 18 659–674 (2017).

24. Balbuena, L., Bowen, R., Baetz, M. & Marwaha, S. Mood Instability and Irritability as Core Symptoms of Major Depression: An Exploration Using Rasch Analysis. *Front. Psychiatry* 7, (2016).

25. Ward, J. *et al.* Genome-wide analysis in UK Biobank identifies four loci associated with mood instability and genetic correlation with major depressive disorder, anxiety

disorder and schizophrenia. *Transl. Psychiatry* 7, 1–9 (2017).

26. Berg, J. J. *et al.* Reduced signal for polygenic adaptation of height in UK Biobank. doi:10.1101/354951.

27. Refoyo-Martínez, A. *et al.* How robust are cross-population signatures of polygenic adaptation in humans? doi:10.1101/2020.07.13.200030.

4d) Identifying candidates for positive selection using patterns of ancient population differentiation

Alba Refoyo Martínez[1], Fernando Racimo[1]

[1]Lundbeck Foundation GeoGenetics Centre, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark

Introduction

We aimed to detect whether there is evidence of positive selection in the past 15,000 years by searching for loci with strong differentiation in allele frequencies both between populations and across time.

Methods

We worked with the imputed dataset (Supplementary Note S2). We applied several variant-level filters: 1) we only used SNPs with MAF > 5% , 2) genotype missingness rate < 50% and 3) variants where <10% individuals had post-imputation genotype probability (GP) <= 0.8. We used pcadapt [1] - a method for detection of allele frequency outliers based on principal component analysis - to search for loci that might have evidence for positive selection in the past. After performing a PCA on the ancient genomic data, we used a scree plot to visualise the percentage of variance explained by each PC.

We performed two scans to look for candidates under positive selection. First, we performed a “eurasian” scan, in which we used the first three principal components of a PCA of all ancient genomes (higher components explained less than 1% of the total variance in allele frequencies) (Figures **S4d.1, S4d.2 and S4d.3**). We also performed a second scan of selection restricting only to ancient West Eurasian individuals (excluding ancient Siberian populations) and we also used the first three components of the PCA (Figures **S4d.4, S4d.5**

and S4d.6). We call this the “west-eur” scan. We also performed a third scan, which we call “hg-neo” scan, in which we only tested for significant loadings corresponding to the component that separates hunter-gatherers (WHG+EHG) from Neolithic farming peoples in the aforementioned Western Eurasian PCA (first principal component). The latter scan should serve to find loci with particularly strong allele frequencies between these groups with two distinct modes of subsistence. Figure **S4d.2** and **S4d.5** shows Manhattan plots for each of the three scans.

We selected the top-scoring 300 SNPs with the lowest p-values and merged them into candidate regions if they were within 100kb of each other. Each region was labelled with the P-value of its highest scoring SNP, which was then used to rank regions. HGNC protein-coding genes within each region were retrieved using *biomaRt* [2]. Table **S4d.1** lists the 25 top-scoring candidate regions from the “eurasian” scan, while Tables **S4d.2** and list the top candidate regions from the “west-eur” and “hg-neo” scan, respectively.

None of our candidate regions seem to be affected by mapping biases. The Fj value for highest scoring SNP in each region is lower than 0.5 (Supplementary Note S4b).

Results

Eurasian scan

In the Eurasian-wide scan, the strongest peak contains the gene SLC24A5, involved in skin and eye pigmentation, and previously reported to be under positive selection in Western Eurasia [3–6] (Figure **S4d.2**). We also recover the region encompassing the *EDAR* gene which has been implicated in numerous studies of positive selection involving East Asian populations [7–9]. Variants in this gene are associated with numerous ectoderm-related traits, including hair thickness and ectodermal dysplasia [10] and tooth morphology [11].

We also found a candidate peak (chr2:25874547-26568094), containing several genes (ASXL2, RAB10, HADHA, GPR113) involved in glucose homeostasis mainly in response to a high fat diet. The ASXL2 gene regulates skeletal, glucose, and adipocyte homeostasis. It promotes adipogenesis, the formation of osteoclasts and insulin resistance [12–14]. A similar activity is carried out by RAB10, it also regulates glucose homeostasis by improving hyperglycemia, regulating at the skeletal muscle level [15–17]. HADHA participates in long fatty acid oxidation for energy production in different tissues [18–20]. Finally, GPR113 participates in energy expenditure in fat tissue and glucose homeostasis. It improves insulin sensitivity and

prevents obesity when it binds to bile acids [21]. High allele frequencies in East Asian populations (Panel B of Figure **S4d.7**).

We found four peaks containing genes associated with cardiovascular disorders and obesity. One of these (chr20:18991679-19454079) overlaps with the SLC24A3 gene. This is a salt sensitivity gene which is significantly expressed in obese individuals [22–24] and it is associated with hypertension [25]. We also find a peak in an intergenic region (chr3:123433220-123889576) containing ROPN1 and KALRN, two genes involved in vascular disorders [26–28]. The alternative allele at the top SNP in this region is at particularly high frequency in ancient Steppe populations. Another candidate region (chr1:234142067-234549596) contains SLC35F3, which codes for a thiamine transport and has been associated with hypertension in a Han Chinese cohort [29,30]. In the same region, we also find COA6, which has high expression in cardiac pathologies [31]. The alternative allele frequency at the top SNP in the region is high in East Asian populations (Panel B of Figure **S4d.7**). Finally, the region (chr10:90592757 - 91009553) contains several genes (CH25H, FAS) associated with obesity and lipid metabolism [32–34], and immune responses [35].

One of the top candidate regions (chr11:131077365-131516733) contains a gene - NTM - involved in neuropsychiatric disorders [25,36], while another region (chr11:44634764 - 45073989) contains a gene associated with schizophrenia, TSPAN18. It has previously shown that the schizophrenia-risk SNPs within this region are highly diverged between Europeans and East Asians [37]. In chromosome 7 (chr7:95947959-96347959), SLC25A13, which is highly associated with citrin deficiency in East Asian populations [38,39]. High alternative allele frequencies in East Asians (Panel A of Figure **S4d.7**) [40,41].

West Eurasian scan

In the scan restricting to ancient populations in Western Eurasia (“west-eur”), we recover three regions which are involved in skin, hair and eye pigmentation, and have been previously implicated in differences in these traits across present-day Eurasians: two in chromosome 15 containing the genes *SLC24A5/MYEF2/CTXN2 and OCA2/HERC2* respectively and one in chromosome 5, containing gene *SLC45A2* [3,42–45]. Alternative allele frequencies shown in Figure **S4d.8**.

The region containing *LCT/MCM6* - responsible for lactase persistence in Europe - is also a candidate region in the selection scan [46–48]. We also recover the TLR-1-6-10 gene cluster,

which is known to be a target of selection in Europe and is associated with the immune response [46–48] (Panel A, Figure **S4d.8).**

Additionally, we find some new potentially important candidate regions for positive selection. The region showing the strongest evidence of selection is located in chromosome 6 (chr6:134192815-134628278), around the SLC2A12 gene, which codes for a glucose transporter that participates in glucose homeostasis [49,50]. Variants in this gene are associated with mean corpuscular levels, heart diseases and height. The alternative allele of the highest-scoring SNP was at high frequency in hunter-gatherers but at much lower frequencies in Neolithic farmer populations and other, more recent, populations (Panel B, Figure **S4d.8**).

Another novel candidate region overlaps with the VAMP 5-8 gene cluster in chromosome 2:85369379-85885211 a region associated with cardiovascular diseases [51,52]. In chr9:27009422-27434948, we found a peak overlapping the TEK gene, which codes for a tyrosine kinase receptor expressed in endothelial cells. This gene has an important role in angiogenesis and cardiovascular development and stability, and it is involved in several vascular disorders [53–55]. Recent studies have also investigated its role in asthma and allergic conjunctivitis [53–56]. Intermediate allele frequencies in Eastern hunter-gatherers in both regions (Panel B,Figure **S4d.8**). Region chr1:227020437 - 227877723 contains CDC42BPA, also known as MRCKα, an important gene involved in iron utilisation [57] is involved in the erythropoiesis regulation [58]. The alternative allele at the top-scoring SNP in this region is at high frequencies found in hunter-gatherer groups, predominantly in eastern hunter-gatherers (Panel B, Figure **S4d.8**).

In chr15:38464638-38992430, we recovered RASGRP1, associated with immunity and related to systemic lupus erythematosus [59], rheumatoid arthritis [60] or Epstein-Barr virus [61] among other disorders. The alternative allele at this SNP is at high frequencies in eastern hunter gatherers and other ancient Baltic populations (Panel B, Figure **S4d.8**). We also found a wide candidate region containing several high-scoring SNPs in chromosome 16:66852047-67871804. The gene that falls in the highest peak of the region is ATP6V0D1, and it plays a very important role in the replication of influenza virus [62,63]. In this region, there are also several genes - *TPPP3/ZDHHC1* and *HSD11B2* - associated with obesity, cardiovascular diseases and hypertension [64–70]. The alternative allele at the top-scoring SNP has intermediate allele frequencies in hunter-gatherer populations, and higher frequencies in eastern hunter-gatherers.

Neolithic vs. hunter-gatherer scan

The “neo-hg” scan specifically recovers patterns of allele differentiation along the axis separating hunter-gatherer and farmer populations in West Eurasia. In this scan, we find a large number of high-scoring regions associated with lipid and sugar metabolism, and various metabolic disorders.

For example, we recover the *FADS* gene cluster, involved in lipid metabolism. This region is presumed to be important in the transition to a diet rich in grains, as a consequence of the expansion of agriculture in Europe and/or the demographic transitions subsequent to it [47,71–73]. This region is also found in the “west-eurasia” scan. The alternative allele frequency at the top-scoring SNP is very high in hunter-gatherer populations.

Another region is located in chromosome 22:31353354-31759255 and also contains genes involved in lipid metabolism: PATZ1, LIMK2, MORC2 and PLA2G3. PATZ1 down-regulates FADS1 [74], LIMK2 shows particularly elevated expression in metabolic syndrome [75], MORC2 plays an important role in cellular lipid metabolism [76–78] and PLA2G3 contributes to atherogenesis [79–81]. The top-scoring SNP in this region has high alternative allele frequencies in Neolithic farmer populations (Panel A, Figure **S4d.9**).

At chromosome 12:6875213-7366672, we find a region with several genes involved in lipid metabolism: PTPN6, EMG1, PHB2, LPCAT3, C1S**.** LPCAT3 is essential in high fat diets and associated with oleic acid levels and linoleic acid [82] and its deficiency alters cholesterol promoting atherosclerosis [83] and plays an important role in hyperuricemia [83,84]. C1S also plays a crucial role in innate immunity [85] and has been recently associated with coronary syndrome [86].

In chromosome 17:8655348-9223981, we found the PIK3 family (PIK3R5, PIK3R6) which is involved in glucose homeostasis and plays an important role in obesity and insulin resistance in type 2 diabetes [87–89].

In chromosome 11:27432440-27832440 we find a region containing BDNF (brain-derived neurotrophic factor), expression of which is associated with obesity [90,91]. It participates in the reduction of free fatty acids, cholesterol and glucose levels and enhances energy expenditure [92]. Its expression has been shown to be suppressed with a high-fat sucrose diet [90,9394]; [90,93]. High alternative allele frequency in HG (Panel A, Figure **S4d.9**).

One of the strongest peaks in this scan corresponds to SLC2A12, coding for a glucose transporter. The next peak is located in chromosome 17:79055998-79469799. SLC38A10, which falls in the tip of the peak, is involved in absorption of amino acids from the GI tract [95] and has been suggested to play a role in pathways involved in neurotransmission [96,97]. BAIAP2's [98,99] and AATK's expression [100,101], which are also found in the same region, are related to high-fat and omega3 fatty acids. Higher alternative allele frequencies are found in hunter-gatherers, in particular in Eastern-HG (Panel A, Figure **S4d.9**).

The highest peak in chromosome 4 contains several genes involved in alcohol metabolism, ADH1B, ADH1C and ADH7 [102,103]. In Panel B, Figure **S4d.9**, we can see the highest alternative allele frequencies for the oldest samples, which correspond to hunter-gatherers.

Two of the top peaks are related to innate immune response in humans. In chromosome 14:73102086-73502086, ZFYVE1 is involved in TLR3-mediated immune response and regulates antiviral response [104,105]. In chromosome 3:98028061-98476960, GPR15 is related to immune tolerance and it regulates the homeostasis in the intestine mucosa [104,106,107]. In chromosome 18:46132358-46558884, SMAD7 is associated with inflammatory bowel diseases such as crohn's disease [108–110].

Two regions are related to brain disorders. In chromosome 1:110588519-111127548, several genes regulate neuronal ion channels: KCNC4, SLC6A17, and STRIP1. Mutations in SLC6A17 cause intellectual disabilities associated with speech impairment and behavioural problems [111], while the KCNC gene family has also been associated with intellectual disability [112].

Figures

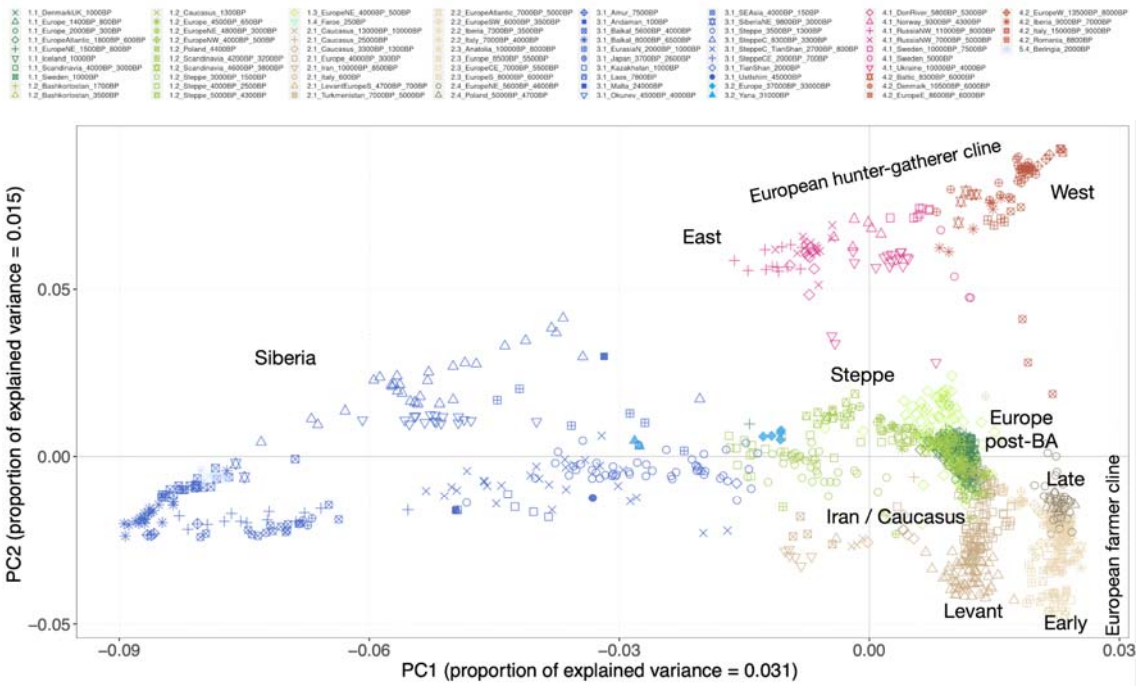


Figure S4d.1. Principal component analysis on 1402 Eurasian samples. The first component explains 3.1% of the variance and separates East Asian, Steppe and European samples. The second component separates farmers and Hunter-gatherers (1.5%).

Eurasian scan

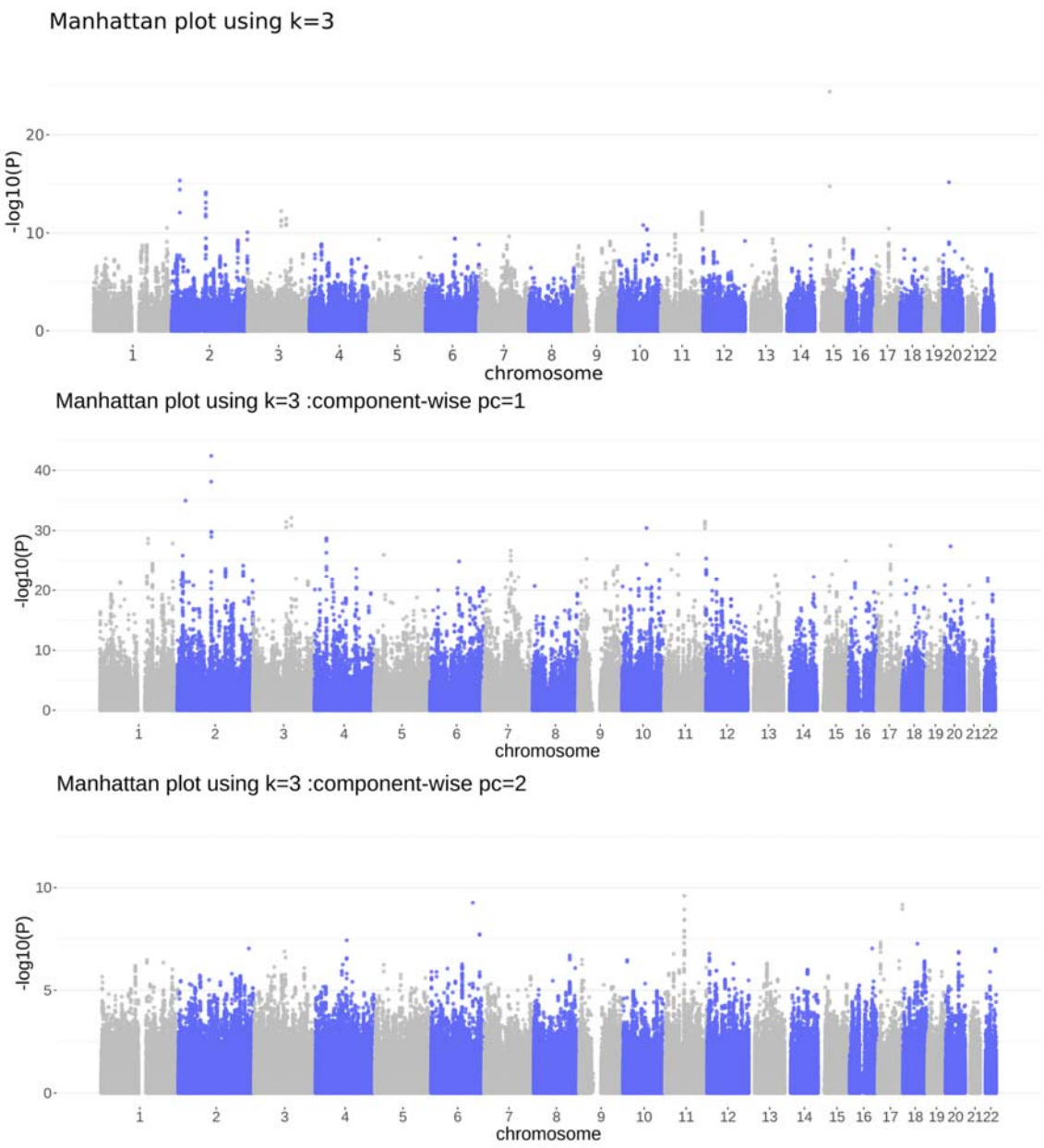


Figure **S4d.2**. Genome scan using pcadapt k=3. A. Method: mahalanobis distances. Manhattan plot scanning the whole genome. B and C. Component-wise genome scans for component PC1 and component PC2, respectively.

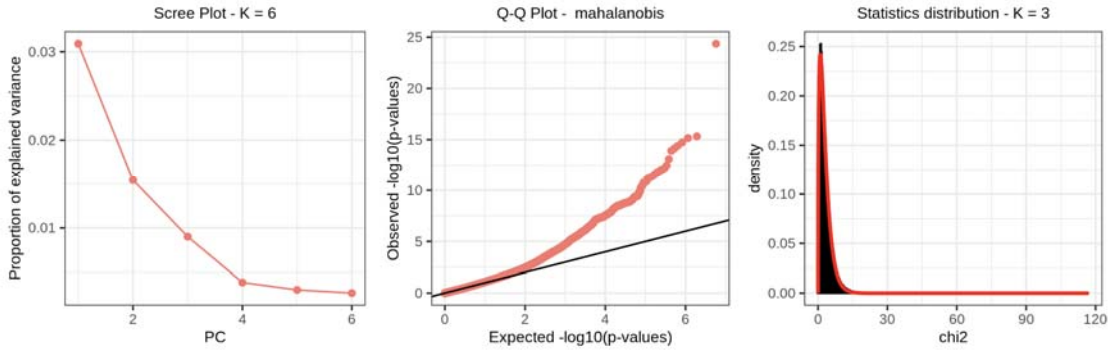


Figure **S4d.3**. A. Scree plot showing proportion of explained variance of the first PCs in the pcadapt analysis. B. Q-Q plot using mahalanobis method for K=3. Distribution of pcadapt scores (k=3) compared to chi-squared distribution with one degree of freedom (red line).

West-Eurasian

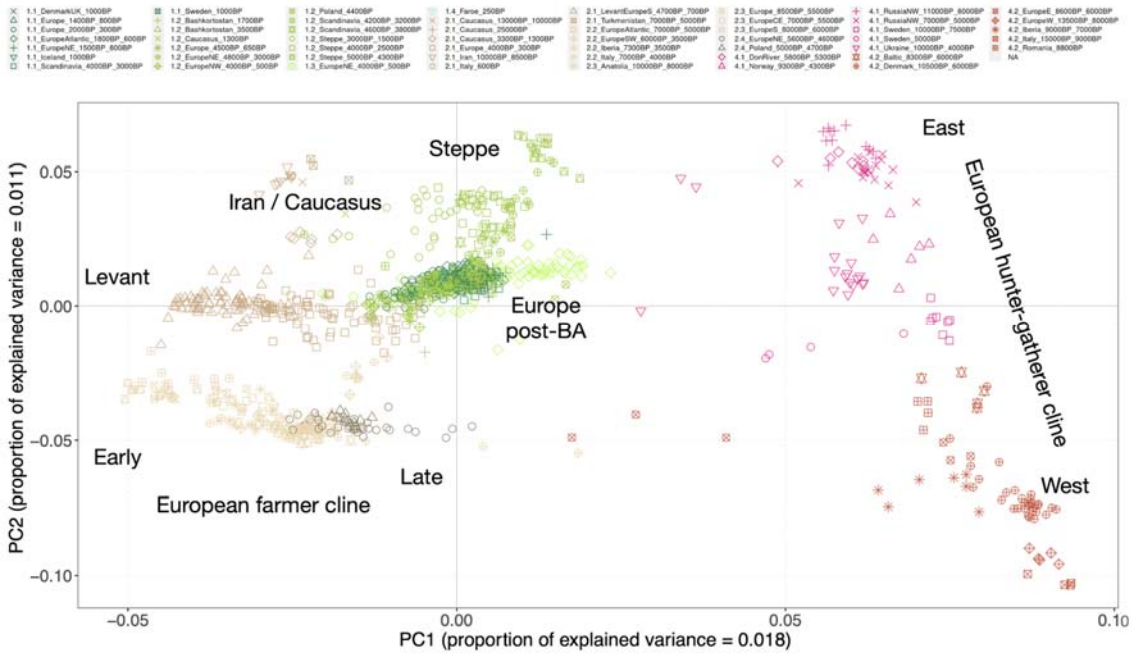


Figure S4d.4. Principal component analysis on 1165 Eurasian samples. The first component explains 1.8% of the variance and separates East Asian, Steppe and European samples. The second component separates farmers and Hunter-gatherers (1.1%).

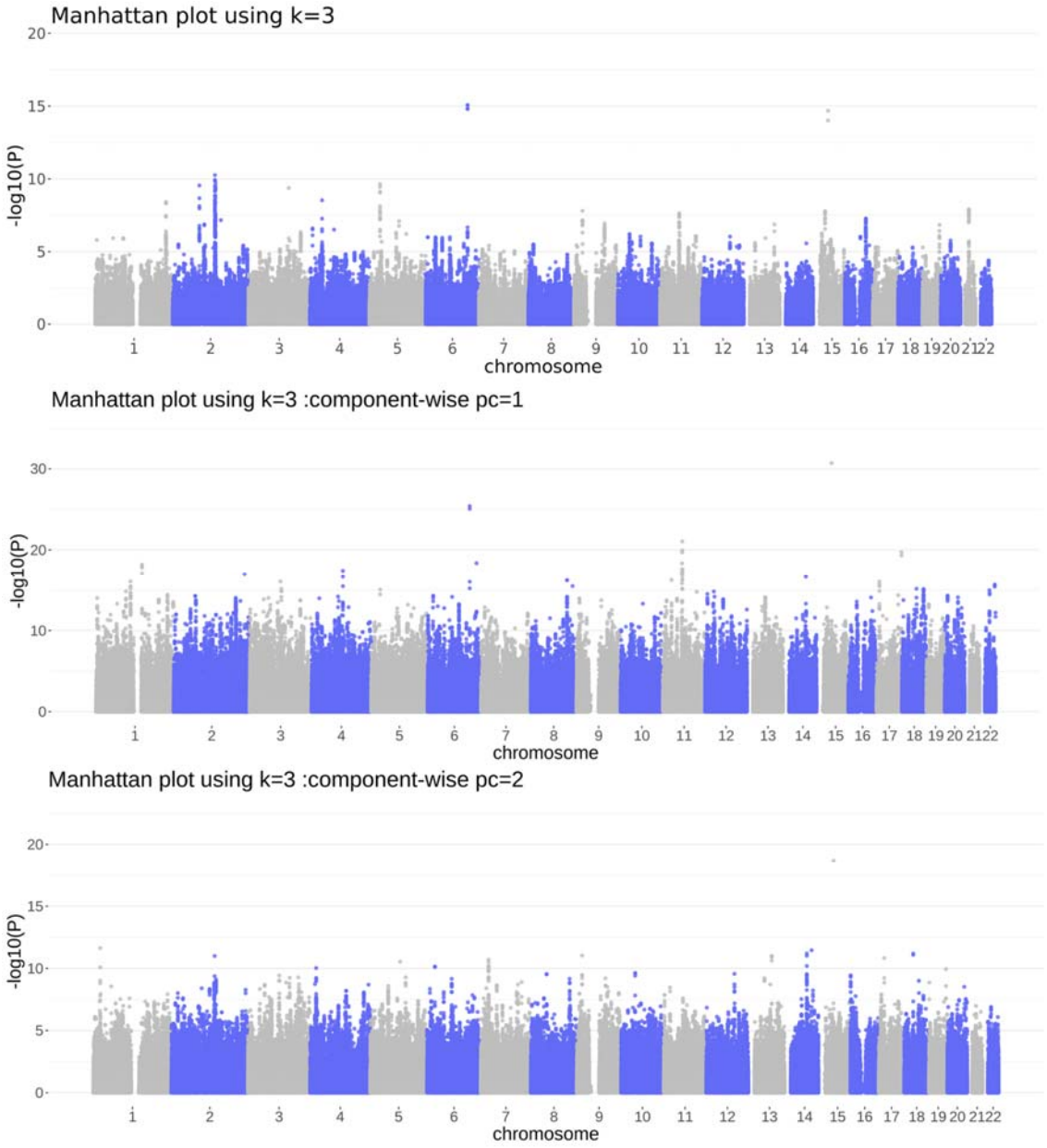


Figure **S4d.5**. Genome scan using pcadapt k=3. A. Method: mahalanobis distances. Manhattan plot scanning the whole genome. B and C. Component-wise genome scans for component PC1 and component PC2, respectively.

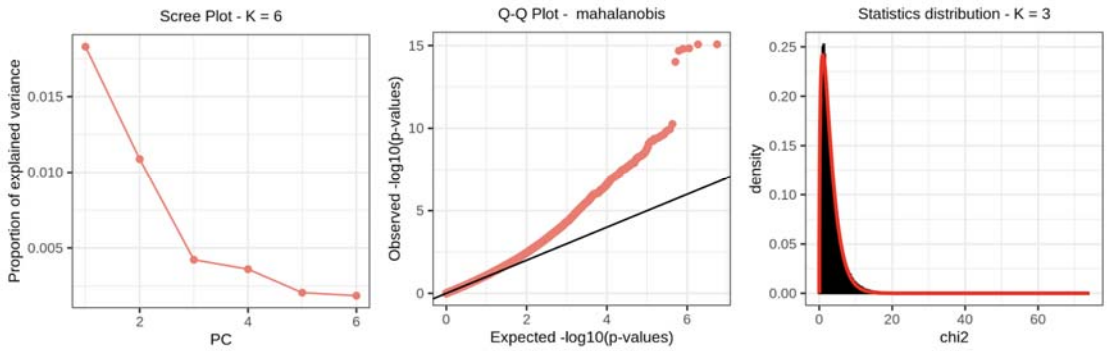


Figure **S4d.6**. A. Scree plot showing proportion of explained variance of the first PCs in the pcadapt analysis. B. Q-Q plot using mahalanobis method for K=3. Distribution of pcadapt scores (k=3) compared to chi-squared distribution with one degree of freedom (red line).

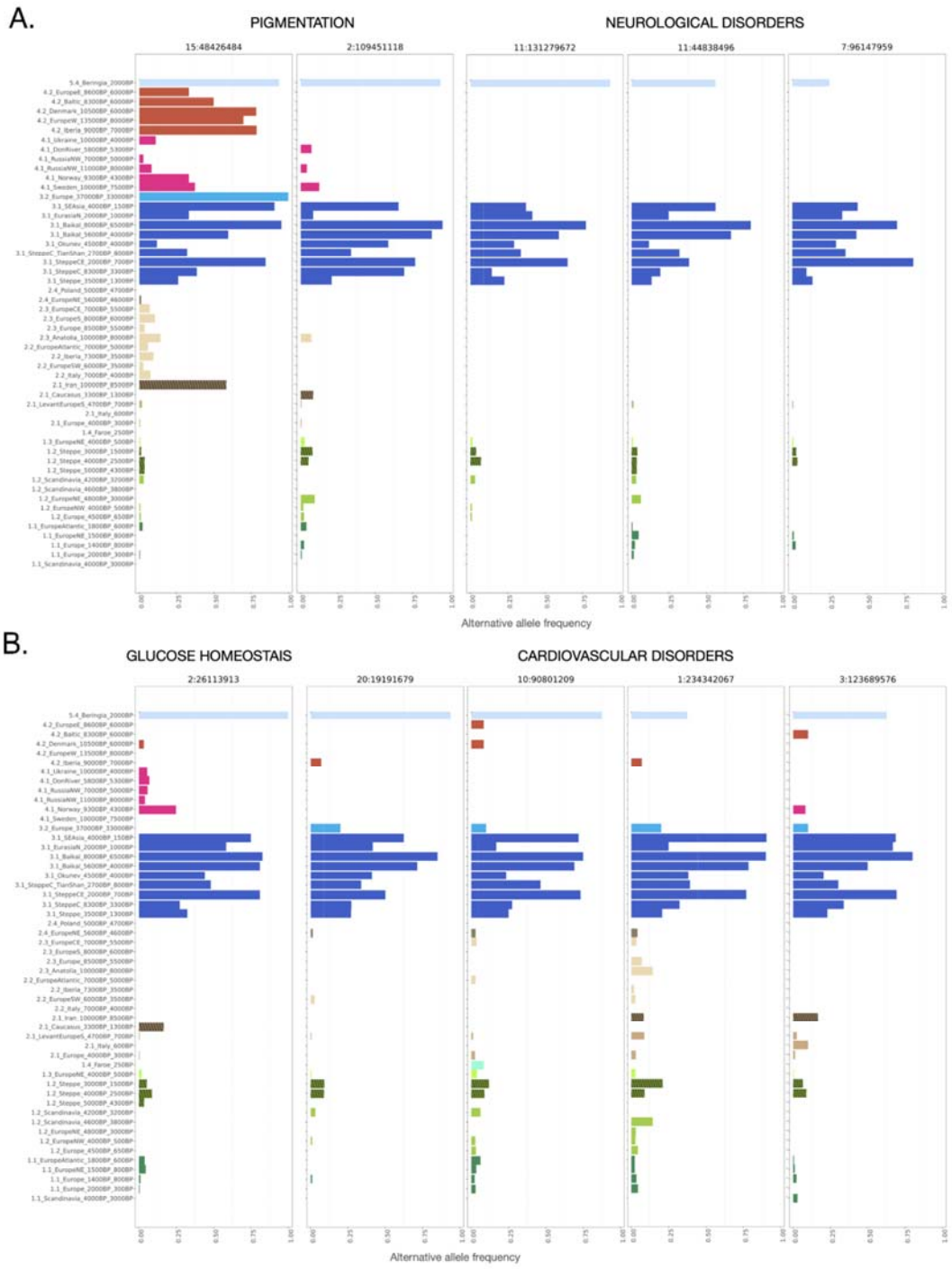


Figure **S4d.7**. Alternative allele frequencies of the position with the lowest p-value in the top regions for the eurasian scan.

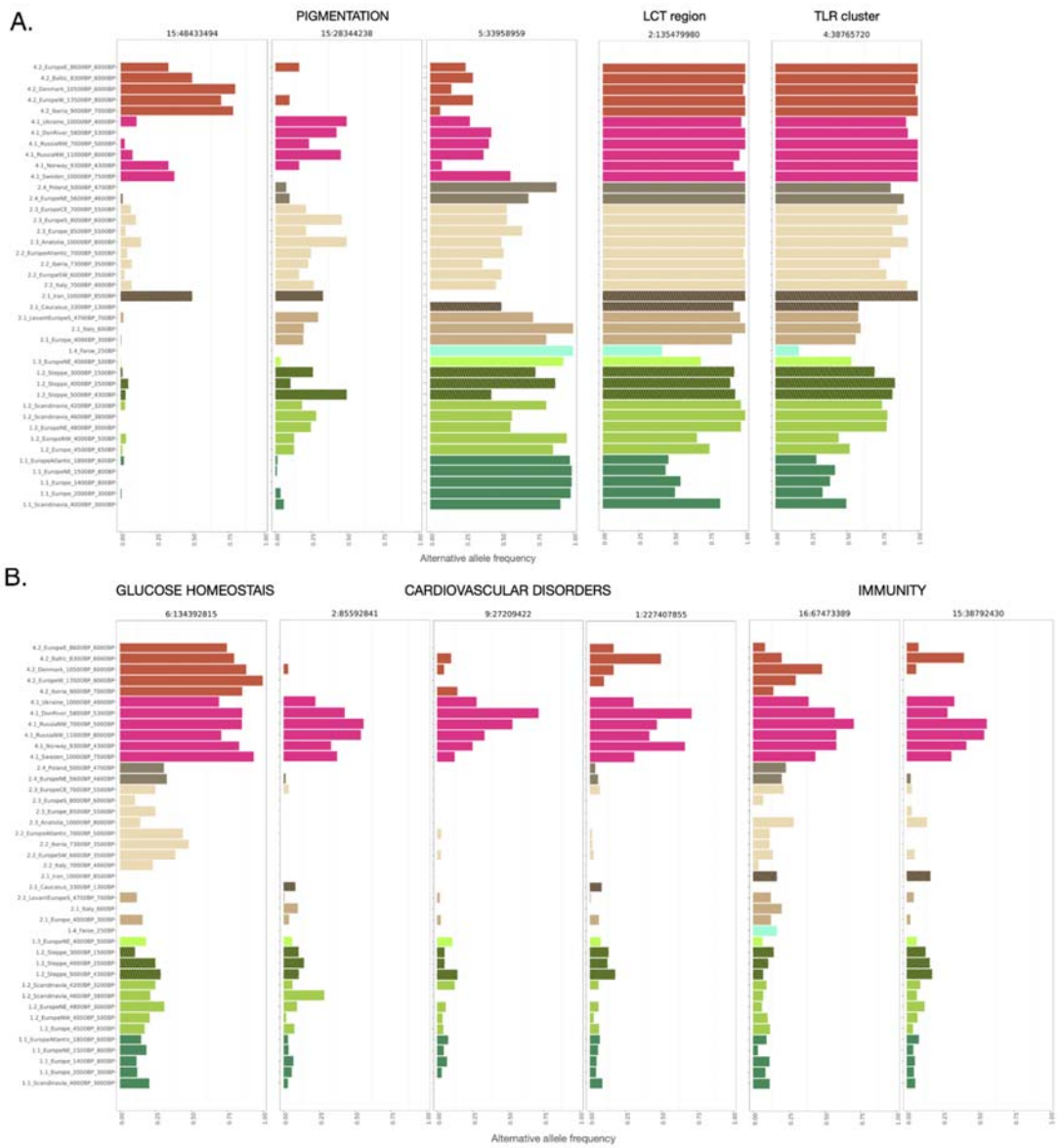


Figure **S4d.8**. Alternative allele frequencies of the position with the lowest p-value in the top regions for the West-eurasian scan.

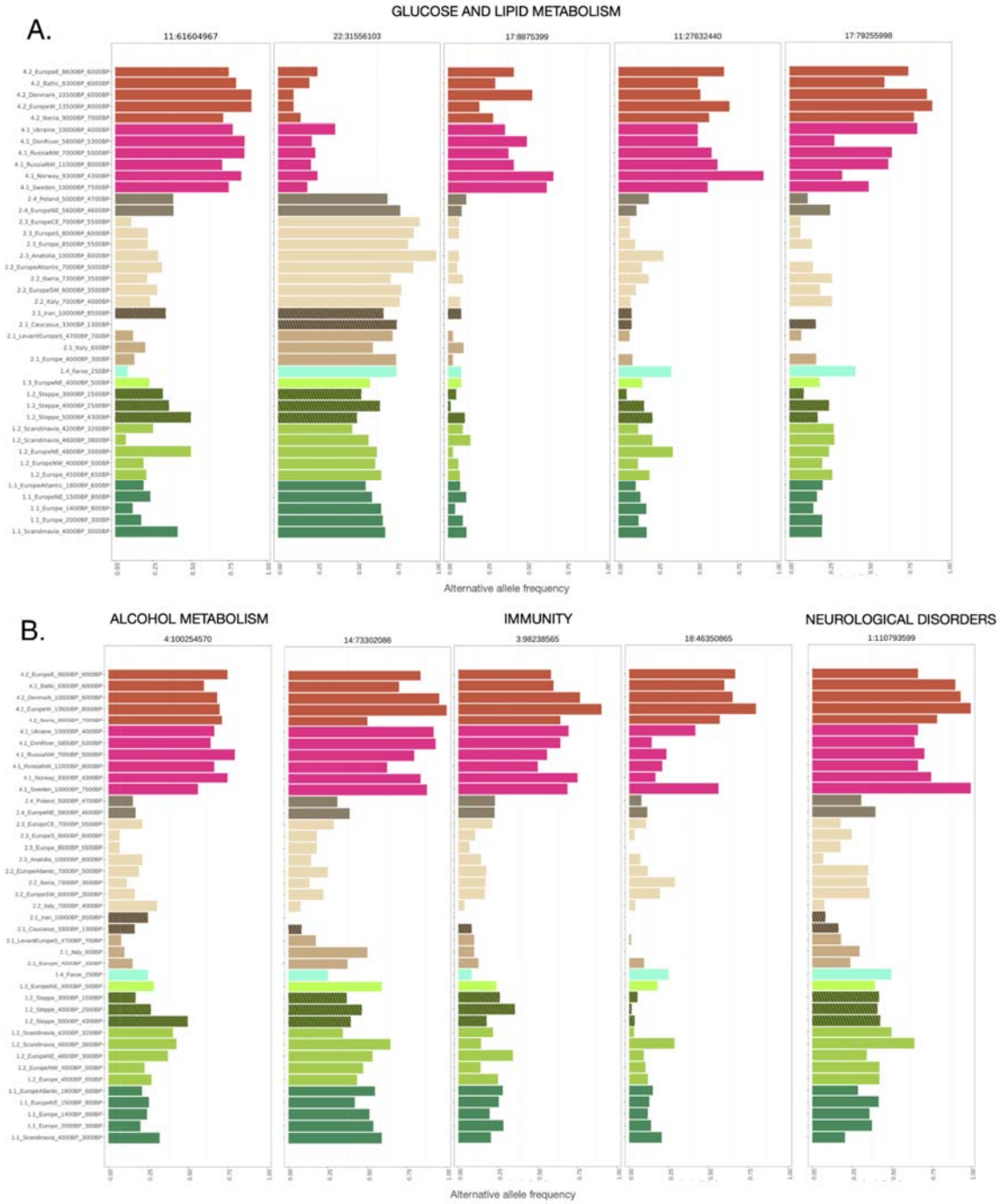


Figure **S4d.9**. Alternative allele frequencies of the position with the lowest p-value in the top regions for the HG West-eurasian scan.

chr	start	end	bestpos	minpvalue	ensembl	hgnc	disease_trait_best	disease_trait
15	48211821	48714309	48426484	4.40E-25	ENSG00000188467, ENSG00000104177, ENSG00000233932, ENSG00000074803, ENSG00000128951, ENSG00000166147	**SLC24A5**, MYEF2, CTXN2, SLC12A1, DUT, FBN1	ENSG00000188467: Hair color, Skin pigmentation, Body mass index, Skin, hair and eye pigmentation (multivariate analysis), Eye color, Eye color (brightness), Eye color (saturation), Iris color (a* coordinate), Skin reflectance (Melanin index), Iris color (b* coordinate)	ENSG00000188467: Hair color, Skin pigmentation, Body mass index, Skin, hair and eye pigmentation (multivariate analysis), Eye color, Eye color (brightness), Eye color (saturation), Iris color (a* coordinate), Skin reflectance (Melanin index), Iris color (b* coordinate); ENSG00000104177: Skin pigmentation, Hair color; ENSG00000233932: ; ENSG00000074803: Systemic lupus erythematosus, Longevity, Skin reflectance (Melanin index), Lymphocyte counts; ENSG00000128951: Protein quantitative trait loci; ENSG00000166147: Breast cancer, Spherical equivalent (joint analysis main effects and education interaction), Height, Pulse pressure, Refractive error, Spherical equivalent, Systolic blood pressure, Thoracic aortic aneurysms and dissections, Colorectal cancer, Spherical equivalent or myopia (age of diagnosis), Central corneal thickness, Intracranial, abdominal aortic or thoracic aortic aneurysm (pleiotropy), Systolic blood pressure x alcohol consumption interaction (2df test), Pulse pressure x alcohol consumption interaction (2df test), Skin reflectance (Melanin index), Spontaneous coronary artery dissection, Waist circumference adjusted for body mass index, Intraocular pressure, Macular thickness, Heel bone mineral density, Lung function (FEV1/FVC)
2	25874547	26568094	26113913	4.68E-16	ENSG00000138101, ENSG000001	DTNB, ASXL2,	ENSG00000084731: Atrial fibrillation, Type 2 diabetes, Height, Red blood cell count	ENSG00000138101: Multiple myeloma, LDL cholesterol, Multiple myeloma (IgH translocation), Type 2 diabetes, Cutaneous

			(rs784 04020)		43970, ENSG000000 84731, ENSG000000 84733, ENSG000001 57833, ENSG000000 84754, ENSG000001 38029, ENSG000001 73567, ENSG000001 38018	**KIF3C**, RAB10 , GARE ML, HADH A, HADH B, GPR1 13, EPT1		malignant melanoma, Nevus count or cutaneous melanoma, B-cell malignancies (chronic lymphocytic leukemia, Hodgkin lymphoma or multiple myeloma) (pleiotropy), C-reactive protein levels, Response to platinum-based chemotherapy (cisplatin), Waist circumference adjusted for body mass index, Body mass index, Height; ENSG00000143970: CTACK levels, Hemoglobin levels, Red cell distribution width, Hemoglobin concentration, Thyroid stimulating hormone levels, Alcohol consumption (drinks per week) (MTAG), Hair colour, Mean corpuscular hemoglobin; ENSG00000084731: Atrial fibrillation, Type 2 diabetes, Height, Red blood cell count; ENSG00000084733: Immune response to smallpox vaccine (IL-6), Mean corpuscular volume, Apolipoprotein B levels, Mean corpuscular hemoglobin; ENSG00000157833: Red cell distribution width; ENSG00000084754: ; ENSG00000138029: ; ENSG00000173567: ; ENSG00000138018: Coronary artery disease, Apolipoprotein A1 levels
20	189916799	19454079	19191679 (rs4141981)	6.92E-16	ENSG000001 85052	SLC24 A3	ENSG00000185052: Matrix metalloproteinase levels, Pulmonary function decline, Age at smoking initiation in chronic obstructive pulmonary disease, Metabolite levels, QT interval (sulfonylurea treatment interaction), Migraine, Pulse pressure, Smoking status (ever vs never smokers), Cataracts (operation), Mean platelet volume, Diastolic blood pressure, Lifetime	ENSG00000185052: Matrix metalloproteinase levels, Pulmonary function decline, Age at smoking initiation in chronic obstructive pulmonary disease, Metabolite levels, QT interval (sulfonylurea treatment interaction), Migraine, Pulse pressure, Smoking status (ever vs never smokers), Cataracts (operation), Mean platelet volume, Diastolic blood pressure, Lifetime smoking index, Psychosis (atypical), Daytime sleepiness, Nicotine dependence and major depression (severity of comorbidity), Nicotine dependence symptom count, Platelet

							smoking index, Psychosis (atypical), Daytime sleepiness, Nicotine dependence and major depression (severity of comorbidity), Nicotine dependence symptom count, Platelet count, Multisite chronic pain, Breast cancer and/or colorectal cancer, Educational attainment (years of education), Educational attainment (MTAG), Highest math class taken (MTAG), Height, Menarche (age at onset), Lung function (FEV1/FVC), Smoking status	count, Multisite chronic pain, Breast cancer and/or colorectal cancer, Educational attainment (years of education), Educational attainment (MTAG), Highest math class taken (MTAG), Height, Menarche (age at onset), Lung function (FEV1/FVC), Smoking status
2	108527043	109781319	109451118 (rs72627476)	7.39E-15	ENSG00000115665, ENSG00000196228, ENSG00000198203, ENSG00000198075, ENSG00000135968, ENSG00000169756, ENSG00000153201, ENSG00000163006, ENSG00000135960, ENSG00000172985	SLC5A7, SULT1C3, SULT1C2, SULT1C4, GCC2, LIMS1, RANBP2, CCDC138, **EDAR**, SH3RF3	ENSG00000163006: Lobe attachment (rater scored), Birth weight	ENSG00000115665: ; ENSG00000196228: Systolic blood pressure, Diastolic blood pressure; ENSG00000198203: Lobe size; ENSG00000198075: Ear protrusion, Lobe attachment, Helix rolling, Lobe attachment (rater-scored or self-reported), Lobe attachment (rater scored); ENSG00000135968: Low density lipoprotein cholesterol levels; ENSG00000169756: LDL cholesterol levels, Apolipoprotein B levels, Low density lipoprotein cholesterol levels, Excessive hairiness, Monocyte chemoattractant protein-1 levels, Total cholesterol levels, LDL cholesterol, Eyebrow thickness, Birth weight, Corneal endothelial cell density; ENSG00000153201: Attention deficit hyperactivity disorder; ENSG00000163006: Lobe attachment (rater scored), Birth weight; ENSG00000135960: Monobrow thickness, Scalp hair shape, Beard thickness, Eyebrow thickness, Ear

								protrusion, Tragus size, Ear morphology, Lobe size, Lobe attachment, Helix rolling, lower facial morphology traits (quantitative measurement), Lobe attachment (rater scored), Straight vs curly hair, Blood protein levels, Gamma glutamyl transferase levels in excessive alcohol consumption, Male-pattern baldness, Monobrow, Thick vs thin eyebrows, Lung function (FEV1), Lung function (FVC), Balding type 1, Hair color; ENSG00000172985: Neurocognitive impairment in HIV-1 infection (dichotomous), Height, Cerebrospinal fluid sTREM-2 levels, Heel bone mineral density, Pediatric bone mineral density (spine), Bitter taste perception (6-n-propylthiouracil) in obesity with metabolic syndrome, Interleukin-5 levels, General cognitive ability, Rapid automised naming of digits, Corneal endothelial cell density, Highest math class taken (MTAG)
3	106950936	107415936	107150936	6.02E-13	ENSG00000138483, ENSG00000114439	CCDC54, BBX	ENSG00000138483:	ENSG00000138483: ; ENSG00000114439: Smoking behaviour, Metabolite levels, Neurociticism, Irritable mood, Experiencing mood swings, Feeling worry, Male-pattern baldness, Depressed affect, Neuroticism, Waist-to-hip ratio adjusted for BMI (adjusted for smoking behaviour), Waist-to-hip ratio adjusted for BMI, Measles, Small cell lung carcinoma, Body mass index, Mood instability, Waist-hip ratio, White blood cell count, Neutrophil count, Platelet count, General cognitive ability, Multisite chronic pain, Monocyte count, Mean platelet volume, Cognitive ability, years of educational attainment or schizophrenia (pleiotropy), Educational attainment (MTAG), Cognitive performance (MTAG), Height, Balding type

								1, Educational attainment (years of education), Highest math class taken (MTAG)
11	131077365	131516733	131279672	7.74E-13	ENSG00000182667	NTM	ENSG00000182667: Bipolar disorder and schizophrenia, Sunburns, Obesity-related traits, Male fertility, Asperger disorder, Educational attainment (years of education), Aggressiveness in attention deficit hyperactivity disorder, Itch intensity from mosquito bite adjusted by bite size, Food addiction, Feeling worry, Refractive error, Smoking status (ever vs never smokers), Leisure sedentary behaviour (television watching), Leisure sedentary behaviour (computer use), Spherical equivalent, Myopia, Cardiac Troponin-T levels, Chronic obstructive pulmonary disease-related biomarkers, Metabolite levels, Serum polyunsaturated fatty acid concentration x sex interaction in metabolic syndrome, Myopia (age of diagnosis), Spherical equivalent or myopia (age of diagnosis), Plasma anti-thyroglobulin levels, C-reactive protein levels, Schizophrenia, Peripheral arterial disease (traffic-related air pollution interaction), Body mass index, Intake of total sugars,	ENSG00000182667: Bipolar disorder and schizophrenia, Sunburns, Obesity-related traits, Male fertility, Asperger disorder, Educational attainment (years of education), Aggressiveness in attention deficit hyperactivity disorder, Itch intensity from mosquito bite adjusted by bite size, Food addiction, Feeling worry, Refractive error, Smoking status (ever vs never smokers), Leisure sedentary behaviour (television watching), Leisure sedentary behaviour (computer use), Spherical equivalent, Myopia, Cardiac Troponin-T levels, Chronic obstructive pulmonary disease-related biomarkers, Metabolite levels, Serum polyunsaturated fatty acid concentration x sex interaction in metabolic syndrome, Myopia (age of diagnosis), Spherical equivalent or myopia (age of diagnosis), Plasma anti-thyroglobulin levels, C-reactive protein levels, Schizophrenia, Peripheral arterial disease (traffic-related air pollution interaction), Body mass index, Intake of total sugars, Familial lung adenocarcinoma, Mental health study participation (completed survey), Reaction time, Smoking initiation (ever regular vs never regular), Age of smoking initiation (MTAG), Cognitive ability, years of educational attainment or schizophrenia (pleiotropy), Smoking initiation (ever regular vs never regular) (MTAG), Pre-treatment viral load in HIV-1 infection, Educational attainment (MTAG), Highest math class taken (MTAG), Smoking status, White blood cell count, Lung function (FEV1/FVC)

							Familial lung adenocarcinoma, Mental health study participation (completed survey), Reaction time, Smoking initiation (ever regular vs never regular), Age of smoking initiation (MTAG), Cognitive ability, years of educational attainment or schizophrenia (pleiotropy), Smoking initiation (ever regular vs never regular) (MTAG), Pre-treatment viral load in HIV-1 infection, Educational attainment (MTAG), Highest math class taken (MTAG), Smoking status, White blood cell count, Lung function (FEV1/FVC)	
3	123433220	123889576	123689576 (rs55935332)	3.47E-12	ENSG00000065534, ENSG00000175455, ENSG00000065371, ENSG00000160145	MYLK, CCDC14, **ROPN1**, KALRN	ENSG00000065371:	ENSG00000065534: Adolescent idiopathic scoliosis, General cognitive ability, Cognitive empathy, Educational attainment (MTAG), Educational attainment (years of education); ENSG00000175455: Intelligence (MTAG), Drug-induced liver injury (fluoroquinolones), General cognitive ability; ENSG00000065371: ; ENSG00000160145: Mean platelet volume, Platelet count, Post bronchodilator FEV1/FVC ratio, Plateletcrit, Schizophrenia, Systolic blood pressure x dichotomous lifestyle risk score interaction (1df test), Systolic blood pressure x dichotomous lifestyle risk score interaction (2df test), Intraocular pressure, Moderate-to-late spontaneous preterm birth, Heschl's gyrus morphology, PR interval in Tripanosoma cruzi seropositivity, Thyroid peroxidase antibody levels, Aspartate

								aminotransferase levels in low alcohol consumption, Hematocrit, Platelet distribution width, Amyotrophic lateral sclerosis, Acute graft versus host disease in bone marrow transplantation (donor effect), Childhood steroid-sensitive nephrotic syndrome, Diffuse large B-cell lymphoma or systemic lupus erythematosus, Marginal zone lymphoma or systemic lupus erythematosus, Longevity (age >99th survival percentile), Red blood cell count, Benign childhood epilepsy with centro-temporal spikes, Lymphocyte counts, Childhood ALL/LBL (acute lymphoblastic leukemia/lymphoblastic lymphoma) treatment-related venous thromboembolism, Mean corpuscular hemoglobin concentration, General cognitive ability, Monocyte count, Urine pH measurement, Low urine pH, Educational attainment (MTAG), Cognitive performance (MTAG), Educational attainment (years of education), Red cell distribution width
10	78689487	79110042	78889487	1.62E-11	ENSG00000156113	KCNMA1	ENSG00000156113: Glucocorticoid-induced osteonecrosis, Lean body mass, Angioedema in response to angiotensin-converting enzyme inhibitor and/or angiotensin receptor blocker, Refractive error, Spherical equivalent, Myopia, Hypospadias, Obesity, Myopia (age of diagnosis), Non-melanoma skin cancer, Spherical equivalent or myopia (age of diagnosis), Male-pattern baldness, Initial	ENSG00000156113: Glucocorticoid-induced osteonecrosis, Lean body mass, Angioedema in response to angiotensin-converting enzyme inhibitor and/or angiotensin receptor blocker, Refractive error, Spherical equivalent, Myopia, Hypospadias, Obesity, Myopia (age of diagnosis), Non-melanoma skin cancer, Spherical equivalent or myopia (age of diagnosis), Male-pattern baldness, Initial pursuit acceleration, Educational attainment, Blood pressure, Mumps, Estimated glomerular filtration rate, Body mass index, Response to ranibizumab in age-related macular degeneration

							pursuit acceleration, Educational attainment, Blood pressure, Mumps, Estimated glomerular filtration rate, Body mass index, Response to ranibizumab in age-related macular degeneration (exudative), Smoking cessation in chronic obstructive pulmonary disease, DNA methylation variation (age effect), Heart rate in heart failure with reduced ejection fraction, Balding type 1, Height	(exudative), Smoking cessation in chronic obstructive pulmonary disease, DNA methylation variation (age effect), Heart rate in heart failure with reduced ejection fraction, Balding type 1, Height
1	234142067	234549596	234342067	3.07E-11	ENSG00000183780, ENSG00000168275, ENSG00000059588	**SLC35F3**, COA6, TARBP1	ENSG00000183780: Post bronchodilator FEV1/FVC ratio, Epstein-Barr virus copy number in lymphoblastoid cell lines, Trunk fat mass, Creatinine levels, Metabolite levels, Intracranial aneurysm, Pediatric bone mineral content (spine), Interleukin-6 levels, Chronic obstructive pulmonary disease or high blood pressure (pleiotropy), Adolescent idiopathic scoliosis, Diverticular disease	ENSG00000183780: Post bronchodilator FEV1/FVC ratio, Epstein-Barr virus copy number in lymphoblastoid cell lines, Trunk fat mass, Creatinine levels, Metabolite levels, Intracranial aneurysm, Pediatric bone mineral content (spine), Interleukin-6 levels, Chronic obstructive pulmonary disease or high blood pressure (pleiotropy), Adolescent idiopathic scoliosis, Diverticular disease; ENSG00000168275: ; ENSG00000059588: Cognitive test performance
17	44595190	45055683	44802774	3.52E-11	ENSG00000238083, ENSG00000185829, ENSG00000073969, ENSG00000108379, ENSG00000158955,	LRRC37A2, ARL17A, **NSF**, WNT3, WNT9B, GOSR	ENSG00000073969: Ovarian cancer in BRCA1 mutation carriers, Parkinson's disease, Sense of smell, Intelligence (MTAG), Neuroticism, Neurociticism, Feeling miserable, Experiencing mood swings, Feeling hurt, Feeling fed-up, Feeling nervous, Feeling worry,	ENSG00000238083: ; ENSG00000185829: Hemoglobin levels, Reaction time, Red cell distribution width, Handedness (Left-handed vs. non-left-handed), Handedness (Right-handed vs. non-right-handed), Mean corpuscular hemoglobin concentration, General cognitive ability, Monocyte count; ENSG00000073969: Ovarian cancer in BRCA1 mutation carriers, Parkinson's disease, Sense of smell, Intelligence

					ENSG00000108433, ENSG00000179673	2, RPRML	Cortical surface area (global PC1), Parkinson's disease or first degree relation to individual with Parkinson's disease, Brain region volumes, Hemoglobin levels, Intelligence, Male-pattern baldness, Worry, White matter microstructure (axial diusivities), Reaction time, White matter microstructure (mean diusivities), Epithelial ovarian cancer, Thyroid stimulating hormone levels, Cortical surface area, General factor of neuroticism, Smoking initiation (ever regular vs never regular), White matter microstructure (radial diusivities), White matter microstructure (fractional anisotropy), General cognitive ability, Macular thickness, Balding type 1	(MTAG), Neuroticism, Neurociticism, Feeling miserable, Experiencing mood swings, Feeling hurt, Feeling fed-up, Feeling nervous, Feeling worry, Cortical surface area (global PC1), Parkinson's disease or first degree relation to individual with Parkinson's disease, Brain region volumes, Hemoglobin levels, Intelligence, Male-pattern baldness, Worry, White matter microstructure (axial diusivities), Reaction time, White matter microstructure (mean diusivities), Epithelial ovarian cancer, Thyroid stimulating hormone levels, Cortical surface area, General factor of neuroticism, Smoking initiation (ever regular vs never regular), White matter microstructure (radial diusivities), White matter microstructure (fractional anisotropy), General cognitive ability, Macular thickness, Balding type 1; ENSG00000108379: Parkinson's disease, Celiac disease, Post bronchodilator FEV1, Hematocrit, Intelligence (MTAG), Coronary artery disease, Irritable mood, Neuroticism, Feeling guilty, Experiencing mood swings, Feeling hurt, Cortical surface area (global PC1), Parkinson's disease or first degree relation to individual with Parkinson's disease, Red blood cell count, Multiple system atrophy, Cognitive function, Alzheimer's disease in APOE e4-carriers, Hemoglobin levels, Depressed affect, Male-pattern baldness, White matter microstructure (axial diusivities), Hemoglobin concentration, Breast cancer, Itch intensity from mosquito bite adjusted by bite size, Handedness (non-right-handed vs right-handed), Handedness (Left-handed vs. non-left-handed), Handedness (left-handed vs. right-handed),

								Alcohol consumption (drinks per week), Reaction time, General factor of neuroticism, Atrial fibrillation, Intracranial volume, Lung function (FEV1), White matter microstructure (radial diusivities), Lung function (FVC), Waist-to-hip ratio adjusted for BMI, General cognitive ability, Snoring, White matter microstructure (fractional anisotropy), Smoking initiation (ever regular vs never regular) (MTAG), Cognitive performance (MTAG); ENSG00000158955: Antineutrophil cytoplasmic antibody-associated vasculitis, Intraocular pressure, Mean corpuscular hemoglobin; ENSG00000108433: Blood pressure, Systolic blood pressure, Nonsyndromic cleft lip with cleft palate, Mean arterial pressure, Coronary artery disease, Blood urea nitrogen levels, QRS duration, QRS complex (Cornell), Pulse pressure, Hemoglobin levels, Myocardial infarction, Aortic root size, Orofacial clefts, Cleft lip with or without cleft palate, Idiopathic pulmonary fibrosis, Atrial fibrillation, Myocardial fractal dimension (slice 2), Myocardial fractal dimension (slice 3), Myocardial fractal dimension (slice 4), Medication use (agents acting on the renin-angiotensin system), Medication use (calcium channel blockers), Height, Cardiovascular disease; ENSG00000179673:
10	90592757	91009553	90801209	4.00E-11	ENSG00000152766, ENSG00000138134, ENSG00000107796, ENSG00000026103,	ANKRD22, STAMBPL1, ACTA2, **FAS**, CH25	ENSG00000026103: Immunoglobulin A, Chronic lymphocytic leukemia, Mosquito bite size, Blood protein levels, Blood protein levels in cardiovascular risk, Juvenile idiopathic arthritis (oligoarticular or rheumatoid	ENSG00000152766: Mucinous adenocarcinoma in colorectal cancer, Late-onset Alzheimer's disease, Accelerometer-based physical activity measurement (average acceleration); ENSG00000138134: Lung cancer, Pulse pressure, Brain region volumes; ENSG00000107796: Lung cancer, Pulse pressure, Chronic inflammatory

					ENSG00000138135, ENSG00000107798	H, LIPA	factor-negative polyarticular), Ankylosing spondylitis, Mean corpuscular volume, Red blood cell count	diseases (ankylosing spondylitis, Crohn's disease, psoriasis, primary sclerosing cholangitis, ulcerative colitis) (pleiotropy), Chronic lymphocytic leukemia; ENSG00000026103: Immunoglobulin A, Chronic lymphocytic leukemia, Mosquito bite size, Blood protein levels, Blood protein levels in cardiovascular risk, Juvenile idiopathic arthritis (oligoarticular or rheumatoid factor-negative polyarticular), Ankylosing spondylitis, Mean corpuscular volume, Red blood cell count; ENSG00000138135: ; ENSG00000107798: Coronary heart disease, Fibrinogen levels, Coronary artery disease (myocardial infarction, percutaneous transluminal coronary angioplasty, coronary artery bypass grafting, angina or chromic ischemic heart disease), Coronary artery disease, Neutrophil count, Blood protein levels, C-reactive protein levels, Sum neutrophil eosinophil counts, Myocardial infarction, Itch intensity from mosquito bite adjusted by bite size, White blood cell count, Red cell distribution width
2	241709643	24234344 1	24208771 2	8.54E-11	ENSG00000130294, ENSG00000172482, ENSG00000172478, ENSG00000162804, ENSG00000122085, ENSG00000115687, ENSG00000115685,	KIF1A, AGXT, C2orf54, SNED1, MTERFD2, PASK, **PPP1R7**, ANO7, HDLBP,	ENSG00000115685:	ENSG00000130294: Nicotine withdrawal symptom count, Response to placebo treatment in childhood asthma (FVC change), Waist circumference adjusted for body mass index, Self-reported math ability, Height; ENSG00000172482: Blood metabolite levels, Height; ENSG00000172478: Major depressive disorder; ENSG00000162804: Sex hormone-binding globulin levels, Blood protein levels, Growth-regulated protein alpha levels, Pharmacokinetics of antipsychotic drugs in severe mental disorder (concentration drug ratio),

					ENSG00000146205, ENSG00000115677, ENSG00000168385, ENSG00000006607	SEPT2, FARP2		Toxicity response to radiotherapy in prostate cancer (hematuria) (time to event), Height; ENSG00000122085: Sex hormone-binding globulin levels, Blood protein levels, Pharmacokinetics of antipsychotic drugs in severe mental disorder (concentration drug ratio); ENSG00000115687: Height; ENSG00000115685: ; ENSG00000146205: Reaction time; ENSG00000115677: Chronic lymphocytic leukemia, Fibrinogen levels, HDL cholesterol levels, Apolipoprotein B levels, Apolipoprotein A1 levels, Male-pattern baldness, Waist circumference adjusted for BMI (adjusted for smoking behaviour), Waist circumference adjusted for BMI (joint analysis main effects and smoking interaction), Waist circumference adjusted for BMI in non-smokers, Waist circumference adjusted for body mass index, Intraocular pressure, Height, Balding type 1, Eosinophil counts; ENSG00000168385: Height, Cerebrospinal fluid immune biomarker levels, Male-pattern baldness, Balding type 1, Lung function (FEV1/FVC); ENSG00000006607: Chronic lymphocytic leukemia, Prostate cancer, Vitiligo, Triglyceride levels, Cerebrospinal fluid immune biomarker levels, Disability (impaired activities of daily living), C-reactive protein levels, Fibrinogen levels, Fibrinogen, Red blood cell count, Low density lipoprotein cholesterol levels, Systolic blood pressure, Educational attainment (MTAG), Educational attainment (years of education), White blood cell count
11	44634764	45073989	44838496	1.42E-10	ENSG00000085117, ENSG000001	CD82, **TSPAN18**,	ENSG00000157570: Schizophrenia, Obstructive sleep apnea trait (average	ENSG00000085117: Hemostatic factors and hematological phenotypes, Red cell distribution width, White matter

					57570, ENSG00000175274	TP53I11	respiratory event duration), Intraocular pressure, Cortical brain region measurements (area, volume and thickness)	microstructure (axial diusivities), White matter microstructure (mean diusivities), White matter microstructure (radial diusivities), Lymphocyte counts, DNA methylation variation (age effect); ENSG00000157570: Schizophrenia, Obstructive sleep apnea trait (average respiratory event duration), Intraocular pressure, Cortical brain region measurements (area, volume and thickness); ENSG00000175274:
7	95947959	96347959	96147959	2.23E-10	ENSG00000004864, ENSG00000197851, ENSG00000127922	SLC25A13, **C7orf76**, SHFM1	ENSG00000197851:	ENSG00000004864: Height, Pork consumption, Oily fish consumption; ENSG00000197851: ; ENSG00000127922: Bone mineral density (hip), Bone mineral density (spine), Femoral neck bone mineral density, Total body bone mineral density, Heel bone mineral density, Total body bone mineral density (age 45-60), Total body bone mineral density (age over 60), Serum platinum levels after completion of cisplatin chemotherapy, Bone mineral density, Nicotine glucouronidation, Chin dimples, Lumbar spine bone mineral density, Bone ultrasound measurement (velocity of sound), Fractures, Cortical surface area, Facial morphology traits (63 three-dimensional facial segments)
15	93385096	93786062	93586062	3.83E-10	ENSG00000173575, ENSG00000182175	CHD2, RGMA	ENSG00000182175: HDL cholesterol levels x short total sleep time interaction (2df test), Blood protein levels, Heel bone mineral density	ENSG00000173575: IgG glycosylation, Schizophrenia, General cognitive ability, Pulse pressure, Mean corpuscular volume, Cognitive performance (MTAG), Self-reported math ability, Self-reported math ability (MTAG), Educational attainment (years of education), Educational attainment (MTAG), Highest math class taken, Highest math class taken (MTAG), Mean corpuscular hemoglobin, Menarche (age at onset); ENSG00000182175: HDL cholesterol levels x short total sleep time

								interaction (2df test), Blood protein levels, Heel bone mineral density
13	88170532	88668426	88411593	4.40E-10	ENSG00000165300	SLITRK5	ENSG00000165300:	ENSG00000165300:
5	31725428	32125428	31925428	4.77E-10	ENSG00000133401, ENSG00000113384	PDZD2, GOLPH3	ENSG00000133401: Obesity-related traits, Myocardial infarction, Height, Chronic obstructive pulmonary disease, Vertical cup-disc ratio (adjusted for vertical disc diameter), Vertical cup-disc ratio (multi-trait analysis), Eotaxin levels, Renal cell carcinoma, Bipolar disorder (body mass index interaction), Response to serotonin reuptake inhibitors in major depressive disorder (plasma drug and metabolite levels), Optic disc size, Vertical cup-disc ratio, Adolescent idiopathic scoliosis, Interleukin-7 levels, Colorectal cancer, Metabolite levels, Corpus callosum central volume, Working memory, Breast cancer specific mortality in estrogen receptor positive breast cancer, Self-reported childhood asthma in adult smokers, Heart rate in heart failure with reduced ejection fraction	ENSG00000133401: Obesity-related traits, Myocardial infarction, Height, Chronic obstructive pulmonary disease, Vertical cup-disc ratio (adjusted for vertical disc diameter), Vertical cup-disc ratio (multi-trait analysis), Eotaxin levels, Renal cell carcinoma, Bipolar disorder (body mass index interaction), Response to serotonin reuptake inhibitors in major depressive disorder (plasma drug and metabolite levels), Optic disc size, Vertical cup-disc ratio, Adolescent idiopathic scoliosis, Interleukin-7 levels, Colorectal cancer, Metabolite levels, Corpus callosum central volume, Working memory, Breast cancer specific mortality in estrogen receptor positive breast cancer, Self-reported childhood asthma in adult smokers, Heart rate in heart failure with reduced ejection fraction; ENSG00000113384: Height
2	211498206	21206 1871	21186 1871	5.45E-10	ENSG00000021826	CPS1	ENSG00000021826: Chronic kidney disease, Fibrinogen, Body mass index in asthmatics, Homocysteine levels, Metabolite levels, Glomerular filtration rate	ENSG00000021826: Chronic kidney disease, Fibrinogen, Body mass index in asthmatics, Homocysteine levels, Metabolite levels, Glomerular filtration rate (creatinine), Betaine levels in individuals undergoing cardiac evaluation,

							(creatinine), Betaine levels in individuals undergoing cardiac evaluation, Glomerular filtration rate in non diabetics (creatinine), Body mass index, Eosinophil percentage of white cells, Mean corpuscular volume, Platelet count, Macular telangiectasia type 2, Metabolite levels (small molecules and protein measures), Plateletcrit, Amino acid levels, Fibrinogen levels, Mean corpuscular hemoglobin, Creatinine levels, Alanine transaminase levels, Serum metabolite concentrations in chronic kidney disease, Urinary metabolite levels in chronic kidney disease, Urinary metabolite modules (eigenmetabolites) in chronic kidney disease, Fat-free mass, HDL cholesterol levels, Appendicular lean mass, Apolipoprotein A1 levels, HDL cholesterol, Plasma homocysteine levels (post-methionine load test), Blood metabolite levels, Serum metabolite levels, Plasma free amino acid levels (adjusted for twenty other PFAAs), Serum 25-Hydroxyvitamin D levels, Eosinophil counts, HDL cholesterol levels x alcohol consumption (regular vs non-regular drinkers) interaction	Glomerular filtration rate in non diabetics (creatinine), Body mass index, Eosinophil percentage of white cells, Mean corpuscular volume, Platelet count, Macular telangiectasia type 2, Metabolite levels (small molecules and protein measures), Plateletcrit, Amino acid levels, Fibrinogen levels, Mean corpuscular hemoglobin, Creatinine levels, Alanine transaminase levels, Serum metabolite concentrations in chronic kidney disease, Urinary metabolite levels in chronic kidney disease, Urinary metabolite modules (eigenmetabolites) in chronic kidney disease, Fat-free mass, HDL cholesterol levels, Appendicular lean mass, Apolipoprotein A1 levels, HDL cholesterol, Plasma homocysteine levels (post-methionine load test), Blood metabolite levels, Serum metabolite levels, Plasma free amino acid levels (adjusted for twenty other PFAAs), Serum 25-Hydroxyvitamin D levels, Eosinophil counts, HDL cholesterol levels x alcohol consumption (regular vs non-regular drinkers) interaction (2df), Eosinophil percentage of granulocytes, Mean platelet volume, Urinary metabolites, Glomerular filtration rate, Estimated glomerular filtration rate, Blood urea nitrogen levels, Estimated glomerular filtration rate in diabetes, Estimated glomerular filtration rate in non-diabetics, HDL cholesterol levels in current drinkers, HDL cholesterol levels x alcohol consumption (drinkers vs non-drinkers) interaction (2df), Blood protein levels, Urinary albumin-to-creatinine ratio, Red blood cell count, Red cell distribution width, White blood cell count, Neutrophil count, Urinary albumin excretion (no hypertensive medication), Urinary albumin

							(2df), Eosinophil percentage of granulocytes, Mean platelet volume, Urinary metabolites, Glomerular filtration rate, Estimated glomerular filtration rate, Blood urea nitrogen levels, Estimated glomerular filtration rate in diabetes, Estimated glomerular filtration rate in non-diabetics, HDL cholesterol levels in current drinkers, HDL cholesterol levels x alcohol consumption (drinkers vs non-drinkers) interaction (2df), Blood protein levels, Urinary albumin-to-creatinine ratio, Red blood cell count, Red cell distribution width, White blood cell count, Neutrophil count, Urinary albumin excretion (no hypertensive medication), Urinary albumin excretion, Glycine levels, Lymphocyte counts, Systolic blood pressure, Urinary potassium to creatinine ratio, Urinary sodium to creatinine ratio, Serum uric acid levels, Urate levels, Height	excretion, Glycine levels, Lymphocyte counts, Systolic blood pressure, Urinary potassium to creatinine ratio, Urinary sodium to creatinine ratio, Serum uric acid levels, Urate levels, Height
12	133566801	133966801	133766801	6.71E-10	ENSG00000198393, ENSG00000198040, ENSG00000196387, ENSG00000214029,	ZNF26, ZNF84, ZNF140, ZNF891,	ENSG00000090612: Waist-hip ratio, Waist-to-hip ratio adjusted for BMI, Type 2 diabetes, Cardiovascular disease	ENSG00000198393: ; ENSG00000198040: Heparin-induced thrombocytopenia; ENSG00000196387: ; ENSG00000214029: Refractive error, Waist-to-hip ratio adjusted for BMI, Intraocular pressure; ENSG00000256223: Type 2 diabetes; ENSG00000090612: Waist-hip ratio, Waist-to-hip ratio adjusted for BMI, Type 2

					ENSG00000256223, ENSG00000090612, ENSG00000227059	ZNF10, **ZNF268**, ANHX		diabetes, Cardiovascular disease; ENSG00000227059:
9	113454364	113913765	113713765	7.08E-10	ENSG00000030304, ENSG00000198121	MUSK, **LPAR1**	ENSG00000198121: Corneal structure, Post bronchodilator FEV1 in COPD, Eosinophil percentage of white cells, Eosinophil counts, Eosinophil percentage of granulocytes, Metabolite levels, Central corneal thickness, Neutrophil percentage of granulocytes, Sum eosinophil basophil counts, Pursuit maintenance gain, Height, Colorectal cancer or advanced adenoma	ENSG00000030304: Heel bone mineral density, Body mass index, Plasma factor V levels in venous thrombosis (conditioned on rs6027), Maximum stenosis, Mean degree of stenosis, Height; ENSG00000198121: Corneal structure, Post bronchodilator FEV1 in COPD, Eosinophil percentage of white cells, Eosinophil counts, Eosinophil percentage of granulocytes, Metabolite levels, Central corneal thickness, Neutrophil percentage of granulocytes, Sum eosinophil basophil counts, Pursuit maintenance gain, Height, Colorectal cancer or advanced adenoma
4	38047254	38736155	38260451	1.39E-09	ENSG00000065882, ENSG00000109787	TBC1D1, KLF3	ENSG00000065882: Amyotrophic lateral sclerosis in C9orf72 mutation negative individuals, Periodontal microbiota, Facial morphology (factor 21, depth of nasal alae), Neutrophil percentage of white cells, Lymphocyte percentage of white cells, Metabolite levels, Verbal declarative memory, Reaction time, Lymphocyte counts, Weight, Response to SSRI (symptom remission), Response to antidepressants (symptom remission), White blood cell count, Eosinophil counts	ENSG00000065882: Amyotrophic lateral sclerosis in C9orf72 mutation negative individuals, Periodontal microbiota, Facial morphology (factor 21, depth of nasal alae), Neutrophil percentage of white cells, Lymphocyte percentage of white cells, Metabolite levels, Verbal declarative memory, Reaction time, Lymphocyte counts, Weight, Response to SSRI (symptom remission), Response to antidepressants (symptom remission), White blood cell count, Eosinophil counts; ENSG00000109787: Eosinophil percentage of white cells, Eosinophil counts, White blood cell count, Sum eosinophil basophil counts, Eosinophil percentage of granulocytes, Neutrophil percentage of granulocytes, Lymphocyte counts, Body mass index, Hand grip strength, Mean

								platelet volume, Mean corpuscular hemoglobin, Red cell distribution width
6	170396266	17079 6266	17059 6266	1.54E-09	ENSG00000198719, ENSG00000112584	DLL1, FAM120B	ENSG00000198719: General risk tolerance (MTAG)	ENSG00000198719: General risk tolerance (MTAG); ENSG00000112584: General risk tolerance (MTAG), Idiopathic dilated cardiomyopathy, Paracentral lobule volume, Menarche (age at onset)
10	94978044	95389 525	95189 525	1.57E-09	ENSG00000138119, ENSG00000138180, ENSG00000186188, ENSG00000138207, ENSG00000095464	MYOF, CEP55, FFAR4, RBP4, PDE6C	ENSG00000138119: Gut microbiome composition (summer), Facial morphology (factor 17, height of vermillion upper lip), Cerebrospinal fluid immune biomarker levels, Metabolite levels	ENSG00000138119: Gut microbiome composition (summer), Facial morphology (factor 17, height of vermillion upper lip), Cerebrospinal fluid immune biomarker levels, Metabolite levels; ENSG00000138180: Lobe attachment (rater-scored or self-reported), Height; ENSG00000186188: Retinol levels, Optic disc area, Waist-to-hip ratio adjusted for BMI, Blood protein levels, Waist-hip ratio, White blood cell count; ENSG00000138207: Optic disc area, Blood protein levels; ENSG00000095464: Preschool internalizing problems, Urinary tract infection frequency
1	168973787	16970 1700	16945 0264	1.71E-09	ENSG00000143153, ENSG00000143156, ENSG00000117475, ENSG00000117477, ENSG00000117479, ENSG00000198734, ENSG00000174175, ENSG00000000460, ENSG00000188404,	ATP1B1, NME7, BLZF1, CCDC181, **SLC19A2**, F5, SELP, C1orf112, SELL, SELE	ENSG00000117479: QT interval	ENSG00000143153: QT interval, Coronary artery disease, Venous thromboembolism, Pulse pressure, Electrocardiographic traits, Systolic blood pressure; ENSG00000143156: Venous thromboembolism, D-dimer levels, Coronary artery disease, QT interval, Pulse pressure, Systolic blood pressure, Mumps, QT dynamics during exercise; ENSG00000117475: Blood protein levels; ENSG00000117477: ; ENSG00000117479: QT interval; ENSG00000198734: Hippocampal atrophy, Activated partial thromboplastin time, Venous thromboembolism, Hemostatic factors and hematological phenotypes, Uric acid levels, Inflammatory bowel disease, Thrombosis, Ischemic stroke, Optic disc area, Blood

					ENSG00000007908			protein levels, Prothrombin time, Vertical cup-disc ratio (multi-trait analysis), Cytokine network levels (multivariate analysis), CTACK levels, Optic disc size, Vertical cup-disc ratio, Peripheral artery disease, Stem cell factor levels, Medication use (antithrombotic agents); ENSG00000174175: Soluble levels of adhesion molecules, Activated partial thromboplastin time, Blood protein levels, Optic disc size, Late-onset Alzheimer's disease; ENSG00000000460: Venous thromboembolism, Acne (severe), Blood protein levels, Intrinsic epigenetic age acceleration, Amyotrophic lateral sclerosis, Tonsillectomy, White blood cell count, Monocyte count, Cardiac Troponin-T levels, Age at menopause, Eosinophil counts; ENSG00000188404: Blood protein levels, Amyotrophic lateral sclerosis, Monocyte count, Eosinophil counts; ENSG00000007908: White blood cell count, Blood protein levels, Age at menopause

Table S4d.1 Eurasia k3

chr	start	end	bestpos	Min(P)	ensembl	hgnc	disease_trait_best	disease_trait
6	134192815	134628278	134392815	8.49E-16	ENSG00000118526, ENSG00000028839, ENSG00000146411, ENSG00000118515	TCF21, TBPL1, **SLC2A12**, SGK1	ENSG00000146411: Coronary artery disease, High chromosomal aberration frequency (chromosome type), FEV1, Lung function (FVC). Mean corpuscular haemoglobin.	ENSG00000118526: Coronary heart disease, Coronary artery disease or ischemic stroke, Coronary artery disease, Coronary artery disease or large artery stroke, PR interval, Medication use (diuretics), Lung function (FEV1/FVC); ENSG00000028839: Mean corpuscular hemoglobin; ENSG00000146411: Coronary artery disease, High chromosomal aberration frequency (chromosome type), FEV1, Lung function (FVC); ENSG00000118515: Immune reponse to smallpox (secreted IFN-alpha), Alzheimer disease and age of onset, Pelvic organ

								prolapse, Pelvic organ prolapse (moderate/severe), Blond vs. brown/black hair color, Schizophrenia (inflammation and infection response interaction), Body mass index (smoking years interaction), Metabolite levels, Adolescent idiopathic scoliosis, Hair color
15	48226484	48633494	48433494	2.09E-15	ENSG00000188467, ENSG00000104177, ENSG00000233932, ENSG00000074803, ENSG00000128951	**SLC24A5, MYEF2**, CTXN2, SLC12A1, DUT	ENSG00000188467: Hair color, Skin pigmentation, Body mass index, Skin, hair and eye pigmentation (multivariate analysis), Eye color, Eye color (brightness), Eye color (saturation), Iris color (a* coordinate), Skin reflectance (Melanin index), Iris color (b* coordinate)	ENSG00000188467: Hair color, Skin pigmentation, Body mass index, Skin, hair and eye pigmentation (multivariate analysis), Eye color, Eye color (brightness), Eye color (saturation), Iris color (a* coordinate), Skin reflectance (Melanin index), Iris color (b* coordinate); ENSG00000104177: Skin pigmentation, Hair color; ENSG00000233932: ; ENSG00000074803: Systemic lupus erythematosus, Longevity, Skin reflectance (Melanin index), Lymphocyte counts; ENSG00000128951: Protein quantitative trait loci
2	135084038	137229668	135479980	5.49E-11	ENSG00000152127, ENSG00000152128, ENSG00000153086, ENSG00000082258, ENSG00000176601, ENSG00000115839, ENSG00000121988, ENSG00000048991, ENSG00000144224, ENSG00000115850, ENSG00000076003, ENSG00000115866, ENSG00000121966	MGAT5, **TMEM163**, ACMSD, CCNT2, MAP3K19, RAB3GAP1, ZRANB3, R3HDM1, UBXN4, LCT, MCM6, DARS, CXCR4	ENSG00000152128: Large artery stroke, Neuroticism, Parkinson's disease or first degree relation to individual with Parkinson's disease, HDL cholesterol levels, Cutaneous melanoma or hair colour, Asthma x air pollution interaction (2df), Hematocrit, Blond vs. brown/black hair color, Spatial memory, Hemoglobin concentration, Low density lipoprotein cholesterol levels, Self-reported math ability, Self-reported math ability (MTAG), Red blood cell count, Hair color	ENSG00000152127: Multiple sclerosis (severity), Subcutaneous adipose tissue, Post bronchodilator FEV1/FVC ratio, Serum alkaline phosphatase levels, N-glycan levels, Blood protein levels, Chronic lymphocytic leukemia or systemic lupus erythematosus, Marginal zone lymphoma or systemic lupus erythematosus, Systemic lupus erythematosus, Eosinophil counts; ENSG00000152128: Large artery stroke, Neuroticism, Parkinson's disease or first degree relation to individual with Parkinson's disease, HDL cholesterol levels, Cutaneous melanoma or hair colour, Asthma x air pollution interaction (2df), Hematocrit, Blond vs. brown/black hair color, Spatial memory, Hemoglobin concentration, Low density lipoprotein cholesterol levels, Self-reported math ability, Self-reported math ability (MTAG), Red blood cell count, Hair color; ENSG00000153086: Obesity-related traits, LDL cholesterol levels, Apolipoprotein B levels, Diisocyanate-induced asthma, Blood metabolite levels, Hematocrit, Free thyroxine concentration, Red blood cell count, Hand grip strength, Diastolic blood pressure; ENSG00000082258: Age at menopause; ENSG00000176601: Colonoscopy-negative controls vs population controls, Corneal structure, Hematocrit, Hemoglobin concentration, Mean corpuscular hemoglobin; ENSG00000115839: Cholesterol, total, Body mass index, Blood metabolite levels, HDL cholesterol levels x alcohol consumption (regular vs non-regular drinkers) interaction

								(2df), LDL cholesterol levels in current drinkers, LDL cholesterol levels x alcohol consumption (drinkers vs non-drinkers) interaction (2df), LDL cholesterol levels x alcohol consumption (regular vs non-regular drinkers) interaction (2df), LDL cholesterol levels, Gut microbiota (bacterial taxa, rank normal transformation method), Sudden cardiac arrest in coronary artery disease; ENSG00000121988: Mosquito bite size, Low density lipoprotein cholesterol levels, Hip circumference, Waist circumference adjusted for body mass index, Sudden cardiac arrest in coronary artery disease, Type 2 diabetes, Height; ENSG00000048991: Mosquito bite size, Blood protein levels, Urinary metabolite levels in chronic kidney disease, Height, LDL cholesterol levels x short total sleep time interaction (2df test), HDL cholesterol levels, Apolipoprotein A1 levels, Red cell distribution width, Hand grip strength; ENSG00000144224: Cholesterol, total, Corneal structure; ENSG00000115850: White blood cell count, Blood protein levels, Osteoarthritis (self-reported); ENSG00000076003: Body mass index, Total cholesterol change in response to fenofibrate in statin-treated type 2 diabetes, Blood protein levels, Hip circumference, 1,5-anhydroglucitol levels, Gut microbiota (bacterial taxa, rank normal transformation method); ENSG00000115866: Mosquito bite size, White blood cell count, Neutrophil count, Monocyte count, Total cholesterol change in response to fenofibrate in statin-treated type 2 diabetes, Systemic lupus erythematosus; ENSG00000121966: Tonsillectomy
5	33699690	34164938	33958959	2.30E-10	ENSG00000151388, ENSG00000182631, ENSG00000164175, ENSG00000242110, ENSG00000082196	ADAMTS12, RXFP3, **SLC45A2**, AMACR, C1QTNF3	ENSG00000164175: Tanning, Skin pigmentation, Black vs. blond hair color, Black vs. red hair color, Melanoma, Hair color, Eye color, Squamous cell carcinoma, Skin colour saturation, Perceived skin darkness, Skin sensitivity to sun, Black vs. non-black hair color, Skin aging (microtopography measurement), Basal cell carcinoma, Cutaneous melanoma or hair colour, Cutaneous malignant melanoma, Nevus	ENSG00000151388: Stroke (pediatric), Mortality in heart failure, Aspartate aminotransferase levels in excessive alcohol consumption, Neurofibrillary tangles, Adolescent idiopathic scoliosis, Height, Hair color; ENSG00000182631: ; ENSG00000164175: Tanning, Skin pigmentation, Black vs. blond hair color, Black vs. red hair color, Melanoma, Hair color, Eye color, Squamous cell carcinoma, Skin colour saturation, Perceived skin darkness, Skin sensitivity to sun, Black vs. non-black hair color, Skin aging (microtopography measurement), Basal cell carcinoma, Cutaneous melanoma or hair colour, Cutaneous malignant melanoma, Nevus count or cutaneous melanoma, Rosacea symptom severity, Low tan response, Skin, hair and eye pigmentation (multivariate

							count or cutaneous melanoma, Rosacea symptom severity, Low tan response, Skin, hair and eye pigmentation (multivariate analysis), Brown vs. black hair color, Blond vs. brown/black hair color, Cutaneous squamous cell carcinoma, Eye color (saturation), Eye color (brightness), Monobrow, Keratinocyte cancer (MTAG), Eye color traits, Skin pigmentation traits, Hair morphology traits, Sunburns	analysis), Brown vs. black hair color, Blond vs. brown/black hair color, Cutaneous squamous cell carcinoma, Eye color (saturation), Eye color (brightness), Monobrow, Keratinocyte cancer (MTAG), Eye color traits, Skin pigmentation traits, Hair morphology traits, Sunburns; ENSG00000242110: Blond vs. brown/black hair color, Longevity; ENSG00000082196: Waist-to-hip ratio adjusted for BMI, Waist-hip ratio
2	85369379	85885211	85592841	2.81E-10	ENSG00000152284, ENSG00000152291, ENSG00000042445, ENSG00000115459, ENSG00000042493, ENSG00000152292, ENSG00000168906, ENSG00000115486, ENSG00000118640, ENSG00000168899, ENSG00000168894, ENSG00000168890, ENSG00000168883, ENSG00000168887, ENSG00000168878	TCF7L1, TGOLN2, RETSAT, **ELMOD3**, CAPG, SH2D6, MAT2A, GGCX, **VAMP8**, **VAMP5**, RNF181, TMEM150A, USP39, C2orf68, SFTPB	ENSG00000115459: Blood protein levels.	ENSG00000152284: Total body bone mineral density, Pulse pressure, Heel bone mineral density, Cervical cancer, Red blood cell count, Lung function (FEV1/FVC), Systolic blood pressure; ENSG00000152291: Ear protrusion, White blood cell count; ENSG00000042445: ; ENSG00000115459: Blood protein levels; ENSG00000042493: ; ENSG00000152292: Mean platelet volume, Platelet count; ENSG00000168906: Basophil count, Coronary artery disease; ENSG00000115486: Coronary artery disease (myocardial infarction, percutaneous transluminal coronary angioplasty, coronary artery bypass grafting, angina or chromic ischemic heart disease), Triglyceride levels, Coronary artery disease, Eosinophil counts; **ENSG00000118640: Eosinophil counts, Prostate cancer, Fat-free mass, Sum eosinophil basophil counts, Coronary artery disease; ENSG00000168899: Parental longevity (father's age at death), Adolescent idiopathic scoliosis, Height;** ENSG00000168894: ; ENSG00000168890: White blood cell count; ENSG00000168883: Neurofibrillary tangles; ENSG00000168887: ; ENSG00000168878: Blood protein levels
3	129487889	129894124	129694124	4.23E-10	ENSG00000172765, ENSG00000170893	TMCC1, **TRH**	ENSG00000170893: waist-hip circumference, Sitting height, Chronic lower respiratory diseases, asthma. rs10934899	ENSG00000172765: Height, Type 2 diabetes, Waist-to-hip ratio adjusted for BMI; ENSG00000170893:
4	38553198	38965720	38765720	2.95E-09	ENSG00000109787, ENSG00000174123, ENSG00000174125,	KLF3, **TLR10**, TLR1, TLR6, FAM114A1	ENSG00000174123: Peripheral arterial disease (traffic-related air pollution interaction), Asthma or	ENSG00000109787: Eosinophil percentage of white cells, Eosinophil counts, White blood cell count, Sum eosinophil basophil counts, Eosinophil percentage of granulocytes, Neutrophil percentage of granulocytes, Lymphocyte counts,

					ENSG00000174130, ENSG00000197712		allergic disease (pleiotropy), Adolescent idiopathic scoliosis	Body mass index, Hand grip strength, Mean platelet volume, Mean corpuscular hemoglobin, Red cell distribution width; ENSG00000174123: Peripheral arterial disease (traffic-related air pollution interaction), Asthma or allergic disease (pleiotropy), Adolescent idiopathic scoliosis; ENSG00000174125: Coronary artery calcified atherosclerotic plaque score in type 2 diabetes, Limited cutaneous systemic scleroderma, Diabetes in response to antihypertensive drug treatment (treatment strategy interaction), Asthma, Allergic sensitization, Self-reported allergy, Asthma and hay fever, Alcohol consumption, Asthma (childhood onset), Allergic rhinitis, Hay fever and/or eczema, Allergy, Allergic disease (asthma, hay fever or eczema), Breast cancer, Asthma onset (childhood vs adult), Composite immunoglobulin trait (IgG/IgM), DNA methylation variation (age effect), Asthma (age of onset), Eczema, Respiratory diseases; ENSG00000174130: Allergic disease (asthma, hay fever or eczema); ENSG00000197712: Alcohol dependence, Moderate or severe diarrhoea in darapladib-treated cardiovascular disease (time to event), Red blood cell count
1	227020437	227877723	227407855	3.77E-09	ENSG00000143801, ENSG00000163050, ENSG00000143776, ENSG00000181450	PSEN2, ADCK3, **CDC42BPA**, ZNF678	ENSG00000143776: Optic disc area, Optic cup area, Blond vs. brown/black hair color, Metabolite levels, Diastolic blood pressure, Myeloid white cell count, Blood urea nitrogen levels, White blood cell count, Neutrophil count, Cardiovascular death or myocardial infarction in response to clopidogrel treatment, Smoking initiation (ever regular vs never regular) (MTAG), Hair color	ENSG00000143801: Worry, Heel bone mineral density; ENSG00000163050: Granulocyte percentage of myeloid white cells, Cataracts (operation), Neutrophil count, Lymphocyte percentage of white cells, Sum basophil neutrophil counts, Waist-to-hip ratio adjusted for BMI, Worry, Sum neutrophil eosinophil counts, Granulocyte count, Gut microbiota (bacterial taxa, hurdle binary method), PR interval, Waist circumference adjusted for body mass index, Estimated glomerular filtration rate, Highest math class taken; ENSG00000143776: Optic disc area, Optic cup area, Blond vs. brown/black hair color, Metabolite levels, Diastolic blood pressure, Myeloid white cell count, Blood urea nitrogen levels, White blood cell count, Neutrophil count, Cardiovascular death or myocardial infarction in response to clopidogrel treatment, Smoking initiation (ever regular vs never regular) (MTAG), Hair color; ENSG00000181450: Height, Hip circumference adjusted for BMI, Body fat distribution (arm fat ratio), Body fat distribution (leg fat ratio), Body fat distribution (trunk fat ratio), Insular cortex volume, Waist circumference adjusted for body mass index

21	26691192	27196935	26935513	1.27E-08	ENSG00000154719, ENSG00000154721, ENSG00000154723, ENSG00000154727	**MRPL39, JAM2,** ATP5J, GABPA	ENSG00000154719: pork intake, arthrosis, Comparative height size at 10, Abdominal hernia	ENSG00000154719: ; ENSG00000154721: Longitudinal change in brain amyloid TP burden, Age at loss of ambulation in Duchenne muscular dystrophy; ENSG00000154723: Alzheimer's disease in hypertension-negative individuals; ENSG00000154727: Gout (normal type), Nicotine dependence symptom count
9	27009422	27434948	27209422	1.56E-08	ENSG00000096872, ENSG00000120156, ENSG00000120160, ENSG00000120162	IFT74, **TEK**, EQTN, MOB3B	ENSG00000120156: Comparative height at 10. Sitting height. IgG glycosylation, Coronary artery disease, Blood protein levels, Blood protein levels in cardiovascular risk, G, Trans fatty acid levels, Endothelial growth factor levels, Cognitive decline (age-related), Clostridium difficile infection in multiple myeloma.	ENSG00000096872: Emphysema annual change measurement in smokers (adjusted lung density); ENSG00000120156: IgG glycosylation, Coronary artery disease, Blood protein levels, Blood protein levels in cardiovascular risk, Schizophrenia, Trans fatty acid levels, Endothelial growth factor levels, Cognitive decline (age-related), Clostridium difficile infection in multiple myeloma; ENSG00000120160: ; ENSG00000120162: Response to TNF antagonist treatment, Urinary symptoms in response to radiotherapy in prostate cancer, Amyotrophic lateral sclerosis, Plasma trimethylamine N-oxide levels, Breast cancer specific mortality in estrogen receptor negative breast cancer, Height, Hair color
15	38464638	38992430	38792430	1.65E-08	ENSG00000166068, ENSG00000171262, ENSG00000172575, ENSG00000175779	SPRED1, FAM98B, **RASGRP1**, C15orf53	ENSG00000172575: Type 1 diabetes, Crohn's disease, Rheumatoid arthritis (ACPA-positive), Rheumatoid arthritis, Multiple sclerosis, Type 2 diabetes, Carboplatin disposition in epthelial ovarian cancer, Autoimmune thyroid disease, Medication use (thyroid preparations), Autoimmune traits (pleiotropy), Autoimmune traits, Hypothyroidism	ENSG00000166068: Birth weight, HDL cholesterol levels x long total sleep time interaction (2df test), Adolescent idiopathic scoliosis, Cognitive performance (processing speed); ENSG00000171262: Systemic lupus erythematosus; ENSG00000172575: Type 1 diabetes, Crohn's disease, Rheumatoid arthritis (ACPA-positive), Rheumatoid arthritis, Multiple sclerosis, Type 2 diabetes, Carboplatin disposition in epthelial ovarian cancer, Autoimmune thyroid disease, Medication use (thyroid preparations), Autoimmune traits (pleiotropy), Autoimmune traits, Hypothyroidism; ENSG00000175779: Cardiac hypertrophy, Metabolic traits, Bipolar disorder or major depressive disorder, Bipolar disorder, Bipolar disorder and schizophrenia, Metabolite levels, Developmental language disorder, Platelet count, Plateletcrit, Parental extreme longevity (95 years and older), Triglyceride change in response to fenofibrate in statin-treated type 2 diabetes, Bipolar I disorder, Alcohol dependence (age at onset), Response to abacavir-containing treatment in HIV-1 infection (virologic failure), Brain region volumes, Bone mineral density, Age at loss of ambulation in Duchenne muscular dystrophy, Follicle stimulating hormone levels in polycystic ovary syndrome, Chronic kidney disease, Estimated

								glomerular filtration rate, Estimated glomerular filtration rate in diabetes, Triglycerides, Circulating fibroblast growth factor 23 levels, Cortical surface area, Leukocyte telomere length, Lymphocyte counts, Non-alcoholic fatty liver disease activity score, Monocyte count, Systemic lupus erythematosus, anorexia nervosa, attention-deficit/hyperactivity disorder, autism spectrum disorder, bipolar disorder, major depression, obsessive-compulsive disorder, schizophrenia, or Tourette syndrome (pleiotropy), Lung function (FVC), Height
11	61264124	62097073	61604967	2.34E-08	ENSG00000204950, ENSG00000011347, ENSG00000134780, ENSG00000124920, ENSG00000134825, ENSG00000168496, ENSG00000134824, ENSG00000149485, ENSG00000221968, ENSG00000167994, ENSG00000167995, ENSG00000167996, ENSG00000149503, ENSG00000168515, ENSG00000124939, ENSG00000124935, ENSG00000110484, ENSG00000197745	LRRC10B, SYT7, DAGLA, MYRF, TMEM258, FEN1, FADS2, FADS1, FADS3, RAB3IL1, BEST1, FTH1, INCENP, SCGB1D1, SCGB2A1, SCGB1D2, SCGB2A2, SCGB1D4	ENSG00000134824: HDL cholesterol, Triglycerides, Fasting blood glucose, Homeostasis model assessment of beta-cell function, Cholesterol, total, LDL cholesterol, Heart rate, Metabolic syndrome, Resting heart rate, Lipid metabolism phenotypes, Hematology traits, Liver enzyme levels (alkaline phosphatase), Comprehensive strength and appendicular lean mass, Metabolite levels, Response to statin therapy, Metabolic traits, Plasma omega-3 polyunsaturated fatty acid levels (docosapentaenoic acid), Plasma omega-3 polyunsaturated fatty acid levels (alphalinolenic acid), Plasma omega-3 polyunsaturated fatty acid level (eicosapentaenoic acid), Platelet count, Fasting blood glucose (BMI interaction), Inflammatory bowel disease, Plasma omega-6 polyunsaturated fatty acid levels (gamma-linolenic acid), Plasma omega-6 polyunsaturated fatty acid levels (dihomo-gamma-linolenic acid), Plasma omega-6 polyunsaturated fatty acid levels (arachidonic	ENSG00000204950: Diastolic blood pressure, Systolic blood pressure, Systolic blood pressure (cigarette smoking interaction), Diastolic blood pressure (cigarette smoking interaction), Mean arterial pressure, Mean arterial pressure x alcohol consumption interaction (2df test), Diastolic blood pressure x alcohol consumption interaction (2df test), Hypertension, Medication use (agents acting on the renin-angiotensin system), Heel bone mineral density; ENSG00000011347: Intelligence (MTAG), Phosphatidylcholine levels, Cholesteryl ester levels, Educational attainment (years of education), Cognitive performance (MTAG), Cognitive performance; ENSG00000134780: Immune reponse to smallpox (secreted IL-2), Plasma omega-3 polyunsaturated fatty acid levels (docosapentaenoic acid), Plasma omega-6 polyunsaturated fatty acid levels (linoleic acid), 3-hydroxypropylmercapturic acid levels in smokers, Refractive error, Cerebrospinal fluid sTREM-2 levels, Spherical equivalent, C-reactive protein levels, Neuroticism, Positive affect, Depressive symptoms, Well-being spectrum (multivariate analysis), Life satisfaction, Depression, Lung function (FVC); ENSG00000124920: Plasma omega-3 polyunsaturated fatty acid levels (docosapentaenoic acid), Plasma omega-3 polyunsaturated fatty acid levels (alphalinolenic acid), Plasma omega-3 polyunsaturated fatty acid level (eicosapentaenoic acid), Metabolite levels, Crohn's disease, Plasma omega-6 polyunsaturated fatty acid levels (gamma-linolenic acid), Plasma omega-6 polyunsaturated fatty acid levels (dihomo-gamma-linolenic acid), Plasma omega-6 polyunsaturated fatty acid levels (arachidonic acid), Hematocrit, Triglycerides, HDL cholesterol, Glycerophospholipid levels, Total cholesterol levels, Moyamoya disease, Resting heart rate, Red blood cell

							acid), Delta-6 desaturase activity, Plasma omega-6 polyunsaturated fatty acid levels (linoleic acid), Eosinophil counts, C-reactive protein levels or LDL-cholesterol levels (pleiotropy), C-reactive protein levels or HDL-cholesterol levels (pleiotropy), C-reactive protein levels or triglyceride levels (pleiotropy), Metabolite levels (lipid measures), Gestational age at birth (child effect), Plateletcrit, Granulocyte percentage of myeloid white cells, Monocyte percentage of white cells, Glycerophospholipid levels, Sphingolipid levels, Vitiligo, Glycated hemoglobin levels, Total cholesterol levels, Heel bone mineral density, Non-albumin protein levels, Albumin-globulin ratio, Hemoglobin A1c levels, Alanine transaminase levels, Triglyceride levels, Breast milk fatty acid composition (maternal genotype effect), Breast milk fatty acid composition (infant genotype effect), Serum metabolite concentrations in chronic kidney disease, Serum metabolite ratios in chronic kidney disease, Low density lipoprotein cholesterol levels, Height, Apolipoprotein B levels, Pulse pressure, LDL cholesterol levels, LDL cholesterol levels x long total sleep time interaction (2df test), HDL cholesterol levels, Triglyceride levels x short total sleep time interaction (2df test), Serum	count, Serum total protein level, Serum metabolite concentrations in chronic kidney disease, Serum metabolite ratios in chronic kidney disease, Triglyceride levels x long total sleep time interaction (2df test), Serum metabolite levels (CMS), Heel bone mineral density, Hemoglobin levels, Spherical equivalent, Blood metabolite levels, Colorectal cancer, Serum omega-3 polyunsaturated fatty acid concentration in metabolic syndrome, Low density lipoprotein cholesterol levels, Serum metabolite levels, Chronic inflammatory diseases (ankylosing spondylitis, Crohn's disease, psoriasis, primary sclerosing cholangitis, ulcerative colitis) (pleiotropy), Hemoglobin concentration, Vaccenic acid (18:1n-7) levels, Gondoic acid (20:1n-9) levels, LDL cholesterol levels, Iron status biomarkers (total iron binding capacity), Trans fatty acid levels, High density lipoprotein cholesterol levels, Red blood cell fatty acid levels, Colorectal cancer or advanced adenoma, Phosphatidylcholine-ether levels, Phosphatidylethanolamine-ether levels, Asthma (adult onset), Asthma, Nasal polyps, Systolic blood pressure, Triglyceride levels, Glycemic traits (pleiotropy), Educational attainment (years of education), Respiratory diseases; ENSG00000134825: Crohn's disease, Metabolic syndrome, Palmitoleic acid (16:1n-7) levels, Stearic acid (18:0) levels, Oleic acid (18:1n-9) levels, Phospholipid levels (plasma), Plasma omega-3 polyunsaturated fatty acid levels (docosapentaenoic acid), Plasma omega-3 polyunsaturated fatty acid level (eicosapentaenoic acid), Plasma omega-3 polyunsaturated fatty acid levels (alphalinolenic acid), Metabolite levels, Plasma omega-6 polyunsaturated fatty acid levels (gamma-linolenic acid), Plasma omega-6 polyunsaturated fatty acid levels (dihomo-gamma-linolenic acid), Plasma omega-6 polyunsaturated fatty acid levels (arachidonic acid), Hematocrit, Triglycerides, HDL cholesterol, Glycerophospholipid levels, Total cholesterol levels, Resting heart rate, Red blood cell count, Serum total protein level, Irritable mood, Serum metabolite concentrations in chronic kidney disease, Serum metabolite ratios in chronic kidney disease, LDL cholesterol levels x short total sleep time interaction (2df test), Triglyceride levels x long total sleep time interaction (2df test), Serum metabolite levels (CMS), Heel bone mineral density, Hemoglobin levels, Spherical equivalent,

							metabolite levels (CMS), HDL cholesterol levels x long total sleep time interaction (2df test), HDL cholesterol levels x short total sleep time interaction (2df test), HDL cholesterol levels x alcohol consumption (regular vs non-regular drinkers) interaction (2df), Triglyceride levels x alcohol consumption (regular vs non-regular drinkers) interaction (2df), Bipolar I disorder, Apolipoprotein A1 levels, Neutrophil count, Colorectal cancer, QRS duration, Crohn's disease, Age-related disease endophenotypes, Sum basophil neutrophil counts, Red cell distribution width, P wave duration, Blood metabolite levels, Blood metabolite ratios, Rheumatoid arthritis, QT interval, Iron status biomarkers (transferrin levels), Serum metabolite levels, Change in serum metabolite levels (CMS), Serum omega-3 polyunsaturated fatty acid concentration in metabolic syndrome, Serum omega-6 to omega-3 polyunsaturated fatty acid ratio in metabolic syndrome, Serum docosahexaenoic fatty acid concentration in metabolic syndrome, Delta-5 desaturase activity response to n3-polyunsaturated fat supplement, Change in serum metabolite levels, LDL cholesterol x physical activity interaction (2df test), High density lipoprotein	Blood metabolite levels, Colorectal cancer, Serum omega-3 polyunsaturated fatty acid concentration in metabolic syndrome, Serum metabolite levels, Blond vs. brown/black hair color, Low density lipoprotein cholesterol levels, Bipolar disorder or major depressive disorder, Chronic inflammatory diseases (ankylosing spondylitis, Crohn's disease, psoriasis, primary sclerosing cholangitis, ulcerative colitis) (pleiotropy), Hemoglobin concentration, Vaccenic acid (18:1n-7) levels, Gondoic acid (20:1n-9) levels, LDL cholesterol levels, Iron status biomarkers (total iron binding capacity), Trans fatty acid levels, High density lipoprotein cholesterol levels, Carboplatin disposition in epthelial ovarian cancer, Red blood cell fatty acid levels, Colorectal cancer or advanced adenoma, Phosphatidylcholine-ether levels, Phosphatidylcholine levels, Phosphatidylethanolamine-ether levels, Cholesteryl ester levels, Asthma (adult onset), Asthma, Nasal polyps, Triglyceride levels, Glycemic traits (pleiotropy), Educational attainment (years of education), Respiratory diseases, Hair color; ENSG00000168496: Plasma omega-3 polyunsaturated fatty acid levels (docosapentaenoic acid), Plasma omega-3 polyunsaturated fatty acid levels (alphalinolenic acid), Plasma omega-3 polyunsaturated fatty acid level (eicosapentaenoic acid), Platelet count, Inflammatory bowel disease, Crohn's disease, Trans fatty acid levels, Metabolite levels, Red blood cell fatty acid levels, Colorectal cancer; ENSG00000134824: HDL cholesterol, Triglycerides, Fasting blood glucose, Homeostasis model assessment of beta-cell function, Cholesterol, total, LDL cholesterol, Heart rate, Metabolic syndrome, Resting heart rate, Lipid metabolism phenotypes, Hematology traits, Liver enzyme levels (alkaline phosphatase), Comprehensive strength and appendicular lean mass, Metabolite levels, Response to statin therapy, Metabolic traits, Plasma omega-3 polyunsaturated fatty acid levels (docosapentaenoic acid), Plasma omega-3 polyunsaturated fatty acid levels (alphalinolenic acid), Plasma omega-3 polyunsaturated fatty acid level (eicosapentaenoic acid), Platelet count, Fasting blood glucose (BMI interaction), Inflammatory bowel disease, Plasma omega-6 polyunsaturated fatty acid levels (gamma-linolenic acid), Plasma omega-6 polyunsaturated fatty acid levels (dihomo-gamma-linolenic acid), Plasma omega-6 polyunsaturated fatty

							cholesterol levels, Asthma, Bipolar disorder, Male-pattern baldness, Nonatopic asthma, Osteoporosis-related phenotypes (MTAG), Granulocyte count, Age-related diseases, mortality and associated endophenotypes, Sum neutrophil eosinophil counts, Gondoic acid (20:1n-9) levels, Sum eosinophil basophil counts, Myeloid white cell count, White blood cell count, Mean platelet volume, Trans fatty acid levels, Red blood cell fatty acid levels, Plasma omega-6 polyunsaturated fatty acid levels (adrenic acid), Laryngeal squamous cell carcinoma, Lung cancer in ever smokers, Triacylglycerol 56:6 levels, Metabolite risk score for predicting weight gain, Fatty acid desaturase activity (serum), Fatty acid desaturase activity (adipose tissue), Phosphatidylcholine levels, Phosphatidylcholine-ether levels, Phosphatidylethanolamine levels, Lysophosphatidylethanolamine levels, Phosphatidylethanolamine-ether levels, Phosphatidylinositol levels, Cholesteryl ester levels, Sphingomyelin levels, Lysophosphatidylcholine levels, Triacylglyceride levels, HDL cholesterol levels in current drinkers, LDL cholesterol levels in current drinkers, Triglyceride levels in current drinkers, HDL cholesterol levels x alcohol consumption (drinkers vs non-	acid levels (arachidonic acid), Delta-6 desaturase activity, Plasma omega-6 polyunsaturated fatty acid levels (linoleic acid), Eosinophil counts, C-reactive protein levels or LDL-cholesterol levels (pleiotropy), C-reactive protein levels or HDL-cholesterol levels (pleiotropy), C-reactive protein levels or triglyceride levels (pleiotropy), Metabolite levels (lipid measures), Gestational age at birth (child effect), Plateletcrit, Granulocyte percentage of myeloid white cells, Monocyte percentage of white cells, Glycerophospholipid levels, Sphingolipid levels, Vitiligo, Glycated hemoglobin levels, Total cholesterol levels, Heel bone mineral density, Non-albumin protein levels, Albumin-globulin ratio, Hemoglobin A1c levels, Alanine transaminase levels, Triglyceride levels, Breast milk fatty acid composition (maternal genotype effect), Breast milk fatty acid composition (infant genotype effect), Serum metabolite concentrations in chronic kidney disease, Serum metabolite ratios in chronic kidney disease, Low density lipoprotein cholesterol levels, Height, Apolipoprotein B levels, Pulse pressure, LDL cholesterol levels, LDL cholesterol levels x long total sleep time interaction (2df test), HDL cholesterol levels, Triglyceride levels x short total sleep time interaction (2df test), Serum metabolite levels (CMS), HDL cholesterol levels x long total sleep time interaction (2df test), HDL cholesterol levels x short total sleep time interaction (2df test), HDL cholesterol levels x alcohol consumption (regular vs non-regular drinkers) interaction (2df), Triglyceride levels x alcohol consumption (regular vs non-regular drinkers) interaction (2df), Bipolar I disorder, Apolipoprotein A1 levels, Neutrophil count, Colorectal cancer, QRS duration, Crohn's disease, Age-related disease endophenotypes, Sum basophil neutrophil counts, Red cell distribution width, P wave duration, Blood metabolite levels, Blood metabolite ratios, Rheumatoid arthritis, QT interval, Iron status biomarkers (transferrin levels), Serum metabolite levels, Change in serum metabolite levels (CMS), Serum omega-3 polyunsaturated fatty acid concentration in metabolic syndrome, Serum omega-6 to omega-3 polyunsaturated fatty acid ratio in metabolic syndrome, Serum docosahexaenoic fatty acid concentration in metabolic syndrome, Delta-5 desaturase activity response to n3-polyunsaturated fat supplement, Change in serum metabolite levels, LDL cholesterol x physical activity

							drinkers) interaction (2df), Triglyceride levels x alcohol consumption (drinkers vs non-drinkers) interaction (2df), LDL cholesterol levels x alcohol consumption (drinkers vs non-drinkers) interaction (2df), LDL cholesterol levels x alcohol consumption (regular vs non-regular drinkers) interaction (2df), Red blood cell count, Aortic valve stenosis, Sleep duration, Hematocrit, Hemoglobin concentration, Total triglycerides levels, Lymphocyte counts, Mean corpuscular hemoglobin concentration, PR interval, Type 2 diabetes, Mean corpuscular volume, IgA levels, Adult onset asthma or type 2 diabetes, Nonatopic asthma or type 2 diabetes, Adult onset asthma or fasting glucose levels, Nonatopic asthma or fasting glucose levels, Asthma (adult onset), Medication use (thyroid preparations), Medication use (adrenergics, inhalants), Urate levels, Gallstone disease, anorexia nervosa, attention-deficit/hyperactivity disorder, autism spectrum disorder, bipolar disorder, major depression, obsessive-compulsive disorder, schizophrenia, or Tourette syndrome (pleiotropy), QT dynamics during exercise, Balding type 1, Hypothyroidism	interaction (2df test), High density lipoprotein cholesterol levels, Asthma, Bipolar disorder, Male-pattern baldness, Nonatopic asthma, Osteoporosis-related phenotypes (MTAG), Granulocyte count, Age-related diseases, mortality and associated endophenotypes, Sum neutrophil eosinophil counts, Gondoic acid (20:1n-9) levels, Sum eosinophil basophil counts, Myeloid white cell count, White blood cell count, Mean platelet volume, Trans fatty acid levels, Red blood cell fatty acid levels, Plasma omega-6 polyunsaturated fatty acid levels (adrenic acid), Laryngeal squamous cell carcinoma, Lung cancer in ever smokers, Triacylglycerol 56:6 levels, Metabolite risk score for predicting weight gain, Fatty acid desaturase activity (serum), Fatty acid desaturase activity (adipose tissue), Phosphatidylcholine levels, Phosphatidylcholine-ether levels, Phosphatidylethanolamine levels, Lysophosphatidylethanolamine levels, Phosphatidylethanolamine-ether levels, Phosphatidylinositol levels, Cholesteryl ester levels, Sphingomyelin levels, Lysophosphatidylcholine levels, Triacylglyceride levels, HDL cholesterol levels in current drinkers, LDL cholesterol levels in current drinkers, Triglyceride levels in current drinkers, HDL cholesterol levels x alcohol consumption (drinkers vs non-drinkers) interaction (2df), Triglyceride levels x alcohol consumption (drinkers vs non-drinkers) interaction (2df), LDL cholesterol levels x alcohol consumption (drinkers vs non-drinkers) interaction (2df), LDL cholesterol levels x alcohol consumption (regular vs non-regular drinkers) interaction (2df), Red blood cell count, Aortic valve stenosis, Sleep duration, Hematocrit, Hemoglobin concentration, Total triglycerides levels, Lymphocyte counts, Mean corpuscular hemoglobin concentration, PR interval, Type 2 diabetes, Mean corpuscular volume, IgA levels, Adult onset asthma or type 2 diabetes, Nonatopic asthma or type 2 diabetes, Adult onset asthma or fasting glucose levels, Nonatopic asthma or fasting glucose levels, Asthma (adult onset), Medication use (thyroid preparations), Medication use (adrenergics, inhalants), Urate levels, Gallstone disease, anorexia nervosa, attention-deficit/hyperactivity disorder, autism spectrum disorder, bipolar disorder, major depression, obsessive-compulsive disorder, schizophrenia, or Tourette syndrome (pleiotropy), QT dynamics during exercise, Balding type 1, Hypothyroidism;

								ENSG00000149485: HDL cholesterol, Triglycerides, Fasting blood glucose, Homeostasis model assessment of beta-cell function, LDL cholesterol, Heart rate, Metabolic syndrome, Resting heart rate, Lipid metabolism phenotypes, Hematology traits, Comprehensive strength and appendicular lean mass, Metabolite levels, Cholesterol, total, Metabolic traits, Plasma omega-3 polyunsaturated fatty acid levels (alphalinolenic acid), Plasma omega-3 polyunsaturated fatty acid level (eicosapentaenoic acid), Plasma omega-3 polyunsaturated fatty acid levels (docosapentaenoic acid), Fasting blood glucose (BMI interaction), Plasma omega-6 polyunsaturated fatty acid levels (gamma-linolenic acid), Plasma omega-6 polyunsaturated fatty acid levels (dihomo-gamma-linolenic acid), Delta-6 desaturase activity, Plasma omega-6 polyunsaturated fatty acid levels (linoleic acid), Eosinophil counts, Platelet count, C-reactive protein levels or HDL-cholesterol levels (pleiotropy), C-reactive protein levels or triglyceride levels (pleiotropy), Metabolite levels (lipid measures), Granulocyte percentage of myeloid white cells, Monocyte percentage of white cells, Glycerophospholipid levels, Sphingolipid levels, Vitiligo, Total cholesterol levels, Alanine transaminase levels, Triglyceride levels, Breast milk fatty acid composition (maternal genotype effect), Breast milk fatty acid composition (infant genotype effect), Serum metabolite concentrations in chronic kidney disease, Serum metabolite ratios in chronic kidney disease, Low density lipoprotein cholesterol levels, Apolipoprotein B levels, Pulse pressure, LDL cholesterol levels, LDL cholesterol levels x long total sleep time interaction (2df test), HDL cholesterol levels, Triglyceride levels x short total sleep time interaction (2df test), Serum metabolite levels (CMS), HDL cholesterol levels x long total sleep time interaction (2df test), HDL cholesterol levels x short total sleep time interaction (2df test), HDL cholesterol levels x alcohol consumption (regular vs non-regular drinkers) interaction (2df), Triglyceride levels x alcohol consumption (regular vs non-regular drinkers) interaction (2df), Bipolar I disorder, Apolipoprotein A1 levels, Age-related disease endophenotypes, Red cell distribution width, Height, Blood metabolite levels, Blood metabolite ratios, Rheumatoid arthritis, Serum metabolite levels, Change in serum metabolite levels (CMS), Serum omega-3 polyunsaturated fatty acid

								concentration in metabolic syndrome, Serum omega-6 to omega-3 polyunsaturated fatty acid ratio in metabolic syndrome, Serum docosahexaenoic fatty acid concentration in metabolic syndrome, Delta-5 desaturase activity response to n3-polyunsaturated fat supplement, Change in serum metabolite levels, LDL cholesterol x physical activity interaction (2df test), High density lipoprotein cholesterol levels, Asthma, Bipolar disorder, Osteoporosis-related phenotypes (MTAG), Granulocyte count, Age-related diseases, mortality and associated endophenotypes, Sum neutrophil eosinophil counts, Sum eosinophil basophil counts, Myeloid white cell count, White blood cell count, Mean platelet volume, Trans fatty acid levels, Red blood cell fatty acid levels, Plasma omega-6 polyunsaturated fatty acid levels (arachidonic acid), Plasma omega-6 polyunsaturated fatty acid levels (adrenic acid), Laryngeal squamous cell carcinoma, Triacylglycerol 56:6 levels, Metabolite risk score for predicting weight gain, Fatty acid desaturase activity (serum), Fatty acid desaturase activity (adipose tissue), Phosphatidylcholine-ether levels, Lysophosphatidylethanolamine levels, Phosphatidylcholine levels, Phosphatidylethanolamine-ether levels, Phosphatidylinositol levels, Cholesteryl ester levels, Sphingomyelin levels, Lysophosphatidylcholine levels, Triacylglyceride levels, HDL cholesterol levels in current drinkers, LDL cholesterol levels in current drinkers, Triglyceride levels in current drinkers, HDL cholesterol levels x alcohol consumption (drinkers vs non-drinkers) interaction (2df), Triglyceride levels x alcohol consumption (drinkers vs non-drinkers) interaction (2df), LDL cholesterol levels x alcohol consumption (drinkers vs non-drinkers) interaction (2df), LDL cholesterol levels x alcohol consumption (regular vs non-regular drinkers) interaction (2df), Red blood cell count, QT interval, Aortic valve stenosis, Sleep duration, Neutrophil count, Total triglycerides levels, Mean corpuscular hemoglobin concentration, Mean corpuscular volume, IgA levels, Gallstone disease; ENSG00000221968: Sphingolipid levels, Metabolite levels, Phosphatidylcholine levels, Cholesteryl ester levels, Sphingolipid d18:1/d18:2 ratio, Urate levels; ENSG00000167994: Sphingolipid levels, Metabolite levels, Schizophrenia, Mean corpuscular hemoglobin; ENSG00000167995: Plasma omega-3 polyunsaturated fatty

								acid level (eicosapentaenoic acid), Metabolite levels; ENSG00000167996: ; ENSG00000149503: Prostate cancer, Breast cancer; ENSG00000168515: ; ENSG00000124939: ; ENSG00000124935: ; ENSG00000110484: ; ENSG00000197745:
16	66852047	67871804	67473389	5.25E-08	ENSG00000159593, ENSG00000168748, ENSG00000172840, ENSG00000166589, ENSG00000166592, ENSG00000166595, ENSG00000172831, ENSG00000172828, ENSG00000172824, ENSG00000067955, ENSG00000125149, ENSG00000237172, ENSG00000102871, ENSG00000135722, ENSG00000102878, ENSG00000140939, ENSG00000196123, ENSG00000179044, ENSG00000205250, ENSG00000102890, ENSG00000125122, ENSG00000168701, ENSG00000135723, ENSG00000135740, ENSG00000196155, ENSG00000168676, ENSG00000159708, ENSG00000159713, ENSG00000159714, ENSG00000176387, **ENSG00000159720**, ENSG00000159723, ENSG00000039523, ENSG00000102974, ENSG00000159753,	NAE1, CA7, PDP2, CDH16, RRAD, FAM96B, CES2, CES3, CES4A, CBFB, C16orf70, B3GNT9, TRADD, FBXL8, HSF4, NOL3, KIAA0895L, EXOC3L1, E2F4, ELMO3, LRRC29, TMEM208, FHOD1, SLC9A5, PLEKHG4, KCTD19, LRRC36, TPPP3, ZDHHC1, HSD11B2, **ATP6V0D1**, AGRP, FAM65A, CTCF, RLTPR, ACD, PARD6A, ENKD1, C16orf86,	ENSG00000159720: Disease progression in age-related macular degeneration, Medication use (thyroid preparations)	ENSG00000159593: HDL cholesterol levels, Apolipoprotein A1 levels; ENSG00000168748: ; ENSG00000172840: Diastolic blood pressure; ENSG00000166589: Blood protein levels in cardiovascular risk, HDL cholesterol levels x alcohol consumption (regular vs non-regular drinkers) interaction (2df), HDL cholesterol levels in current drinkers, HDL cholesterol levels x alcohol consumption (drinkers vs non-drinkers) interaction (2df); ENSG00000166592: ; ENSG00000166595: ; ENSG00000172831: ; ENSG00000172828: Heel bone mineral density, Blood protein levels; ENSG00000172824: Hemoglobin levels; ENSG00000067955: Mean corpuscular hemoglobin, Breast cancer specific mortality in estrogen receptor positive breast cancer, Blood protein levels, Red cell distribution width; ENSG00000125149: Adolescent idiopathic scoliosis, Height; ENSG00000237172: ; ENSG00000102871: ; ENSG00000135722: ; ENSG00000102878: ; ENSG00000140939: ; ENSG00000196123: Mean corpuscular volume, Heel bone mineral density; ENSG00000179044: Hemoglobin levels, Monocyte count; ENSG00000205250: Mean corpuscular volume, Mean corpuscular hemoglobin; ENSG00000102890: Intraocular pressure; ENSG00000125122: HDL cholesterol, Male-pattern baldness; ENSG00000168701: ; ENSG00000135723: Waist circumference adjusted for body mass index, Waist-to-hip ratio adjusted for BMI; ENSG00000135740: ; ENSG00000196155: Heel bone mineral density, Blood protein levels; ENSG00000168676: Waist circumference adjusted for body mass index, Waist circumference adjusted for BMI (adjusted for smoking behaviour), Waist circumference adjusted for BMI (joint analysis main effects and smoking interaction), Waist circumference adjusted for BMI in non-smokers, Body mass index, Height, Hypothyroidism; ENSG00000159708: Waist circumference adjusted for body mass index, Adolescent idiopathic scoliosis, Blood protein levels, Waist circumference

					ENSG00000102977, ENSG00000102981, ENSG00000124074, ENSG00000159761, ENSG00000141098, ENSG00000141084, ENSG00000102904, ENSG00000102901	GFOD2, RANBP10, TSNAXIP1, CENPT		adjusted for BMI (joint analysis main effects and physical activity interaction), Waist circumference adjusted for BMI in active individuals, Waist-to-hip ratio adjusted for BMI, Waist-hip ratio, Waist-to-hip ratio adjusted for BMI (additive genetic model), Mean corpuscular hemoglobin, Eosinophil counts; ENSG00000159713: ; ENSG00000159714: Bone mineral density; ENSG00000176387: ; ENSG00000159720: Disease progression in age-related macular degeneration, Medication use (thyroid preparations); ENSG00000159723: ; ENSG00000039523: Heel bone mineral density, HDL cholesterol, Hemoglobin levels, Mean platelet volume, Lung function (FVC), Lung function (FEV1/FVC); ENSG00000102974: Emotional recognition, Mean corpuscular hemoglobin concentration; ENSG00000159753: HDL cholesterol levels x long total sleep time interaction (2df test), HDL cholesterol, Myopia (age of diagnosis), Lymphocyte counts, Eosinophil counts; ENSG00000102977: Obesity-related traits, Hemoglobin levels; ENSG00000102981: ; ENSG00000124074: Blood protein levels; ENSG00000159761: ; ENSG00000141098: HDL cholesterol, Male-pattern baldness, High density lipoprotein cholesterol levels; ENSG00000141084: Hematocrit, HDL cholesterol, Empathy quotient, Hemoglobin concentration, High density lipoprotein cholesterol levels; ENSG00000102904: Balding type 1; ENSG00000102901: Red blood cell count, Mean corpuscular hemoglobin
2	154307589	154707589	154507589	6.80E-08	ENSG00000177519	RPRM	ENSG00000177519:	ENSG00000177519:
5	94974878	95385933	95174878	8.07E-08	ENSG00000175449, ENSG00000145757, ENSG00000164292, ENSG00000173221, ENSG00000236882, ENSG00000118985	RFESD, SPATA9, RHOBTB3, GLRX, C5orf27, ELL2	ENSG00000236882:	ENSG00000175449: Blood protein levels, Pharmacokinetics of antiepileptic drugs in severe mental disorder (concentration drug ratio); ENSG00000145757: Blood protein levels, Lung function (FEV1/FVC); ENSG00000164292: Metabolite levels, Adolescent idiopathic scoliosis; ENSG00000173221: Metabolite levels, Adolescent idiopathic scoliosis; ENSG00000236882: ; ENSG00000118985: IgG glycosylation, Multiple myeloma, Multiple myeloma and monoclonal gammopathy, Coronary artery calcified atherosclerotic plaque score in type 2 diabetes, Non-albumin protein levels, Albumin-globulin ratio, B-cell malignancies (chronic lymphocytic leukemia, Hodgkin lymphoma or multiple myeloma) (pleiotropy), Serum total protein level, IgG digalactosylation phenotypes (multivariate analysis), IgG sialylation phenotypes

								(multivariate analysis), Mean corpuscular hemoglobin, IgA levels, Mean corpuscular volume
11	60728079	61159993	60928079	9.36E-08	ENSG00000013725, **ENSG00000110448**, ENSG00000167987, ENSG00000229859, ENSG00000229183, ENSG00000256713, ENSG00000167992, ENSG00000167986, ENSG00000149476, ENSG00000162144, ENSG00000149483, ENSG00000187049	**CD6, CD5, VPS37C**, PGA3, PGA4, PGA5, VWCE, DDB1, DAK, CYB561A3, TMEM138, TMEM216	ENSG00000167987: Rheumatoid arthritis (ACPA-positive), Rheumatoid arthritis, Adolescent idiopathic scoliosis, Periventricular white matter hyperintensities	ENSG00000013725: Multiple sclerosis, Inflammatory bowel disease, Crohn's disease, Ulcerative colitis, Chronic inflammatory diseases (ankylosing spondylitis, Crohn's disease, psoriasis, primary sclerosing cholangitis, ulcerative colitis) (pleiotropy), CD6 levels; ENSG00000110448: ; ENSG00000167987: Rheumatoid arthritis (ACPA-positive), Rheumatoid arthritis, Adolescent idiopathic scoliosis, Periventricular white matter hyperintensities; ENSG00000229859: ; ENSG00000229183: ; ENSG00000256713: Height; ENSG00000167992: ; ENSG00000167986: ; ENSG00000149476: ; ENSG00000162144: ; ENSG00000149483: ; ENSG00000187049:
9	98012608	98509513	98296403	1.14E-07	ENSG00000158169, ENSG00000185920	FANCC, PTCH1	ENSG00000185920: Pulmonary function, Height, Pulmonary function (smoking interaction), Bone mineral density (spine), Bone mineral density (hip), Waist circumference adjusted for body mass index, Hip circumference adjusted for BMI, Nonsyndromic cleft lip with cleft palate, Birth weight, Heel bone mineral density, Neurociticism, Neuroticism, Feeling hurt, Feeling worry, Alanine aminotransferase levels in low alcohol consumption, Appendicular lean mass, Cortical surface area (visual PC2), Birth weight (MTAG), Birth length (MTAG), Monobrow, Intelligence, Worry, Lung function (FVC), Waist circumference adjusted for BMI (adjusted for smoking behaviour), Waist circumference adjusted for BMI (joint analysis main effects and smoking interaction), Waist	ENSG00000158169: Cortical surface area (visual PC2), Hematocrit, Waist circumference adjusted for body mass index, Heel bone mineral density, Height; ENSG00000185920: Pulmonary function, Height, Pulmonary function (smoking interaction), Bone mineral density (spine), Bone mineral density (hip), Waist circumference adjusted for body mass index, Hip circumference adjusted for BMI, Nonsyndromic cleft lip with cleft palate, Birth weight, Heel bone mineral density, Neurociticism, Neuroticism, Feeling hurt, Feeling worry, Alanine aminotransferase levels in low alcohol consumption, Appendicular lean mass, Cortical surface area (visual PC2), Birth weight (MTAG), Birth length (MTAG), Monobrow, Intelligence, Worry, Lung function (FVC), Waist circumference adjusted for BMI (adjusted for smoking behaviour), Waist circumference adjusted for BMI (joint analysis main effects and smoking interaction), Waist circumference adjusted for BMI in non-smokers, 3-month functional outcome in ischaemic stroke (modified Rankin score), Cortical surface area, Depressive symptoms, Positive affect, Offspring birth weight, Lung function (FEV1/FVC), Well-being spectrum (multivariate analysis), White matter microstructure (fractional anisotropy), Type 2 diabetes, Life satisfaction, Macular thickness, Sensitivity to environmental stress and adversity, Cognitive performance, Cognitive performance (MTAG), Highest math class taken (MTAG), White blood cell count, Sunburns

							circumference adjusted for BMI in non-smokers, 3-month functional outcome in ischaemic stroke (modified Rankin score), Cortical surface area, Depressive symptoms, Positive affect, Offspring birth weight, Lung function (FEV1/FVC), Well-being spectrum (multivariate analysis), White matter microstructure (fractional anisotropy), Type 2 diabetes, Life satisfaction, Macular thickness, Sensitivity to environmental stress and adversity, Cognitive performance, Cognitive performance (MTAG), Highest math class taken (MTAG), White blood cell count, Sunburns	
2	101127642	101528728	101328728	1.32E-07	ENSG00000115539, ENSG00000170485	PDCL3, NPAS2	ENSG00000115539:	ENSG00000115539: ; ENSG00000170485: 3-hydroxy-1-methylpropylmercapturic acid levels in smokers, Facial morphology (factor 15, philtrum width), Intraocular pressure, Serum 25-Hydroxyvitamin D levels, Intraocular pressure and central corneal thickness (multi-trait analysis), Pulse pressure, Chronotype, Educational attainment (years of education), Educational attainment (MTAG), Waist-hip ratio, White blood cell count
13	97953799	98367921	98153799	1.33E-07	ENSG00000139793, ENSG00000125249	MBNL2, RAP2A	ENSG00000125249:	ENSG00000139793: Alcoholism (alcohol use disorder factor score), Alcoholism (alcohol dependence factor score), Energy expenditure (24h), Platelet reactivity in response to clopidogrel treatment, Diastolic blood pressure, Pre-treatment viral load in HIV-1 infection, Self-reported math ability, Self-reported math ability (MTAG), Highest math class taken (MTAG), Menarche (age at onset); ENSG00000125249:
19	54178627	54585820	54385437	1.43E-07	ENSG00000142405, ENSG00000179820, ENSG00000126583, ENSG00000105605, ENSG00000142408, ENSG00000130433,	NLRP12, MYADM, PRKCG, CACNG7, CACNG8, CACNG6,	ENSG00000179820:	ENSG00000142405: Granulocyte percentage of myeloid white cells, Blood protein levels, Macrophage Migration Inhibitory Factor levels, Monocyte percentage of white cells, Monocyte count; ENSG00000179820: ; ENSG00000126583: ; ENSG00000105605: General cognitive ability, Educational attainment (MTAG), Cognitive performance (MTAG), Self-reported math ability, Self-reported math ability (MTAG),

					ENSG00000189068, ENSG00000248385	VSTM1, TARM1		Highest math class taken (MTAG); ENSG00000142408: ; ENSG00000130433: ; ENSG00000189068: Blood protein levels; ENSG00000248385:
4	7809737	8218920	8016164	2.53E-07	ENSG00000196526, ENSG00000163995, ENSG00000125089	AFAP1, ABLIM2, SH3TC1	ENSG00000163995: Post bronchodilator FEV1 in COPD, Cognitive function, Triacylglyceride levels, Early onset periodontitis x smoking status interaction, Irritable bowel syndrome, Diffusing capacity of carbon monoxide, Alcohol consumption (drinks per week) (MTAG), Response to cognitive-behavioural therapy in major depressive disorder	ENSG00000196526: Post bronchodilator FEV1/FVC ratio, Intraocular pressure, Pulse pressure, Glaucoma (primary open-angle), Glaucoma, Glaucoma (multi-trait analysis), Alanine aminotransferase levels in excessive alcohol consumption, HDL cholesterol levels x short total sleep time interaction (2df test), Granulocyte-colony stimulating factor levels, Lung function (FEV1/FVC), Diverticular disease, Mean platelet volume, Platelet count, Medication use (antiglaucoma preparations and miotics), Height, Systolic blood pressure; ENSG00000163995: Post bronchodilator FEV1 in COPD, Cognitive function, Triacylglyceride levels, Early onset periodontitis x smoking status interaction, Irritable bowel syndrome, Diffusing capacity of carbon monoxide, Alcohol consumption (drinks per week) (MTAG), Response to cognitive-behavioural therapy in major depressive disorder; ENSG00000125089: Response to metformin (IC50), Idiopathic downbeat nystagmus, Low density lipoprotein cholesterol levels
4	76831352	77231352	77031352	3.04E-07	ENSG00000138744, ENSG00000198301, ENSG00000138755, ENSG00000156219, ENSG00000169245, ENSG00000169248, ENSG00000138750, ENSG00000138760, ENSG00000189157, ENSG00000118804	NAAA, SDAD1, CXCL9, ART3, CXCL10, CXCL11, NUP54, **SCARB2**, FAM47E, FAM47E-STBD1	ENSG00000138750: Deliberate self-harm	ENSG00000138744: Coronary artery calcified atherosclerotic plaque (130 HU threshold) in type 2 diabetes, Blood protein levels, Neurological blood protein biomarker levels; ENSG00000198301: Longevity, Monokine induced by gamma interferon levels, Hippocampal volume, Interferon gamma-induced protein 10 levels; ENSG00000138755: Blood protein levels; ENSG00000156219: Blood protein levels, C-X-C motif chemokine 10 levels, Monokine induced by gamma interferon levels, Hippocampal volume, Neonatal cytokine/chemokine levels (fetal genetic effect), Type 2 diabetes, Body mass index, Mean corpuscular hemoglobin; ENSG00000169245: Blood protein levels, C-X-C motif chemokine 10 levels, Neonatal cytokine/chemokine levels (fetal genetic effect); ENSG00000169248: Blood protein levels; ENSG00000138750: Deliberate self-harm; ENSG00000138760: Body mass index, Parkinson's disease or first degree relation to individual with Parkinson's disease, Body mass index (joint analysis main effects and smoking interaction), BMI (adjusted for smoking behaviour); ENSG00000189157: Parkinson's disease,

								Parkinson's disease or first degree relation to individual with Parkinson's disease, HDL cholesterol levels x long total sleep time interaction (2df test), Platelet count, Mean corpuscular hemoglobin, Mean corpuscular volume, Blood protein levels; ENSG00000118804:
3	167638903	168093774	167852482	4.62E-07	ENSG00000173905	GOLIM4	ENSG00000173905: Response to metformin (IC50), Metabolite levels, Cerebrospinal fluid t-tau levels in Alzheimer's disease dementia, Self-reported math ability, Height, Self-reported math ability (MTAG)	ENSG00000173905: Response to metformin (IC50), Metabolite levels, Cerebrospinal fluid t-tau levels in Alzheimer's disease dementia, Self-reported math ability, Height, Self-reported math ability (MTAG)
5	85631480	86067680	85867680	4.80E-07	ENSG00000127184	COX7C	ENSG00000127184:	ENSG00000127184:
15	28144238	28735266	28344238	5.46E-07	ENSG00000104044, ENSG00000128731, ENSG00000153684	OCA2, HERC2, GOLGA8F	ENSG00000128731: Black vs. blond hair color, Black vs. red hair color, Eye color, Vitiligo, Hair color, Tanning, Eye color traits, Blond vs. brown hair color, Blue vs. green eyes, Blue vs. brown eyes, Iris color, Multiple myeloma (IgH translocation), Alzheimer disease and age of onset, Squamous cell carcinoma, Skin colour saturation, Perceived skin darkness, Skin sensitivity to sun, Blond vs non-blond hair color, Brown vs. non-brown hair color, Light vs. dark hair color, Basal cell carcinoma, Osteoarthritis of the hip (with total joint replacement), Intraocular pressure, Skin pigmentation (conditioned on rs1426654 and rs35397), Skin pigmentation, Cutaneous melanoma or hair colour, Cutaneous malignant melanoma, Nevus count or cutaneous melanoma, Diisocyanate-induced asthma, Refractive astigmatism, Corneal	ENSG00000104044: Black vs. blond hair color, Black vs. red hair color, Eye color, Lung function (forced expiratory flow during mid-portion (25% and 75%) of forced vital capacity), Squamous cell carcinoma, Post bronchodilator FEV1/FVC ratio, Blond vs non-blond hair color, Brown vs. non-brown hair color, Light vs. dark hair color, Red vs non-red hair color, Facial morphology (factor 19), Uveal melanoma, Iris color (b* coordinate), Iris color (L* coordinate), Iris color (a* coordinate), Skin pigmentation (conditioned on rs1426654 and rs35397), Cataracts (operation), Cutaneous malignant melanoma, Central retinal vein equivalent, Central retinal arteriolar equivalent, Low tan response, Eye color (brightness), Eye color (hue), Skin, hair and eye pigmentation (multivariate analysis), Blond vs. brown/black hair color, Brown vs. black hair color, Cutaneous squamous cell carcinoma, Skin pigmentation, Eye color (saturation), Melanoma, Shingles, Eye color traits, Skin pigmentation traits, Macular thickness, Hair color, Sunburns; ENSG00000128731: Black vs. blond hair color, Black vs. red hair color, Eye color, Vitiligo, Hair color, Tanning, Eye color traits, Blond vs. brown hair color, Blue vs. green eyes, Blue vs. brown eyes, Iris color, Multiple myeloma (IgH translocation), Alzheimer disease and age of onset, Squamous cell carcinoma, Skin colour saturation, Perceived skin darkness, Skin sensitivity to sun, Blond vs non-blond hair color, Brown vs. non-brown hair color, Light vs. dark hair color, Basal cell carcinoma, Osteoarthritis of the hip (with total joint replacement), Intraocular pressure, Skin pigmentation

							astigmatism, Bone mineral content, Low tan response, Rosacea symptom severity, Eye color (hue), Skin, hair and eye pigmentation (multivariate analysis), Brown vs. black hair color, Red vs. brown/black hair color, Blond vs. brown/black hair color, Eye color (saturation), Eye color (brightness), Monobrow, Colorectal cancer, Colonoscopy-negative controls vs population controls, Keratinocyte cancer (MTAG), Glaucoma, Iris color (b* coordinate), Iris heterochromicity, Iris color (L* coordinate), Type 2 diabetes, Skin pigmentation traits, Hair morphology traits, Sunburns	(conditioned on rs1426654 and rs35397), Skin pigmentation, Cutaneous melanoma or hair colour, Cutaneous malignant melanoma, Nevus count or cutaneous melanoma, Diisocyanate-induced asthma, Refractive astigmatism, Corneal astigmatism, Bone mineral content, Low tan response, Rosacea symptom severity, Eye color (hue), Skin, hair and eye pigmentation (multivariate analysis), Brown vs. black hair color, Red vs. brown/black hair color, Blond vs. brown/black hair color, Eye color (saturation), Eye color (brightness), Monobrow, Colorectal cancer, Colonoscopy-negative controls vs population controls, Keratinocyte cancer (MTAG), Glaucoma, Iris color (b* coordinate), Iris heterochromicity, Iris color (L* coordinate), Type 2 diabetes, Skin pigmentation traits, Hair morphology traits, Sunburns; ENSG00000153684:

Table S4d.2. Top 25 regions from west-eurasia scan.

chr	start	end	bestpos	minpvalue	ensembl	hgnc	disease_trait_best	disease_trait
15	48233494	48633494	48433494	2.02E-31	ENSG00000188467, ENSG00000104177, ENSG00000233932, ENSG00000074803, ENSG00000128951	SLC24A5, MYEF2, CTXN2, SLC12A1, DUT	ENSG00000188467: Hair color, Skin pigmentation, Body mass index, Skin, hair and eye pigmentation (multivariate analysis), Eye color, Eye color (brightness), Eye color (saturation), Iris color (a* coordinate), Skin reflectance (Melanin index), Iris color (b* coordinate)	ENSG00000188467: Hair color, Skin pigmentation, Body mass index, Skin, hair and eye pigmentation (multivariate analysis), Eye color, Eye color (brightness), Eye color (saturation), Iris color (a* coordinate), Skin reflectance (Melanin index), Iris color (b* coordinate); ENSG00000104177: Skin pigmentation, Hair color; ENSG00000233932: ; ENSG00000074803: Systemic lupus erythematosus, Longevity, Skin reflectance (Melanin index), Lymphocyte counts; ENSG00000128951: Protein quantitative trait loci
6	134192815	134829070	134392815	4.09E-26	ENSG00000118526, ENSG00000028839, ENSG00000146411, ENSG00000118515	TCF21, TBPL1, SLC2A12, SGK1	ENSG00000146411: Coronary artery disease, High chromosomal aberration frequency (chromosome type), FEV1, Lung function (FVC)	ENSG00000118526: Coronary heart disease, Coronary artery disease or ischemic stroke, Coronary artery disease, Coronary artery disease or large artery stroke, PR interval, Medication use (diuretics), Lung function (FEV1/FVC); ENSG00000028839: Mean corpuscular hemoglobin; ENSG00000146411: Coronary artery disease, High chromosomal aberration frequency (chromosome type), FEV1, Lung function (FVC);

								ENSG00000118515: Immune reponse to smallpox (secreted IFN-alpha), Alzheimer disease and age of onset, Pelvic organ prolapse, Pelvic organ prolapse (moderate/severe), Blond vs. brown/black hair color, Schizophrenia (inflammation and infection response interaction), Body mass index (smoking years interaction), Metabolite levels, Adolescent idiopathic scoliosis, Hair color
11	61264124	62263869	61770303	8.70E-22	ENSG00000204950, ENSG00000011347, ENSG00000134780, ENSG00000124920, ENSG00000134825, ENSG00000168496, ENSG00000134824, ENSG00000149485, ENSG00000221968, ENSG00000167994, ENSG00000167995, ENSG00000167996, ENSG00000149503, ENSG00000168515, ENSG00000124939, ENSG00000124935, ENSG00000110484, ENSG00000197745, ENSG00000162174, ENSG00000149021, ENSG00000124942	LRRC10B, SYT7, DAGLA, MYRF, TMEM258, FEN1, FADS2, FADS1, FADS3, RAB3IL1, BEST1, FTH1, INCENP, SCGB1D1, SCGB2A1, SCGB1D2, SCGB2A2, SCGB1D4, ASRGL1, SCGB1A1, AHNAK	ENSG00000167996:	ENSG00000204950: Diastolic blood pressure, Systolic blood pressure, Systolic blood pressure (cigarette smoking interaction), Diastolic blood pressure (cigarette smoking interaction), Mean arterial pressure, Mean arterial pressure x alcohol consumption interaction (2df test), Diastolic blood pressure x alcohol consumption interaction (2df test), Hypertension, Medication use (agents acting on the renin-angiotensin system), Heel bone mineral density; ENSG00000011347: Intelligence (MTAG), Phosphatidylcholine levels, Cholesteryl ester levels, Educational attainment (years of education), Cognitive performance (MTAG), Cognitive performance; ENSG00000134780: Immune reponse to smallpox (secreted IL-2), Plasma omega-3 polyunsaturated fatty acid levels (docosapentaenoic acid), Plasma omega-6 polyunsaturated fatty acid levels (linoleic acid), 3-hydroxypropylmercapturic acid levels in smokers, Refractive error, Cerebrospinal fluid sTREM-2 levels, Spherical equivalent, C-reactive protein levels, Neuroticism, Positive affect, Depressive symptoms, Well-being spectrum (multivariate analysis), Life satisfaction, Depression, Lung function (FVC); ENSG00000124920: Plasma omega-3 polyunsaturated fatty acid levels (docosapentaenoic acid), Plasma omega-3 polyunsaturated fatty acid levels (alphalinolenic acid), Plasma omega-3 polyunsaturated fatty acid level (eicosapentaenoic acid), Metabolite levels, Crohn's disease, Plasma omega-6 polyunsaturated fatty acid levels (gamma-linolenic acid), Plasma omega-6 polyunsaturated fatty acid levels (dihomo-gamma-linolenic acid), Plasma omega-6 polyunsaturated fatty acid levels (arachidonic acid), Hematocrit, Triglycerides, HDL cholesterol, Glycerophospholipid levels, Total cholesterol levels, Moyamoya disease, Resting heart rate, Red blood cell count, Serum total protein level, Serum metabolite concentrations in chronic kidney disease, Serum metabolite ratios in chronic kidney disease, Triglyceride levels x long total sleep time interaction (2df test), Serum metabolite levels (CMS), Heel bone mineral density, Hemoglobin levels, Spherical equivalent, Blood metabolite levels, Colorectal cancer, Serum omega-3 polyunsaturated fatty acid concentration in metabolic syndrome, Low density lipoprotein cholesterol levels, Serum metabolite levels, Chronic inflammatory diseases (ankylosing spondylitis, Crohn's disease, psoriasis, primary sclerosing

								cholangitis, ulcerative colitis) (pleiotropy), Hemoglobin concentration, Vaccenic acid (18:1n-7) levels, Gondoic acid (20:1n-9) levels, LDL cholesterol levels, Iron status biomarkers (total iron binding capacity), Trans fatty acid levels, High density lipoprotein cholesterol levels, Red blood cell fatty acid levels, Colorectal cancer or advanced adenoma, Phosphatidylcholine-ether levels, Phosphatidylethanolamine-ether levels, Asthma (adult onset), Asthma, Nasal polyps, Systolic blood pressure, Triglyceride levels, Glycemic traits (pleiotropy), Educational attainment (years of education), Respiratory diseases; ENSG00000134825: Crohn's disease, Metabolic syndrome, Palmitoleic acid (16:1n-7) levels, Stearic acid (18:0) levels, Oleic acid (18:1n-9) levels, Phospholipid levels (plasma), Plasma omega-3 polyunsaturated fatty acid levels (docosapentaenoic acid), Plasma omega-3 polyunsaturated fatty acid level (eicosapentaenoic acid), Plasma omega-3 polyunsaturated fatty acid levels (alphalinolenic acid), Metabolite levels, Plasma omega-6 polyunsaturated fatty acid levels (gamma-linolenic acid), Plasma omega-6 polyunsaturated fatty acid levels (dihomo-gamma-linolenic acid), Plasma omega-6 polyunsaturated fatty acid levels (arachidonic acid), Hematocrit, Triglycerides, HDL cholesterol, Glycerophospholipid levels, Total cholesterol levels, Resting heart rate, Red blood cell count, Serum total protein level, Irritable mood, Serum metabolite concentrations in chronic kidney disease, Serum metabolite ratios in chronic kidney disease, LDL cholesterol levels x short total sleep time interaction (2df test), Triglyceride levels x long total sleep time interaction (2df test), Serum metabolite levels (CMS), Heel bone mineral density, Hemoglobin levels, Spherical equivalent, Blood metabolite levels, Colorectal cancer, Serum omega-3 polyunsaturated fatty acid concentration in metabolic syndrome, Serum metabolite levels, Blond vs. brown/black hair color, Low density lipoprotein cholesterol levels, Bipolar disorder or major depressive disorder, Chronic inflammatory diseases (ankylosing spondylitis, Crohn's disease, psoriasis, primary sclerosing cholangitis, ulcerative colitis) (pleiotropy), Hemoglobin concentration, Vaccenic acid (18:1n-7) levels, Gondoic acid (20:1n-9) levels, LDL cholesterol levels, Iron status biomarkers (total iron binding capacity), Trans fatty acid levels, High density lipoprotein cholesterol levels, Carboplatin disposition in epthelial ovarian cancer, Red blood cell fatty acid levels, Colorectal cancer or advanced adenoma, Phosphatidylcholine-ether levels, Phosphatidylcholine levels, Phosphatidylethanolamine-ether levels, Cholesteryl ester levels, Asthma (adult onset), Asthma, Nasal polyps, Triglyceride levels, Glycemic traits (pleiotropy), Educational attainment (years of education), Respiratory diseases, Hair color; ENSG00000168496: Plasma omega-3 polyunsaturated fatty acid levels

								(docosapentaenoic acid), Plasma omega-3 polyunsaturated fatty acid levels (alphalinolenic acid), Plasma omega-3 polyunsaturated fatty acid level (eicosapentaenoic acid), Platelet count, Inflammatory bowel disease, Crohn's disease, Trans fatty acid levels, Metabolite levels, Red blood cell fatty acid levels, Colorectal cancer; ENSG00000134824: HDL cholesterol, Triglycerides, Fasting blood glucose, Homeostasis model assessment of beta-cell function, Cholesterol, total, LDL cholesterol, Heart rate, Metabolic syndrome, Resting heart rate, Lipid metabolism phenotypes, Hematology traits, Liver enzyme levels (alkaline phosphatase), Comprehensive strength and appendicular lean mass, Metabolite levels, Response to statin therapy, Metabolic traits, Plasma omega-3 polyunsaturated fatty acid levels (docosapentaenoic acid), Plasma omega-3 polyunsaturated fatty acid levels (alphalinolenic acid), Plasma omega-3 polyunsaturated fatty acid level (eicosapentaenoic acid), Platelet count, Fasting blood glucose (BMI interaction), Inflammatory bowel disease, Plasma omega-6 polyunsaturated fatty acid levels (gamma-linolenic acid), Plasma omega-6 polyunsaturated fatty acid levels (dihomo-gamma-linolenic acid), Plasma omega-6 polyunsaturated fatty acid levels (arachidonic acid), Delta-6 desaturase activity, Plasma omega-6 polyunsaturated fatty acid levels (linoleic acid), Eosinophil counts, C-reactive protein levels or LDL-cholesterol levels (pleiotropy), C-reactive protein levels or HDL-cholesterol levels (pleiotropy), C-reactive protein levels or triglyceride levels (pleiotropy), Metabolite levels (lipid measures), Gestational age at birth (child effect), Plateletcrit, Granulocyte percentage of myeloid white cells, Monocyte percentage of white cells, Glycerophospholipid levels, Sphingolipid levels, Vitiligo, Glycated hemoglobin levels, Total cholesterol levels, Heel bone mineral density, Non-albumin protein levels, Albumin-globulin ratio, Hemoglobin A1c levels, Alanine transaminase levels, Triglyceride levels, Breast milk fatty acid composition (maternal genotype effect), Breast milk fatty acid composition (infant genotype effect), Serum metabolite concentrations in chronic kidney disease, Serum metabolite ratios in chronic kidney disease, Low density lipoprotein cholesterol levels, Height, Apolipoprotein B levels, Pulse pressure, LDL cholesterol levels, LDL cholesterol levels x long total sleep time interaction (2df test), HDL cholesterol levels, Triglyceride levels x short total sleep time interaction (2df test), Serum metabolite levels (CMS), HDL cholesterol levels x long total sleep time interaction (2df test), HDL cholesterol levels x short total sleep time interaction (2df test), HDL cholesterol levels x alcohol consumption (regular vs non-regular drinkers) interaction (2df), Triglyceride levels x alcohol consumption (regular vs non-regular drinkers) interaction (2df), Bipolar I disorder, Apolipoprotein

								A1 levels, Neutrophil count, Colorectal cancer, QRS duration, Crohn's disease, Age-related disease endophenotypes, Sum basophil neutrophil counts, Red cell distribution width, P wave duration, Blood metabolite levels, Blood metabolite ratios, Rheumatoid arthritis, QT interval, Iron status biomarkers (transferrin levels), Serum metabolite levels, Change in serum metabolite levels (CMS), Serum omega-3 polyunsaturated fatty acid concentration in metabolic syndrome, Serum omega-6 to omega-3 polyunsaturated fatty acid ratio in metabolic syndrome, Serum docosahexaenoic fatty acid concentration in metabolic syndrome, Delta-5 desaturase activity response to n3-polyunsaturated fat supplement, Change in serum metabolite levels, LDL cholesterol x physical activity interaction (2df test), High density lipoprotein cholesterol levels, Asthma, Bipolar disorder, Male-pattern baldness, Nonatopic asthma, Osteoporosis-related phenotypes (MTAG), Granulocyte count, Age-related diseases, mortality and associated endophenotypes, Sum neutrophil eosinophil counts, Gondoic acid (20:1n-9) levels, Sum eosinophil basophil counts, Myeloid white cell count, White blood cell count, Mean platelet volume, Trans fatty acid levels, Red blood cell fatty acid levels, Plasma omega-6 polyunsaturated fatty acid levels (adrenic acid), Laryngeal squamous cell carcinoma, Lung cancer in ever smokers, Triacylglycerol 56:6 levels, Metabolite risk score for predicting weight gain, Fatty acid desaturase activity (serum), Fatty acid desaturase activity (adipose tissue), Phosphatidylcholine levels, Phosphatidylcholine-ether levels, Phosphatidylethanolamine levels, Lysophosphatidylethanolamine levels, Phosphatidylethanolamine-ether levels, Phosphatidylinositol levels, Cholesteryl ester levels, Sphingomyelin levels, Lysophosphatidylcholine levels, Triacylglyceride levels, HDL cholesterol levels in current drinkers, LDL cholesterol levels in current drinkers, Triglyceride levels in current drinkers, HDL cholesterol levels x alcohol consumption (drinkers vs non-drinkers) interaction (2df), Triglyceride levels x alcohol consumption (drinkers vs non-drinkers) interaction (2df), LDL cholesterol levels x alcohol consumption (drinkers vs non-drinkers) interaction (2df), LDL cholesterol levels x alcohol consumption (regular vs non-regular drinkers) interaction (2df), Red blood cell count, Aortic valve stenosis, Sleep duration, Hematocrit, Hemoglobin concentration, Total triglycerides levels, Lymphocyte counts, Mean corpuscular hemoglobin concentration, PR interval, Type 2 diabetes, Mean corpuscular volume, IgA levels, Adult onset asthma or type 2 diabetes, Nonatopic asthma or type 2 diabetes, Adult onset asthma or fasting glucose levels, Nonatopic asthma or fasting glucose levels, Asthma (adult onset), Medication use (thyroid preparations), Medication use (adrenergics, inhalants), Urate levels,

								Gallstone disease, anorexia nervosa, attention-deficit/hyperactivity disorder, autism spectrum disorder, bipolar disorder, major depression, obsessive-compulsive disorder, schizophrenia, or Tourette syndrome (pleiotropy), QT dynamics during exercise, Balding type 1, Hypothyroidism; ENSG00000149485: HDL cholesterol, Triglycerides, Fasting blood glucose, Homeostasis model assessment of beta-cell function, LDL cholesterol, Heart rate, Metabolic syndrome, Resting heart rate, Lipid metabolism phenotypes, Hematology traits, Comprehensive strength and appendicular lean mass, Metabolite levels, Cholesterol, total, Metabolic traits, Plasma omega-3 polyunsaturated fatty acid levels (alphalinolenic acid), Plasma omega-3 polyunsaturated fatty acid level (eicosapentaenoic acid), Plasma omega-3 polyunsaturated fatty acid levels (docosapentaenoic acid), Fasting blood glucose (BMI interaction), Plasma omega-6 polyunsaturated fatty acid levels (gamma-linolenic acid), Plasma omega-6 polyunsaturated fatty acid levels (dihomo-gamma-linolenic acid), Delta-6 desaturase activity, Plasma omega-6 polyunsaturated fatty acid levels (linoleic acid), Eosinophil counts, Platelet count, C-reactive protein levels or HDL-cholesterol levels (pleiotropy), C-reactive protein levels or triglyceride levels (pleiotropy), Metabolite levels (lipid measures), Granulocyte percentage of myeloid white cells, Monocyte percentage of white cells, Glycerophospholipid levels, Sphingolipid levels, Vitiligo, Total cholesterol levels, Alanine transaminase levels, Triglyceride levels, Breast milk fatty acid composition (maternal genotype effect), Breast milk fatty acid composition (infant genotype effect), Serum metabolite concentrations in chronic kidney disease, Serum metabolite ratios in chronic kidney disease, Low density lipoprotein cholesterol levels, Apolipoprotein B levels, Pulse pressure, LDL cholesterol levels, LDL cholesterol levels x long total sleep time interaction (2df test), HDL cholesterol levels, Triglyceride levels x short total sleep time interaction (2df test), Serum metabolite levels (CMS), HDL cholesterol levels x long total sleep time interaction (2df test), HDL cholesterol levels x short total sleep time interaction (2df test), HDL cholesterol levels x alcohol consumption (regular vs non-regular drinkers) interaction (2df), Triglyceride levels x alcohol consumption (regular vs non-regular drinkers) interaction (2df), Bipolar I disorder, Apolipoprotein A1 levels, Age-related disease endophenotypes, Red cell distribution width, Height, Blood metabolite levels, Blood metabolite ratios, Rheumatoid arthritis, Serum metabolite levels, Change in serum metabolite levels (CMS), Serum omega-3 polyunsaturated fatty acid concentration in metabolic syndrome, Serum omega-6 to omega-3 polyunsaturated fatty acid ratio in metabolic syndrome, Serum docosahexaenoic fatty acid concentration in metabolic

								syndrome, Delta-5 desaturase activity response to n3-polyunsaturated fat supplement, Change in serum metabolite levels, LDL cholesterol x physical activity interaction (2df test), High density lipoprotein cholesterol levels, Asthma, Bipolar disorder, Osteoporosis-related phenotypes (MTAG), Granulocyte count, Age-related diseases, mortality and associated endophenotypes, Sum neutrophil eosinophil counts, Sum eosinophil basophil counts, Myeloid white cell count, White blood cell count, Mean platelet volume, Trans fatty acid levels, Red blood cell fatty acid levels, Plasma omega-6 polyunsaturated fatty acid levels (arachidonic acid), Plasma omega-6 polyunsaturated fatty acid levels (adrenic acid), Laryngeal squamous cell carcinoma, Triacylglycerol 56:6 levels, Metabolite risk score for predicting weight gain, Fatty acid desaturase activity (serum), Fatty acid desaturase activity (adipose tissue), Phosphatidylcholine-ether levels, Lysophosphatidylethanolamine levels, Phosphatidylcholine levels, Phosphatidylethanolamine-ether levels, Phosphatidylinositol levels, Cholesteryl ester levels, Sphingomyelin levels, Lysophosphatidylcholine levels, Triacylglyceride levels, HDL cholesterol levels in current drinkers, LDL cholesterol levels in current drinkers, Triglyceride levels in current drinkers, HDL cholesterol levels x alcohol consumption (drinkers vs non-drinkers) interaction (2df), Triglyceride levels x alcohol consumption (drinkers vs non-drinkers) interaction (2df), LDL cholesterol levels x alcohol consumption (drinkers vs non-drinkers) interaction (2df), LDL cholesterol levels x alcohol consumption (regular vs non-regular drinkers) interaction (2df), Red blood cell count, QT interval, Aortic valve stenosis, Sleep duration, Neutrophil count, Total triglycerides levels, Mean corpuscular hemoglobin concentration, Mean corpuscular volume, IgA levels, Gallstone disease; ENSG00000221968: Sphingolipid levels, Metabolite levels, Phosphatidylcholine levels, Cholesteryl ester levels, Sphingolipid d18:1/d18:2 ratio, Urate levels; ENSG00000167994: Sphingolipid levels, Metabolite levels, Schizophrenia, Mean corpuscular hemoglobin; ENSG00000167995: Plasma omega-3 polyunsaturated fatty acid level (eicosapentaenoic acid), Metabolite levels; ENSG00000167996: ; ENSG00000149503: Prostate cancer, Breast cancer; ENSG00000168515: ; ENSG00000124939: ; ENSG00000124935: ; ENSG00000110484: ; ENSG00000197745: ; ENSG00000162174: 3-hydroxypropylmercapturic acid levels in smokers, Trunk fat mass, Body fat mass; ENSG00000149021: Waist-to-hip ratio adjusted for BMI, Waist-hip ratio; ENSG00000124942: Alzheimer disease and age of onset, Heel bone mineral density, Alzheimer's disease (late onset), HDL cholesterol levels, Cutaneous melanoma or hair colour, Waist-to-hip ratio adjusted for BMI, Blond vs. brown/black hair color, C-reactive protein levels, Lung function (FEV1),

								Waist-hip ratio, Waist circumference adjusted for body mass index, Hair color
17	79055998	79469799	79255998	1.91E-20	ENSG00000175866, ENSG00000181409, ENSG00000141577, ENSG00000167302, ENSG00000224877, ENSG00000157637, ENSG00000185332	BAIAP2, AATK, AZI1, ENTHD2, C17orf89, SLC38A10, TMEM105	ENSG00000157637: Longevity, IgG glycosylation, Serum 25-Hydroxyvitamin D levels, Menarche (age at onset)	ENSG00000175866: Serum uric acid levels in response to allopurinol in gout, Neurociticism, Neuroticism, Feeling tense, Fat-free mass, Memory performance, Depression, Worry, Depressed affect, Body mass index, General factor of neuroticism, Positive affect, Well-being spectrum (multivariate analysis), Breast cancer, Depressive symptoms (MTAG), Depressive symptoms, General cognitive ability, Sensitivity to environmental stress and adversity, Cognitive performance (MTAG), Self-reported math ability, Self-reported math ability (MTAG); ENSG00000181409: Obesity-related traits, Neurociticism, Feeling tense, Beef consumption, Fat-free mass, Worry, Neuroticism, Body mass index, Depressive symptoms, Neutrophil count, Life satisfaction, Sensitivity to environmental stress and adversity, Height, Systolic blood pressure; ENSG00000141577: IgG glycosylation, Frontotemporal dementia, IgG galactosylation phenotypes (multivariate analysis), IgG N-glycosylation phenotypes (multivariate analysis), Red blood cell count; ENSG00000167302: Blood protein levels; ENSG00000224877: IgG fucosylation phenotypes (multivariate analysis); ENSG00000157637: Longevity, IgG glycosylation, Serum 25-Hydroxyvitamin D levels, Menarche (age at onset); ENSG00000185332: Intake of total sugars
11	60562850	61159993	60762850	4.62E-19	ENSG00000172689, ENSG00000110104, ENSG00000183134, ENSG00000149506, ENSG00000110107, ENSG00000110108, ENSG00000006118, ENSG00000110446, ENSG00000013725, ENSG00000110448, ENSG00000167987, ENSG00000229859, ENSG00000229183, ENSG00000256713, ENSG00000167992, ENSG00000167986, ENSG00000149476, ENSG00000162144,	MS4A10, CCDC86, PTGDR2, ZP1, PRPF19, TMEM109, TMEM132A, SLC15A3, CD6, CD5, VPS37C, PGA3, PGA4, PGA5, VWCE, DDB1, DAK, CYB561A3, TMEM138, TMEM216	ENSG00000013725: Multiple sclerosis, Inflammatory bowel disease, Crohn's disease, Ulcerative colitis, Chronic inflammatory diseases (ankylosing spondylitis, Crohn's disease, psoriasis, primary sclerosing cholangitis, ulcerative colitis) (pleiotropy), CD6 levels	ENSG00000172689: ; ENSG00000110104: ; ENSG00000183134: ; ENSG00000149506: Response to ziprazidone in schizophrenia; ENSG00000110107: ; ENSG00000110108: Blood protein levels; ENSG00000006118: Blood protein levels; ENSG00000110446: ; ENSG00000013725: Multiple sclerosis, Inflammatory bowel disease, Crohn's disease, Ulcerative colitis, Chronic inflammatory diseases (ankylosing spondylitis, Crohn's disease, psoriasis, primary sclerosing cholangitis, ulcerative colitis) (pleiotropy), CD6 levels; ENSG00000110448: ; ENSG00000167987: Rheumatoid arthritis (ACPA-positive), Rheumatoid arthritis, Adolescent idiopathic scoliosis, Periventricular white matter hyperintensities; ENSG00000229859: ; ENSG00000229183: ; ENSG00000256713: Height; ENSG00000167992: ; ENSG00000167986: ; ENSG00000149476: ; ENSG00000162144: ; ENSG00000149483: ; ENSG00000187049:

					ENSG00000149483, ENSG00000187049			
1	147576333	148018267	147805232	6.54E-19	ENSG00000203836, ENSG00000255963, ENSG00000122497	NBPF24, PPIAL4A, NBPF14	ENSG00000255963:	ENSG00000203836: ; ENSG00000255963: ; ENSG00000122497:
4	100054570	100521443	100254570	3.90E-18	ENSG00000198099, ENSG00000172955, ENSG00000187758, ENSG00000196616, ENSG00000196344, ENSG00000138813, ENSG00000145331, ENSG00000138823	ADH4, ADH6, ADH1A, ADH1B, ADH7, C4orf17, TRMT10A, MTTP	ENSG00000196616: Conduct disorder (maternal expressed emotions interaction), Esophageal cancer (alcohol interaction), Oral cavity and pharyngeal cancer, Alcohol dependence, Oropharynx cancer, Oral cavity cancer, Cerebrospinal fluid clusterin levels, Body mass index, Low density lipoprotein cholesterol levels, Risk-taking tendency (4-domain principal component model), LDL cholesterol levels, Relative fat intake, Relative protein intake, Apolipoprotein B levels, HDL cholesterol levels, Esophageal cancer, Alcohol consumption, Serum 25-Hydroxyvitamin D levels, C-reactive protein levels, Heel bone mineral density, Alcohol consumption (drinks per week), Alcohol consumption (drinkers vs non-drinkers), Alcohol consumption (heavy vs. light/non-drinkers), Pulse pressure, Alcohol consumption over the past year, Major depression and alcohol dependence, Lung cancer, Alcohol consumption in current drinkers, Blood urea nitrogen levels, Waist-hip ratio, Maximum habitual alcohol consumption, Regular attendance at a pub or social club, Urinary sodium excretion, Predicted visceral adipose tissue, Total cholesterol levels, Alcohol use disorder, Alcohol use	ENSG00000198099: Esophageal cancer (alcohol interaction), Alcohol dependence, Triglyceride levels, Serum metabolite levels, Blood protein levels, Alcohol use disorder, Platelet count, Eosinophil counts; ENSG00000172955: QRS duration; ENSG00000187758: Urinary metabolite levels in chronic kidney disease, Serum 25-Hydroxyvitamin D levels; ENSG00000196616: Conduct disorder (maternal expressed emotions interaction), Esophageal cancer (alcohol interaction), Oral cavity and pharyngeal cancer, Alcohol dependence, Oropharynx cancer, Oral cavity cancer, Cerebrospinal fluid clusterin levels, Body mass index, Low density lipoprotein cholesterol levels, Risk-taking tendency (4-domain principal component model), LDL cholesterol levels, Relative fat intake, Relative protein intake, Apolipoprotein B levels, HDL cholesterol levels, Esophageal cancer, Alcohol consumption, Serum 25-Hydroxyvitamin D levels, C-reactive protein levels, Heel bone mineral density, Alcohol consumption (drinks per week), Alcohol consumption (drinkers vs non-drinkers), Alcohol consumption (heavy vs. light/non-drinkers), Pulse pressure, Alcohol consumption over the past year, Major depression and alcohol dependence, Lung cancer, Alcohol consumption in current drinkers, Blood urea nitrogen levels, Waist-hip ratio, Maximum habitual alcohol consumption, Regular attendance at a pub or social club, Urinary sodium excretion, Predicted visceral adipose tissue, Total cholesterol levels, Alcohol use disorder, Alcohol use disorder (consumption score), Hemoglobin concentration, Problematic alcohol use, Alcohol consumption (drinks per week) (MTAG), Problematic alcohol use (MTAG), Alcohol use disorder (total score), Alcohol use disorder (dependence and problematic use scores), Bitter alcoholic beverage consumption, Alcohol dependence symptom count, Alcohol dependence (tolerance), Alcohol dependence (desire to cut drinking), Cardiovascular disease, Red blood cell count, Mean corpuscular hemoglobin, Systolic blood pressure; ENSG00000196344: Oral cavity and pharyngeal cancer, Blood protein levels, Maximum habitual alcohol consumption; ENSG00000138813: Metabolite levels (MHPG), Alcohol consumption (drinks per week), Pre-treatment viral load in HIV-1 infection, Height; ENSG00000145331: Gut microbiome composition (summer); ENSG00000138823: Celiac disease,

							disorder (consumption score), Hemoglobin concentration, Problematic alcohol use, Alcohol consumption (drinks per week) (MTAG), Problematic alcohol use (MTAG), Alcohol use disorder (total score), Alcohol use disorder (dependence and problematic use scores), Bitter alcoholic beverage consumption, Alcohol dependence symptom count, Alcohol dependence (tolerance), Alcohol dependence (desire to cut drinking), Cardiovascular disease, Red blood cell count, Mean corpuscular hemoglobin, Systolic blood pressure	Triglyceride levels, Maximum habitual alcohol consumption, HDL cholesterol, Alcohol use disorder, Lung function (FVC)
2	226448377	226848377	226648377	1.06E-17	ENSG00000144460	NYAP2	ENSG00000144460: Neurociticism, Docetaxel-induced peripheral neuropathy in metastatic castrate-resistant prostate cancer, HDL cholesterol levels, Neuroticism, Smoking initiation (ever regular vs never regular), Age of smoking initiation (MTAG), Positive affect, Well-being spectrum (multivariate analysis), Depressive symptoms, Gut microbiota (bacterial taxa, hurdle binary method), Smoking cessation (MTAG), Life satisfaction, Smoking initiation (ever regular vs never regular) (MTAG), Educational attainment (MTAG), Cognitive performance (MTAG), Educational attainment (years of education), Highest math class taken (MTAG), Smoking status	ENSG00000144460: Neurociticism, Docetaxel-induced peripheral neuropathy in metastatic castrate-resistant prostate cancer, HDL cholesterol levels, Neuroticism, Smoking initiation (ever regular vs never regular), Age of smoking initiation (MTAG), Positive affect, Well-being spectrum (multivariate analysis), Depressive symptoms, Gut microbiota (bacterial taxa, hurdle binary method), Smoking cessation (MTAG), Life satisfaction, Smoking initiation (ever regular vs never regular) (MTAG), Educational attainment (MTAG), Cognitive performance (MTAG), Educational attainment (years of education), Highest math class taken (MTAG), Smoking status
14	73102086	73504565	73302086	2.25E-17	ENSG00000205683, ENSG00000119599, ENSG00000165861	DPF3, DCAF4, ZFYVE1	ENSG00000205683: 3-hydroxy-1-methylpropylmercapturic acid levels in smokers, Disease progression in age-related macular degeneration,	ENSG00000205683: 3-hydroxy-1-methylpropylmercapturic acid levels in smokers, Disease progression in age-related macular degeneration, Pulse pressure, Metabolite levels, Rosacea symptom severity, Adolescent idiopathic scoliosis, C-reactive protein levels, Stem cell factor levels, Renal

							Pulse pressure, Metabolite levels, Rosacea symptom severity, Adolescent idiopathic scoliosis, C-reactive protein levels, Stem cell factor levels, Renal cell carcinoma, Intake of sweets, Body mass index, Atrial fibrillation, Hematocrit, Central corneal thickness, Hemoglobin concentration, Tuberculosis, Systemic lupus erythematosus, Intraocular pressure, Red cell distribution width	cell carcinoma, Intake of sweets, Body mass index, Atrial fibrillation, Hematocrit, Central corneal thickness, Hemoglobin concentration, Tuberculosis, Systemic lupus erythematosus, Intraocular pressure, Red cell distribution width; ENSG00000119599: Leukocyte telomere length, Eosinophil counts, Body mass index, Systolic blood pressure; ENSG00000165861: LDL cholesterol levels, Resistance to Mycobacterium tuberculosis in HIV-positive individuals measured by a negative tuberculin skin test (continuous), Red blood cell count, Body mass index, Intelligence, Mean corpuscular hemoglobin, Red cell distribution width
11	27418490	27933143	27632440	5.74E-17	ENSG00000205213, ENSG00000148943, ENSG00000176697	LGR4, LIN7C, BDNF	ENSG00000176697: Obesity, Body mass index, Smoking behavior, Weight, Childhood body mass index, Menarche (age at onset), Menopause (age at onset), Coronary artery disease, Feeling nervous, Smoking initiation, Snoring, Risk-taking tendency (4-domain principal component model), General risk tolerance (MTAG), Triglyceride levels, Smoking status (ever vs never smokers), Body fat percentage, Fat-free mass, Diastolic blood pressure, Body mass index (joint analysis main effects and physical activity interaction), Coffee consumption, Worry, C-reactive protein levels, Systolic blood pressure, Hip circumference, Body mass index in physically active individuals, Body mass index in physically inactive individuals, Body mass index (SNP x SNP interaction), BMI (adjusted for smoking behaviour), BMI in non-smokers, BMI in smokers, Body mass index (joint analysis main effects and smoking interaction), Waist-hip ratio, Smoking initiation (ever regular vs	ENSG00000205213: Male-pattern baldness, Heel bone mineral density, Spontaneous preterm birth with premature rupture of membranes, Urate levels in obese individuals, Blond vs. brown/black hair color, Waist-hip ratio, Body mass index, Balding type 1, Lung function (FVC), Hair color; ENSG00000148943: Feeling miserable, Hip circumference, Survival in pancreatic cancer, Heel bone mineral density, Body mass index; ENSG00000176697: Obesity, Body mass index, Smoking behavior, Weight, Childhood body mass index, Menarche (age at onset), Menopause (age at onset), Coronary artery disease, Feeling nervous, Smoking initiation, Snoring, Risk-taking tendency (4-domain principal component model), General risk tolerance (MTAG), Triglyceride levels, Smoking status (ever vs never smokers), Body fat percentage, Fat-free mass, Diastolic blood pressure, Body mass index (joint analysis main effects and physical activity interaction), Coffee consumption, Worry, C-reactive protein levels, Systolic blood pressure, Hip circumference, Body mass index in physically active individuals, Body mass index in physically inactive individuals, Body mass index (SNP x SNP interaction), BMI (adjusted for smoking behaviour), BMI in non-smokers, BMI in smokers, Body mass index (joint analysis main effects and smoking interaction), Waist-hip ratio, Smoking initiation (ever regular vs never regular), Metabolic syndrome, Predicted visceral adipose tissue, Basal metabolic rate variance, Body mass index variance, Basal metabolic rate, Hand grip strength, Type 2 diabetes, Problematic alcohol use (MTAG), Diverticular disease, Alcohol consumption, Smoking cessation (MTAG), Age of smoking initiation (MTAG), Cigarettes smoked per day (MTAG), Smoking initiation (ever regular vs never regular) (MTAG), Educational attainment (years of education), Heel bone mineral density, Educational attainment (MTAG), Highest math class taken, Highest math class taken (MTAG), Smoking status

							never regular), Metabolic syndrome, Predicted visceral adipose tissue, Basal metabolic rate variance, Body mass index variance, Basal metabolic rate, Hand grip strength, Type 2 diabetes, Problematic alcohol use (MTAG), Diverticular disease, Alcohol consumption, Smoking cessation (MTAG), Age of smoking initiation (MTAG), Cigarettes smoked per day (MTAG), Smoking initiation (ever regular vs never regular) (MTAG), Educational attainment (years of education), Heel bone mineral density, Educational attainment (MTAG), Highest math class taken, Highest math class taken (MTAG), Smoking status	
1	110586660	111157709	110793599	8.40E-17	ENSG00000143093, ENSG00000156150, ENSG00000186150, ENSG00000197106, ENSG00000116396, ENSG00000162775, ENSG00000168679, ENSG00000134248, ENSG00000143125, ENSG00000143105, ENSG00000177301	STRIP1, ALX3, UBL4B, SLC6A17, KCNC4, RBM15, SLC16A4, LAMTOR5, PROK1, KCNA10, KCNA2	ENSG00000116396: Self-rated health, Oropharynx cancer, Gut microbiota (functional units), Red blood cell count, Educational attainment (MTAG), Cognitive performance (MTAG), Self-reported math ability, Self-reported math ability (MTAG), Educational attainment (years of education), Highest math class taken (MTAG)	ENSG00000143093: Optic disc size, Cognitive performance (MTAG), Educational attainment (years of education); ENSG00000156150: Optic disc size; ENSG00000186150: ; ENSG00000197106: Keratinocyte cancer (MTAG), Basal cell carcinoma, Educational attainment (years of education), Sunburns; ENSG00000116396: Self-rated health, Oropharynx cancer, Gut microbiota (functional units), Red blood cell count, Educational attainment (MTAG), Cognitive performance (MTAG), Self-reported math ability, Self-reported math ability (MTAG), Educational attainment (years of education), Highest math class taken (MTAG); ENSG00000162775: ; ENSG00000168679: Adolescent idiopathic scoliosis; ENSG00000134248: ; ENSG00000143125: ; ENSG00000143105: Bitter taste perception (phenylthiocarbamide) in obesity with metabolic syndrome; ENSG00000177301: Nonsyndromic cleft lip with cleft palate
3	98038565	98438565	98238565	8.59E-17	ENSG00000196098, ENSG00000206536, ENSG00000232382, ENSG00000231861, ENSG00000080822, ENSG00000080819, ENSG00000154165	OR5K4, OR5K3, OR5K1, OR5K2, CLDND1, CPOX, GPR15	ENSG00000080822: Post bronchodilator FEV1/FVC ratio	ENSG00000196098: ; ENSG00000206536: ; ENSG00000232382: ; ENSG00000231861: ; ENSG00000080822: Post bronchodilator FEV1/FVC ratio; ENSG00000080819: ; ENSG00000154165:
17	8655348	9223981	8875399	9.12E-17	ENSG00000183318, ENSG00000185156,	SPDYE4, MFSD6L,	ENSG00000141506: 3-hydroxypropylmercapturic acid	ENSG00000183318: ; ENSG00000185156: ; ENSG00000174083: ; ENSG00000141506: 3-hydroxypropylmercapturic acid levels in smokers,

					ENSG00000174083, ENSG00000141506, ENSG00000065320, ENSG00000170310	PIK3R6, PIK3R5, NTN1, STX8	levels in smokers, Metabolite levels, Eosinophil counts, Autoimmune thyroid disease, Atypical femoral fracture in phosphonate treatment, Hypothyroidism	Metabolite levels, Eosinophil counts, Autoimmune thyroid disease, Atypical femoral fracture in phosphonate treatment, Hypothyroidism; ENSG00000065320: Orofacial clefts, Breast cancer (prognosis), Neuroticism, Percentage gas trapping, 3-hydroxypropylmercapturic acid levels in smokers, Lobe attachment (rater-scored or self-reported), Heel bone mineral density, Feeling hurt, Blood protein levels, Nonsyndromic cleft lip with or without cleft palate, Cleft lip with or without cleft palate, Colonoscopy-negative controls vs population controls, Rostral middle frontal gyrus volume, Total grey matter volume, Cortex volume, White matter microstructure (fractional anisotropy), Bisphosphonate-associated atypical femoral fracture, Sensitivity to environmental stress and adversity; ENSG00000170310: antipsychotic drug dosage in schizophrenia or schizoaffective disorder, HDL cholesterol change in response to fenofibrate in statin-treated type 2 diabetes, Diabetic kidney disease, Obstructive sleep apnea trait (apnea hypopnea index), Borderline personality disorder, White matter microstructure (mode of anisotropy), Gut microbiota (bacterial taxa, hurdle binary method), Blood protein levels, Height
8	133319098	133719098	133519098	3.22E-16	ENSG00000184156, ENSG00000129295, ENSG00000165071	KCNQ3, LRRC6, TMEM71	ENSG00000129295: Adolescent idiopathic scoliosis, Cognitive performance (MTAG)	ENSG00000184156: Coronary artery calcification, LDL peak particle diameter (total fat intake interaction), QT interval, Interleukin-8 levels, Spontaneous adipocyte lipolysis, Height, Stimulated adipocyte lipolysis, Major depressive disorder, Rate of cognitive decline in Alzheimer's disease, Type 2 diabetes; ENSG00000129295: Adolescent idiopathic scoliosis, Cognitive performance (MTAG); ENSG00000165071: Thyroid stimulating hormone levels, Hyperthyroidism, Smoking initiation (ever regular vs never regular) (MTAG), Age of smoking initiation (MTAG), Educational attainment (MTAG), Educational attainment (years of education), Height
3	100737908	101365506	101082807	6.97E-16	ENSG00000081148, ENSG00000138468, ENSG00000174173, ENSG00000081154	IMPG2, SENP7, TRMT10C, PCNP	ENSG00000138468: Risk-taking tendency (4-domain principal component model), Mosaic loss of chromosome Y (Y chromosome dosage), Appendicular lean mass, Mean corpuscular hemoglobin, Alcohol consumption (drinks per week), Diastolic blood pressure, Red blood cell count, Platelet count, Alzheimer's disease (late onset), Bitter alcoholic beverage	ENSG00000081148: Inflammatory bowel disease, Crohn's disease, Mean corpuscular hemoglobin; ENSG00000138468: Risk-taking tendency (4-domain principal component model), Mosaic loss of chromosome Y (Y chromosome dosage), Appendicular lean mass, Mean corpuscular hemoglobin, Alcohol consumption (drinks per week), Diastolic blood pressure, Red blood cell count, Platelet count, Alzheimer's disease (late onset), Bitter alcoholic beverage consumption, Mean corpuscular volume; ENSG00000174173: Eosinophil counts; ENSG00000081154:

							consumption, Mean corpuscular volume	
18	46132358	46550865	46350865	7.05E-16	ENSG00000134030, ENSG00000101665	CTIF, SMAD7	ENSG00000101665: Colorectal cancer, Hematocrit, Creatinine levels, Glomerular filtration rate, Normal facial asymmetry (angle of surface orientation score), Heel bone mineral density, Red blood cell count, Hemoglobin concentration, Hemoglobin levels, Parental longevity (at least one long-lived parent), Colorectal cancer or advanced adenoma, Estimated glomerular filtration rate, Atrial fibrillation, Eosinophil counts, Asthma, Mean corpuscular hemoglobin, Mean corpuscular volume, Red cell distribution width, Eczema	ENSG00000134030: IgG glycosylation, Lung function (FVC), Response to carboplatin in ovarian cancer (MTT IC50), Male-pattern baldness, Red blood cell count, Adolescent idiopathic scoliosis, Automobile speeding propensity, Circulating odd-numbered chain saturated fatty acid levels (C15:0), Migraine with aura, White blood cell count, Iris heterochromicity, Lumbar spine bone mineral density (integral), Rate of cognitive decline in Alzheimer's disease, Height, Smoking status; ENSG00000101665: Colorectal cancer, Hematocrit, Creatinine levels, Glomerular filtration rate, Normal facial asymmetry (angle of surface orientation score), Heel bone mineral density, Red blood cell count, Hemoglobin concentration, Hemoglobin levels, Parental longevity (at least one long-lived parent), Colorectal cancer or advanced adenoma, Estimated glomerular filtration rate, Atrial fibrillation, Eosinophil counts, Asthma, Mean corpuscular hemoglobin, Mean corpuscular volume, Red cell distribution width, Eczema
22	31356103	31756103	31556103	1.03E-15	ENSG00000133422, ENSG00000183963, ENSG00000185133, ENSG00000100078, ENSG00000138942, ENSG00000182541, ENSG00000100100, ENSG00000100105	MORC2, SMTN, INPP5J, PLA2G3, RNF185, LIMK2, PIK3IP1, PATZ1	ENSG00000100078: Serum 25-Hydroxyvitamin D levels	ENSG00000133422: ; ENSG00000183963: Lung function (FEV1/FVC); ENSG00000185133: Gut microbiota (beta diversity); ENSG00000100078: Serum 25-Hydroxyvitamin D levels; ENSG00000138942: Apolipoprotein A1 levels, Waist circumference, Waist-to-hip ratio adjusted for BMI, Neutrophil count, Waist circumference adjusted for body mass index, Waist-hip ratio; ENSG00000182541: Type 2 diabetes nephropathy, Eosinophil percentage of white cells, Eosinophil counts, Paclitaxel-induced neuropathy, Multiple sclerosis, Sum eosinophil basophil counts, Eosinophil percentage of granulocytes, Neutrophil percentage of granulocytes, Schizophrenia, Metabolite levels, Reaction time; ENSG00000100100: ; ENSG00000100105: Height, Reticulocyte fraction of red cells, Reticulocyte count, Red cell distribution width, Eosinophil counts
3	101491814	101891814	101691814	1.40E-15	ENSG00000144815, ENSG00000144802, ENSG00000170044	NXPE3, NFKBIZ, ZPLD1	ENSG00000144802: Ulcerative colitis, Colorectal cancer, Spatial memory	ENSG00000144815: ; ENSG00000144802: Ulcerative colitis, Colorectal cancer, Spatial memory; ENSG00000170044: Post bronchodilator FEV1/FVC ratio, Adolescent idiopathic scoliosis, C-reactive protein levels, Waist-to-hip circumference ratio (alcohol intake interaction), Response to ranibizumab in age-related macular degeneration (exudative)
12	27574693	27974693	27774693	1.53E-15	ENSG00000029153, ENSG00000165935, ENSG00000110841, ENSG00000174236,	ARNTL2, SMCO2, PPFIBP1, REP15,	ENSG00000110841: Cerebral cortical growth, Tinnitus in cisplatin-treated testicular cancer	ENSG00000029153: Body fat percentage, Waist-to-hip ratio adjusted for BMI, Waist-hip ratio; ENSG00000165935: Obstructive sleep apnea trait (apnea hypopnea index), Lung cancer in ever smokers; ENSG00000110841: Cerebral cortical growth, Tinnitus in cisplatin-treated

					ENSG00000061794, ENSG00000205693, ENSG00000087448	MRPS35, MANSC4, KLHL42		testicular cancer; ENSG00000174236: ; ENSG00000061794: Type 2 diabetes, Type 2 diabetes (adjusted for BMI), Lung function (FEV1/FVC); ENSG00000205693: Blood protein levels; ENSG00000087448: Pulse pressure, Resting heart rate

Table S4d.3 West Eurasia.cw_hg

References

1. Luu, K., Bazin, E. & Blum, M. G. B. pcadapt: an R package to perform genome scans for selection based on principal component analysis. *Mol. Ecol. Resour.* **17**, 67–77 (2017).

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

3. Wilde, S. *et al.* Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 4832–4837 (2014).

4. Pickrell, J. K. *et al.* Signals of recent positive selection in a worldwide sample of human populations. *Genome Research* vol. 19 826–837 (2009).

5. Lao, O., de Gruijter, J. M., van Duijn, K., Navarro, A. & Kayser, M. Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms. *Ann. Hum. Genet.* **71**, 354–369 (2007).

6. Herchuelz, A. *Sodium-calcium exchange and the plasma membrane Ca2+-ATPase in cell function: fifth international conference*. (Wiley-Blackwell, 2007).

7. Bryk, J. *et al.* Positive Selection in East Asians for an EDAR Allele that Enhances NF-κB Activation. *PLoS ONE* vol. 3 e2209 (2008).

8. Hider, J. L. *et al.* Exploring signatures of positive selection in pigmentation candidate genes in populations of East Asian ancestry. *BMC Evol. Biol.* **13**, 150 (2013).

9. Duforet-Frebourg, N., Luu, K., Laval, G., Bazin, E. & Blum, M. G. B. Detecting Genomic Signatures of Natural Selection with Principal Component Analysis: Application to the 1000 Genomes Data. *Mol. Biol. Evol.* **33**, 1082–1093 (2016).

10. Prashanth, S. & Deshmukh, S. Ectodermal Dysplasia: A Genetic Review. *International Journal of Clinical Pediatric Dentistry* vol. 5 197–202 (2012).

11. Kimura, R. *et al.* A common variation in EDAR is a genetic determinant of shovel-shaped incisors. *Am. J. Hum. Genet.* **85**, 528–535 (2009).

12. Izawa, T. *et al.* ASXL2 Regulates Glucose, Lipid, and Skeletal Homeostasis. *Cell Rep.* **11**, 1625–1637 (2015).

13. Zou, W. *et al.* Myeloid-specific Asxl2 deletion limits diet-induced obesity by regulating energy expenditure. *J. Clin. Invest.* **130**, 2644–2656 (2020).

14. Park, U.-H., Yoon, S. K., Park, T., Kim, E.-J. & Um, S.-J. Additional Sex Comb-like (ASXL) Proteins 1 and 2 Play Opposite Roles in Adipogenesis via Reciprocal Regulation of Peroxisome Proliferator-activated Receptor γ. *Journal of Biological Chemistry* vol. 286 1354–1363 (2011).

15. Ponsuksili, S. *et al.* Epigenome-wide skeletal muscle DNA methylation profiles at the background of distinct metabolic types and ryanodine receptor variation in pigs. *BMC Genomics* **20**, 492 (2019).

16. Samad, M. B. *et al.* [6]-Gingerol, from Zingiber officinale, potentiates GLP-1 mediated glucose-stimulated insulin secretion pathway in pancreatic β-cells and increases RAB8/RAB10-regulated membrane presentation of GLUT4 transporters in skeletal muscle to improve hyperglycemia in Leprdb/db type 2 diabetic mice. *BMC Complementary and Alternative Medicine* vol. 17 (2017).

17. Vazirani, R. P. *et al.* Disruption of Adipose Rab10-Dependent Insulin Signaling Causes Hepatic Insulin Resistance. *Diabetes* **65**, 1577–1589 (2016).

18. Hsieh, P. *et al.* Exome Sequencing Provides Evidence of Polygenic Adaptation to a Fat-Rich Animal Diet in Indigenous Siberian Populations. *Mol. Biol. Evol.* **34**, 2913–2926 (2017).

19. Thapa, D. *et al.* The protein acetylase GCN5L1 modulates hepatic fatty acid oxidation activity via acetylation of the mitochondrial β-oxidation enzyme HADHA. *J. Biol. Chem.* **293**, 17676–17684 (2018).

20. Baloni, P. *et al.* Genome-scale metabolic model of the rat liver predicts effects of diet restriction. *Sci. Rep.* **9**, 9807 (2019).

21. Ong, H. S. & Yim, H. C. H. Microbial Factors in Inflammatory Diseases and Cancers. *Regulation of Inflammatory Signaling in Health and Disease* 153–174 (2017) doi:10.1007/978-981-10-5987-2_7.

22. Logsdon, B. A., Hoffman, G. E. & Mezey, J. G. Mouse obesity network reconstruction with a variational Bayes algorithm to employ aggressive false positive control. *BMC Bioinformatics* **13**, 53 (2012).

23. Pei, Y.-F. *et al.* Genomic variants at 20p11 associated with body fat mass in the European population. *Obesity* **25**, 757–764 (2017).

24. Sabir, J. S. M. *et al.* Unraveling the role of salt-sensitivity genes in obesity with integrated network biology and co-expression analysis. *PLoS One* **15**, e0228400 (2020).

25. Wu, H. *et al.* Transcriptome Sequencing to Detect the Potential Role of Long Noncoding RNAs in Salt-Sensitive Hypertensive Rats. *Biomed Res. Int.* **2019**, 2816959 (2019).

26. Wang, L. *et al.* Peakwide Mapping on Chromosome 3q13 Identifies the Kalirin Gene as a Novel Candidate Gene for Coronary Artery Disease. *The American Journal of Human Genetics* vol. 80 650–663 (2007).

27. Ikram, M. A., Seshadri, S. & Bis, J. C. Genomewide Association Studies of Stroke. *Journal of Vascular Surgery* vol. 50 467 (2009).

28. Krug, T. *et al.* Kalirin: a novel genetic risk factor for ischemic stroke. *Hum. Genet.* **127**, 513–523 (2010).

29. Zang, X.-L. *et al.* Association of a SNP in SLC35F3 Gene with the Risk of Hypertension in a Chinese Han Population. *Frontiers in Genetics* vol. 7 (2016).

30. Zhang, K. *et al.* Genetic implication of a novel thiamine transporter in human hypertension. *J. Am. Coll. Cardiol.* **63**, 1542–1555 (2014).

31. Pacheu-Grau, D. *et al.* COA6 Facilitates Cytochrome c Oxidase Biogenesis as Thiol-reductase for Copper Metallochaperones in Mitochondria. *J. Mol. Biol.* **432**, 2067–2079 (2020).

32. Russo, L. *et al.* Cholesterol 25-hydroxylase (CH25H) as a promoter of adipose tissue inflammation in obesity and diabetes. *Mol Metab* **39**, 100983 (2020).
33. Zhao, J., Chen, J., Li, M., Chen, M. & Sun, C. Multifaceted Functions of CH25H and 25HC to Modulate the Lipid Metabolism, Immune Responses, and Broadly Antiviral Activities. *Viruses* **12**, (2020).
34. Demir, A., Kahraman, R., Candan, G. & Ergen, A. The role of FAS gene variants in inflammatory bowel disease. *Turk. J. Gastroenterol.* **31**, 356–361 (2020).
35. Rieux-Laucat, F., Magérus-Chatinet, A. & Neven, B. The Autoimmune Lymphoproliferative Syndrome with Defective FAS or FAS-Ligand Functions. *Journal of Clinical Immunology* vol. 38 558–568 (2018).
36. Karis, K. *et al.* Altered Expression Profile of IgLON Family of Neural Cell Adhesion Molecules in the Dorsolateral Prefrontal Cortex of Schizophrenic Patients. *Front. Mol. Neurosci.* **11**, 8 (2018).
37. Liu, J., Li, M. & Su, B. GWAS-identified schizophrenia risk SNPs atTSPAN18are highly diverged between Europeans and East Asians. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* vol. 171 1032–1040 (2016).
38. Fu, H.-Y. *et al.* The mutation spectrum of the SLC25A13 gene in Chinese infants with intrahepatic cholestasis and aminoacidemia. *J. Gastroenterol.* **46**, 510–518 (2011).
39. Chen, J.-L., Zhang, Z.-H., Li, B.-X., Cai, Z. & Zhou, Q.-H. Bioinformatic and functional analysis of promoter region of human SLC25A13 gene. *Gene* **693**, 69–75 (2019).
40. Vitoria, I. *et al.* Citrin deficiency in a Romanian child living in Spain highlights the worldwide distribution of this defect and illustrates the value of nutritional therapy. *Molecular Genetics and Metabolism* vol. 110 181–183 (2013).
41. Fiermonte, G. *et al.* An adult with type 2 citrullinemia presenting in Europe. *N. Engl. J. Med.* **358**, 1408–1409 (2008).
42. Sarkar, A. & Nandineni, M. R. Association of common genetic variants with

human skin color variation in Indian populations. *Am. J. Hum. Biol.* **30**, (2018).

43. Edwards, M. *et al.* Association of the OCA2 polymorphism His615Arg with melanin content in east Asian populations: further evidence of convergent evolution of skin pigmentation. *PLoS Genet.* **6**, e1000867 (2010).

44. Eiberg, H. *et al.* Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression. *Human Genetics* vol. 123 177–187 (2008).

45. Sturm, R. A. *et al.* A Single SNP in an Evolutionary Conserved Region within Intron 86 of the HERC2 Gene Determines Human Blue-Brown Eye Color. *The American Journal of Human Genetics* vol. 82 424–431 (2008).

46. Allentoft, M. E. *et al.* Population genomics of Bronze Age Eurasia. *Nature* **522**, 167–172 (2015).

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

48. Barreiro, L. B. *et al.* Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. *PLoS Genet.* **5**, e1000562 (2009).

49. Astiz, M. & Oster, H. GLUT12-A promising new target for the treatment of insulin resistance in obesity and type 2 diabetes. *Acta physiologica* vol. 226 e13329 (2019).

50. Waller, A. P. *et al.* GLUT12 functions as a basal and insulin-independent glucose transporter in the heart. *Biochim. Biophys. Acta* **1832**, 121–127 (2013).

51. Joehanes, R. *et al.* Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies. *Genome Biol.* **18**, 16 (2017).

52. CARDIoGRAMplusC4D Consortium *et al.* Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat. Genet.* **45**, 25–33 (2013).

53. Limaye, N. *et al.* Somatic mutations in angiopoietin receptor gene TEK cause solitary and multiple sporadic venous malformations. *Nat. Genet.* **41**, 118–124 (2009).

54. Jones, N., Iljin, K., Dumont, D. J. & Alitalo, K. Tie receptors: new modulators of angiogenic and lymphangiogenic responses. *Nature Reviews Molecular Cell Biology* vol. 2 257–267 (2001).

55. Dumont, D. J. *et al.* Dominant-negative and targeted null mutations in the endothelial receptor tyrosine kinase, tek, reveal a critical role in vasculogenesis of the embryo. *Genes Dev.* **8**, 1897–1909 (1994).

56. Gál, Z. *et al.* Investigation of the Possible Role of Tie2 Pathway and TEK Gene in Asthma and Allergic Conjunctivitis. *Frontiers in Genetics* vol. 11 (2020).

57. Cmejla, R. *et al.* Human MRCKalpha is regulated by cellular iron levels and interferes with transferrin iron uptake. *Biochem. Biophys. Res. Commun.* **395**, 163–167 (2010).

58. Richard, C. *et al.* Myotonic dystrophy kinase-related CDC42-binding kinase α, a new transferrin receptor type 2-binding partner, is a regulator of erythropoiesis. *Am. J. Hematol.* (2021) doi:10.1002/ajh.26104.

59. Molineros, J. E. *et al.* Mechanistic Characterization of RASGRP1 Variants Identifies an hnRNP-K-Regulated Transcriptional Enhancer Contributing to SLE Susceptibility. *Frontiers in Immunology* vol. 10 (2019).

60. Potier, M. L. *et al.* RasGRP1 and RasGRP3 expression in lymphocytes of rheumatoid arthritis patients. *Annals of the Rheumatic Diseases* vol. 71 A54.2–A54 (2012).

61. Somekh, I. *et al.* Correction to: Novel Mutations in RASGRP1 Are Associated with Immunodeficiency, Immune Dysregulation, and EBV-Induced Lymphoma. *J. Clin. Immunol.* **38**, 711 (2018).

62. Karlas, A. *et al.* Genome-wide RNAi screen identifies human host factors crucial for influenza virus replication. *Nature* **463**, 818–822 (2010).

63. Hao, L. *et al.* Drosophila RNAi screen identifies host genes important for influenza virus replication. *Nature* **454**, 890–893 (2008).

64. Mariniello, B. *et al.* Analysis of the 11β-Hydroxysteroid Dehydrogenase Type

2 Gene (HSD11B2) in Human Essential Hypertension*. *Am. J. Hypertens.* **18**, 1091–1098 (2005).

65. Zeller, T. *et al.* Transcriptome-Wide Analysis Identifies Novel Associations With Blood Pressure. *Hypertension* **70**, 743–750 (2017).

66. Gunaratnam, K., Vidal, C., Gimble, J. M. & Duque, G. Mechanisms of palmitate-induced lipotoxicity in human osteoblasts. *Endocrinology* **155**, 108–116 (2014).

67. Lumish, H. S., O'Reilly, M. & Reilly, M. P. Sex Differences in Genomic Drivers of Adipose Distribution and Related Cardiometabolic Disorders: Opportunities for Precision Medicine. *Arterioscler. Thromb. Vasc. Biol.* **40**, 45–60 (2020).

68. Ng, M. C. Y. *et al.* Discovery and fine-mapping of adiposity loci using high density imputation of genome-wide association studies in individuals of African ancestry: African Ancestry Anthropometry Genetics Consortium. *PLoS Genet.* **13**, e1006719 (2017).

69. Ueda, K. *et al.* Renal Dysfunction Induced by Kidney-Specific Gene Deletion of as a Primary Cause of Salt-Dependent Hypertension. *Hypertension* **70**, 111–118 (2017).

70. Agarwal, A. K. *et al.* CA-Repeat polymorphism in intron 1 of HSD11B2 : effects on gene expression and salt sensitivity. *Hypertension* **36**, 187–194 (2000).

71. Buckley, M. T. *et al.* Selection in Europeans on Fatty Acid Desaturases Associated with Dietary Changes. *Mol. Biol. Evol.* **34**, 1307–1318 (2017).

72. Ye, K., Gao, F., Wang, D., Bar-Yosef, O. & Keinan, A. Dietary adaptation of FADS genes in Europe varied across time and geography. *Nat Ecol Evol* **1**, 167 (2017).

73. Mathieson, S. & Mathieson, I. FADS1 and the Timing of Human Adaptation to Agriculture. *Mol. Biol. Evol.* **35**, 2957–2970 (2018).

74. Pan, G. *et al.* PATZ1 down-regulates FADS1 by binding to rs174557 and is opposed by SP1/SREBP1c. *Nucleic Acids Res.* **45**, 2408–2422 (2017).

75. Tabur, S. *et al.* Evidence for elevated (LIMK2 and CFL1) and suppressed

(ICAM1, EZR, MAP2K2, and NOS3) gene expressions in metabolic syndrome. *Endocrine* **53**, 465–470 (2016).

76. Fairn, G. D. & McMaster, C. R. Emerging roles of the oxysterol-binding protein family in metabolism, transport, and signaling. *Cell. Mol. Life Sci.* **65**, 228–236 (2008).

77. Lehto, M. & Olkkonen, V. M. The OSBP-related proteins: a novel protein family involved in vesicle transport, cellular lipid metabolism, and cell signalling. *Biochim. Biophys. Acta* **1631**, 1–11 (2003).

78. Sánchez-Solana, B., Li, D.-Q. & Kumar, R. Cytosolic functions of MORC2 in lipogenesis and adipogenesis. *Biochim. Biophys. Acta* **1843**, 316–326 (2014).

79. Sartipy, P., Camejo, G., Svensson, L. & Hurt-Camejo, E. Phospholipase A2 modification of lipoproteins: potential effects on atherogenesis. *Adv. Exp. Med. Biol.* **507**, 3–7 (2002).

80. Dennis, E. A., Cao, J., Hsu, Y.-H., Magrioti, V. & Kokotos, G. Phospholipase A2 Enzymes: Physical Structure, Biological Function, Disease Implication, Chemical Inhibition, and Therapeutic Intervention. *Chem. Rev.* **111**, 6130 (2011).

81. Wyles, J. P., Perry, R. J. & Ridgway, N. D. Characterization of the sterol-binding domain of oxysterol-binding protein (OSBP)-related protein 4 reveals a novel role in vimentin organization. *Experimental Cell Research* vol. 313 1426–1437 (2007).

82. Tintle, N. L. *et al.* A genome-wide association study of saturated, mono- and polyunsaturated red blood cell fatty acids in the Framingham Heart Offspring Study. *Prostaglandins Leukot. Essent. Fatty Acids* **94**, 65–72 (2015).

83. Thomas, C. *et al.* LPCAT3 deficiency in hematopoietic cells alters cholesterol and phospholipid homeostasis and promotes atherosclerosis. *Atherosclerosis* **275**, 409–418 (2018).

84. Liu, N. *et al.* Hyperuricemia induces lipid disturbances mediated by LPCAT3 upregulation in the liver. *FASEB J.* (2020) doi:10.1096/fj.202000950R.

85. Godahewa, G. I., Bathige, S. D. N. K., Herath, H. M. L. P. B., Noh, J. K. &

Lee, J. Characterization of rock bream (Oplegnathus fasciatus) complement components C1r and C1s in terms of molecular aspects, genomic modulation, and immune responsive transcriptional profiles following bacterial and viral pathogen exposure. *Fish Shellfish Immunol.* **46**, 656–668 (2015).

86. Dai, D.-F. *et al.* Plasma concentration of SCUBE1, a novel platelet protein, is elevated in patients with acute coronary syndrome and ischemic stroke. *J. Am. Coll. Cardiol.* **51**, 2173–2180 (2008).

87. Huang, X., Liu, G., Guo, J. & Su, Z. The PI3K/AKT pathway in obesity and type 2 diabetes. *Int. J. Biol. Sci.* **14**, 1483–1496 (2018).

88. Zhong, X. *et al.* LNK deficiency decreases obesity-induced insulin resistance by regulating GLUT4 through the PI3K-Akt-AS160 pathway in adipose tissue. *Aging* **12**, 17150–17166 (2020).

89. López-Gómez, C. *et al.* Oleic Acid Protects Against Insulin Resistance by Regulating the Genes Related to the PI3K Signaling Pathway. *J. Clin. Med. Res.* **9**, (2020).

90. Sandrini, L. *et al.* Association between Obesity and Circulating Brain-Derived Neurotrophic Factor (BDNF) Levels: Systematic Review of Literature and Meta-Analysis. *Int. J. Mol. Sci.* **19**, (2018).

91. Lommatzsch, M. *et al.* The impact of age, weight and gender on BDNF levels in human platelets and plasma. *Neurobiol. Aging* **26**, 115–123 (2005).

92. Tsuchida, A. *et al.* Brain-derived neurotrophic factor ameliorates lipid metabolism in diabetic mice. *Diabetes Obes. Metab.* **4**, 262–269 (2002).

93. Wu, A., Molteni, R., Ying, Z. & Gomez-Pinilla, F. A saturated-fat diet aggravates the outcome of traumatic brain injury on hippocampal plasticity and cognitive function by reducing brain-derived neurotrophic factor. *Neuroscience* **119**, 365–375 (2003).

94. Molteni, R., Barnard, R. J., Ying, Z., Roberts, C. K. & Gómez-Pinilla, F. A high-fat, refined sugar diet reduces hippocampal brain-derived neurotrophic factor,

neuronal plasticity, and learning. *Neuroscience* **112**, 803–814 (2002).

95. Graber, T. G., Borack, M. S., Reidy, P. T., Volpi, E. & Rasmussen, B. B. Essential amino acid ingestion alters expression of genes associated with amino acid sensing, transport, and mTORC1 regulation in human skeletal muscle. *Nutr. Metab.* **14**, 35 (2017).

96. Hellsten, S. V., Hägglund, M. G., Eriksson, M. M. & Fredriksson, R. The neuronal and astrocytic protein SLC38A10 transports glutamine, glutamate, and aspartate, suggesting a role in neurotransmission. *FEBS Open Bio* **7**, 730–746 (2017).

97. Tripathi, R., Hosseini, K., Arapi, V., Fredriksson, R. & Bagchi, S. SLC38A10 (SNAT10) is Located in ER and Golgi Compartments and Has a Role in Regulating Nascent Protein Synthesis. *International Journal of Molecular Sciences* vol. 20 6265 (2019).

98. Schmidt, S. *et al.* Effect of omega-3 polyunsaturated fatty acids on the cytoskeleton: an open-label intervention study. *Lipids Health Dis.* **14**, 4 (2015).

99. Zulkafli, I. S., Waddell, B. J. & Mark, P. J. Postnatal Dietary Omega-3 Fatty Acid Supplementation Rescues Glucocorticoid-Programmed Adiposity, Hypertension, and Hyperlipidemia in Male Rat Offspring Raised on a High-Fat Diet. *Endocrinology* vol. 154 3110–3117 (2013).

100. Aqil, M., Mallik, S., Bandyopadhyay, S., Maulik, U. & Jameel, S. Transcriptomic Analysis of mRNAs in Human Monocytic Cells Expressing the HIV-1 Nef Protein and Their Exosomes. *Biomed Res. Int.* **2015**, (2015).

101. Kirpich, I. A. *et al.* Integrated hepatic transcriptome and proteome analysis of mice with high-fat diet-induced nonalcoholic fatty liver disease. *J. Nutr. Biochem.* **22**, 38–45 (2011).

102. Li, H. *et al.* Diversification of the ADH1B gene during expansion of modern humans. *Ann. Hum. Genet.* **75**, 497–507 (2011).

103. Polimanti, R. & Gelernter, J. ADH1B: From alcoholism, natural selection, and cancer to the human phenome. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **177**,

113–125 (2018).

104. Zhong, X. *et al.* The zinc-finger protein ZFYVE1 modulates TLR3-mediated signaling by facilitating TLR3 ligand binding. *Cell. Mol. Immunol.* **17**, 741–752 (2020).

105. Yang, Y. *et al.* The RNA-binding protein Mex3B is a coreceptor of Toll-like receptor 3 in innate antiviral response. *Cell Res.* **26**, 288–303 (2016).

106. Kim, S. V. *et al.* GPR15-mediated homing controls immune homeostasis in the large intestine mucosa. *Science* **340**, 1456–1459 (2013).

107. Nguyen, L. P. *et al.* Role and species-specific expression of colon T cell homing receptor GPR15 in colitis. *Nature Immunology* vol. 16 207–213 (2015).

108. Monteleone, G., Boirivant, M., Pallone, F. & MacDonald, T. T. TGF-β1 and Smad7 in the regulation of IBD. *Mucosal Immunology* vol. 1 S50–S53 (2008).

109. Kennedy, B. W. C. Mongersen, an Oral SMAD7 Antisense Oligonucleotide, and Crohn's Disease. *The New England journal of medicine* vol. 372 2461 (2015).

110. Garo, L. P. *et al.* Smad7 Controls Immunoregulatory PDL2/1-PD1 Signaling in Intestinal Inflammation and Autoimmunity. *Cell Rep.* **28**, 3353–3366.e5 (2019).

111. Iqbal, Z. *et al.* Homozygous SLC6A17 mutations cause autosomal-recessive intellectual disability with progressive tremor, speech impairment, and behavioral problems. *Am. J. Hum. Genet.* **96**, 386–396 (2015).

112. Park, J. *et al.* KCNC1-related disorders: new de novo variants expand the phenotypic spectrum. *Ann Clin Transl Neurol* **6**, 1319–1326 (2019).

4e) Correlation between components of variation in population structure and components of variation in SNP-trait association

Alba Refoyo Martínez[1], Fernando Racimo[1]

[1]Lundbeck Foundation GeoGenetics Centre, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark

Introduction

Ancient Western and Central Eurasian populations show strong patterns of genome-wide differentiation, which are even stronger than those observed among present-day Western and Central Eurasians (Supplementary Note S4d). We were interested in determining whether variants that contribute to differentiation between populations may also be associated with complex traits. This could perhaps serve to evince the genetic component of trait differences among these ancient populations.

Recently, Tanigawa et al., 2019[1] developed a method (DeGAs) by which SNP-trait associations from thousands of phenotypes could be represented in a low-dimensional space, yielding a smaller set of latent components of genetic associations. They performed GWAS in 337,199 White British individuals in the UK Biobank study for 2,138 phenotypes, obtained Z-scores from each SNP and each trait, and then performed truncated singular-value decomposition (TSVD) on the matrix of Z-scores, which yielded 100 components of variation. The decomposition allows one to see how much a particular variant or trait contributes to the variation explained by that component, and to rank the components by their amount of contribution to the total variation in SNP-trait associations.

By looking at these components of genetic association in the UK Biobank and comparing them with our components of variation in ancient population structure, we can begin to understand which sets of trait-associated variants may have been important during recent human evolution.

Methods

We carried out a principal component analysis (PCA) on the 1,165 West-Eurasian imputed genomes, using *pcadapt* [2]. At the variant level we filtered out, 1) variants with MAF < 5%, 2) genotype missingness rate > 50% and, 3) variants with a genotype probability lower than 0.8 in more than 10% of the samples (Figure 1).

Our aim is to study the correlation between our population structure components, captured in the first two components, and the 100 components of SNP-trait associations from the DeGAs analysis, which were obtained from [1]. Henceforth, we will refer to the components of population structure from the ancient genome PCA with the label “PS” (i.e. the first principal component is PS1, the second is PS2, etc.). We will refer to the components of trait-association variation from Tanigawa et al., 2019[1] as “DG” (i.e. the first component is DG1, the second is DG2, etc.). We applied the Pearson correlation between the loadings of the two analyses (Figure 3).

We aimed to test whether the correlations observed were significantly different from those that would be observed under a model in which there was no association between a particular DG component and a particular PS component. For this, we obtained p-values using a randomization scheme. We randomised the sign of the 1000 DGs loadings accounting for the structure of linkage disequilibrium (LD) along the genome, by dividing our genome into 1Mb blocks and randomising the sign of all SNPs within each block in unison recalculating the correlation with our first two components (PS1-PS2).

For each combination of PSx and DGy (where x and y are indices over all PS and DG components, respectively), we obtained a P-value using the following equation:

$$P = \frac{1 + \Sigma_i^N I(|cor_x^i| > |cor_x^y|)}{1 + N}$$

Here, cor^y_x is the true correlation between PSx and DGy, cor^y_x is the correlation between PSx and DGy after randomising signs, I() is an indicator function and N is the number of randomised samples used, which was set to 1,000.

Results

We first performed a PCA on the genotypes of the 1,165 imputed ancient West-Eurasian individuals (Supplementary Note S4d). The first component (1.8% of total variance explained) represents a gradient separating Neolithic farmer populations from hunter-gatherer genomes, while the second component (1.1%) captures a gradient separating ancient East Asians, ancient Steppe populations and ancient Western Eurasians (Figure **S4d.4**).

We computed the correlation between the first 8 components of population structure and the 100 trait-association components. Figure **S4e.1.** shows the distribution of all DG correlations with the first 8 PS components (PS1 to PS8). For this analysis, we focused on PS1 and PS2. PS1 captures a gradient separating ancient Neolithic farmers from Mesolithic hunter-gatherer genomes. The trait components most correlated with PS1 are DG89, DG71, DG22, DG15, DG52, DG5, DG82, DG2, DG51, DG99, DG69, DG12, DG84, DG83 and DG79 (Table XX and top four 4 in Figure **S4e.2**). These components have in common that are driven by anthropometric and lifestyle and environmental traits, with an important contribution from diet-related measures. They are mainly correlated with five different broader categories: 1) lifestyle and environment, mainly representing food intake, cereal and fruit, and time spent outdoor and sun exposure; 2) verbal interview about mental health

described by worrier/anxious/nervous feelings and sleep duration; 3) impedance, body measurement and bone mineral density; 4) lung capacity and asthma, mostly represented by the FEV1 measure; 5) blood pressure measures, hypertension and cholesterol.

PS2 separates ancient East Asian and European samples. The significant top correlations of DG components with PS2, top four shown in Figure **S4e.3**, are DG38, DG12, DG40, DG84, DG56, DG82, DG36, DG15, DG28. and DG68. One of them was also significant for PS1. These components are mainly driven by anthropometric measures and cardiovascular measurement and disorders. They can be classified into the same categories previously described. The main difference with PS1 is that more components are mainly explained by the lifestyle and environmental relation and impedance measure while in PS2, blood pressure and cardiovascular disorders and skin-sun exposure relation have higher contributions.

We computed p-values using two randomization schemes to test whether these correlations were significantly different from zero. The first test relies on the randomization of the sign of the loadings before calculating the correlations between the loadings representing components of structure and the loadings from the DeGAs analysis. We first divided the genome into 5Mb blocks and then randomised the signs of the SNPs in each block in the same direction, to account for LD. We also use a bootstrapping approach also dividing the genome into blocks.

We observe there are 22 significant correlations with PS1 while there are 7 significant correlations with PS2. Significant p-values for the block-based randomization for PS1 and PS2 are shown in Table S4e.1 and Table S4e.2 respectively.

Discussion

Among the significantly PS1-associated trait components, there are many components involving traits related to lifestyle and environment: eg. water intake, cereal intake and time spent outdoors. This signal may perhaps be related to differences in diet and lifestyle between ancient hunter-gatherers and farmers [3,4]. We also find traits related to haematological measurements, forced expiratory volume (see also Supplementary Note S4d), body fat, anxiety and hypertension, among others.

However, we should also keep in mind that population stratification in the original UK Biobank dataset may affect these results. For example, if present-day British individuals with different amounts of Neolithic farmer ancestry happen to live in different areas of Britain with markedly different lifestyles, one could presume that a spurious signal of correlation between

those lifestyle differences and amount of Neolithic ancestry could be generated. Indeed, while non-significant, we find that some of the trait association components that are most correlated with PS1 have to do with lifestyle and environmental traits such as tea intake, and mental health, and these may, in turn, may be correlated with variation in PS1 in present-day British people (due, perhaps, to more recent immigrants from Asia in urban settings, for example).

The strongest DG correlations with the first two PCs performed on the European populations from the 1000 Genomes Project (FIN, CEU, GBR, TSI and IBS) are quite different to the ones with the ancient samples. Only DG52, which is mainly explained by asthma and bone mineral density traits, is shared with the West-Eurasian study. The first component, which separates Finnish from the other european populations, is strongly correlated with those DGs in which haematological measurements contribute between 40-92% (Table S4e.3). The second component, that separates the British and Utah residents (CEPH) with Northern and Western European ancestry from the Southern European and Finnish populations. This component is correlated with DG components in which spirometry traits that measure lung capacity, lifestyle and environment, such as water, coffee and tea intake, and body measurement and impedance traits, are the ones contributing most (Table S4e.4).

Figures

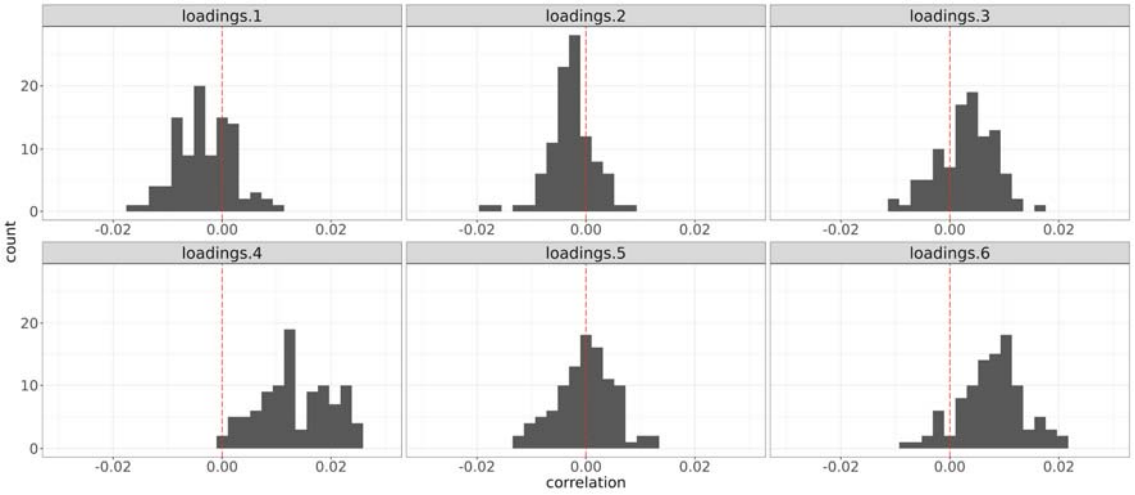
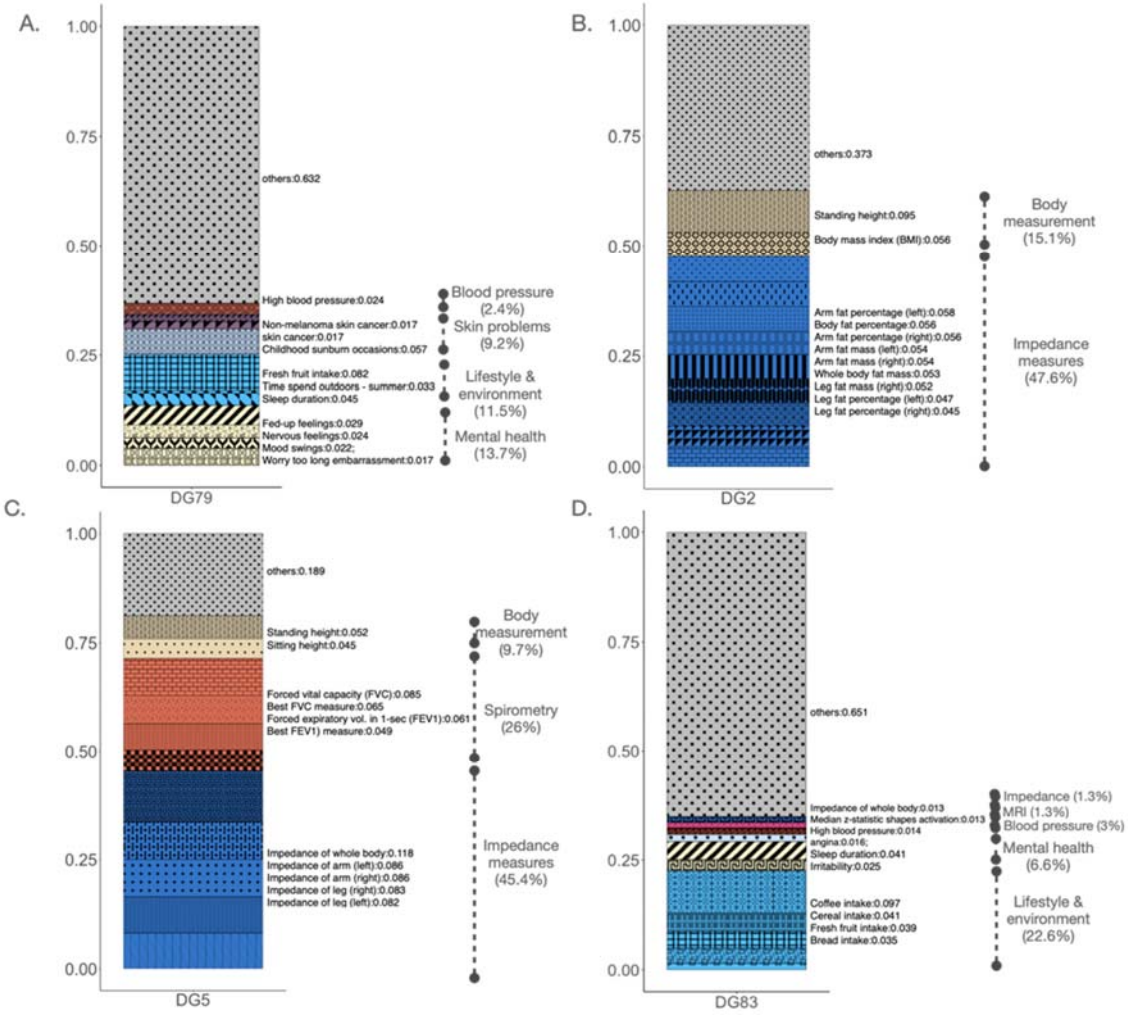


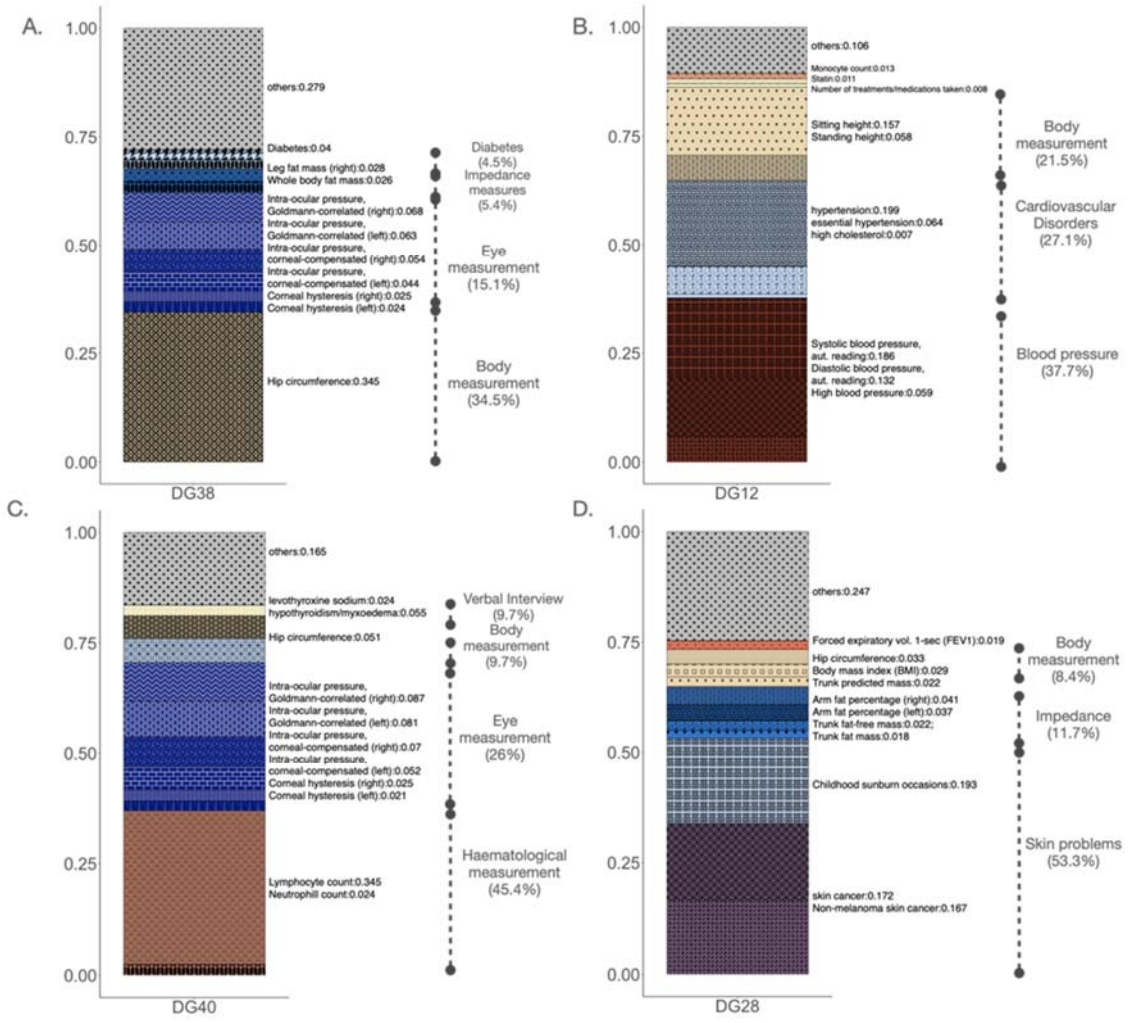
Figure S4e.1. Pearson correlation between the 6 loadings from the PCA analysis (PS1-PS6) and the 100 loadings from the DeGAs analysis (DG1-DG100).

5043



5044

5045 **Figure S4e.2. PS1.** Top 4 UK Biobank trait-association components from DeGAs that have
5046 the highest correlation (P < 5e-3) with the principal component separating ancient farmer
5047 from hunter-gatherer populations



5048

5049 **Figure S4e.3. PS2.** Top 4 UK Biobank trait-association components from DeGAs that have
5050 the highest correlation (P < 5e-3) with the principal component separating East Asian and
5051 European samples.

5052

5053

Neolithic farmer vs. hunter-gatherer genomes gradient				
DG	**Correlation**	**P-value**	**contribution_category**	**contribution_phenotypes**
79	-0.017	0.001	others:0.632; Blood pressure:0.024; Cancer:0.034; Non-cancer Illness:0.057; Lifestyle and environment:0.115; Verbal Interview:0.137	others:0.632; High blood pressure:0.024; Non-melanoma skin cancer:0.017; skin cancer:0.017; Childhood sunburn occasions:0.057; Fresh fruit intake:0.082; Time spend outdoors in summer:0.033; Sleep duration:0.045; Fed-up feelings:0.029; Nervous feelings:0.024; Mood swings:0.022; Worry too long after embarrassment:0.017
2	-0.014	0.001	others:0.373; Body measurement:0.151; Impedance measures:0.476	others:0.373; Standing height:0.095; Body mass index (BMI):0.056; Arm fat percentage (left):0.058; Body fat percentage:0.056; Arm fat percentage (right):0.056; Arm fat mass (left):0.054; Arm fat mass (right):0.054; Whole body fat mass:0.053; Leg fat mass (right):0.052; Leg fat percentage (left):0.047; Leg fat percentage (right):0.045
5	-0.013	0.003	others:0.189; Body measurement:0.097; Spirometry:0.26; Impedance measures:0.454	others:0.189; Standing height:0.052; Sitting height:0.045; Forced vital capacity (FVC):0.085; Forced vital capacity (FVC), Best measure:0.065; Forced expiratory volume in 1-second (FEV1):0.061; Forced expiratory volume in 1-second (FEV1), Best measure:0.049; Impedance of whole body:0.118; Impedance of arm (left):0.086; Impedance of arm (right):0.086; Impedance of leg (right):0.083; Impedance of leg (left):0.082
83	-0.013	0.003	others:0.651; Impedance measures:0.013; Brain MRI:0.013; Blood pressure:0.014; Non-cancer Illness:0.016; Verbal Interview:0.066; Lifestyle and environment:0.226	others:0.651; Impedance of whole body:0.013; Median z-statistic (in group-defined mask) for shapes activation:0.013; High blood pressure:0.014; angina:0.016; Sleep duration:0.041; Irritability:0.025; Coffee intake:0.097; Cereal intake:0.041; Fresh fruit intake:0.039; Bread intake:0.035; Water intake:0.015
71	-0.012	0.003	others:0.506; Non-cancer Illness:0.041; Hematological measurement:0.073; Verbal Interview:0.108; Lifestyle and environment:0.273	others:0.506; cholelithiasis/gall stones:0.024; cholecystitis:0.017; Monocyte percentage:0.036; Eosinophill percentage:0.02; Eosinophill count:0.016; **Number of treatments/medications taken:0.108; Fresh fruit intake:0.104; Time spend outdoors in summer:0.074**; Tea intake:0.045; Time spent outdoors in winter:0.033; Coffee intake:0.016
89	-0.012	0.005	others:0.58; Arterial stiffness:0.034; Spirometry:0.061; Lifestyle and environment:0.067; Non-cancer Illness:0.077; Verbal Interview:0.181	others:0.58; Pulse wave peak to peak time:0.017; Pulse wave Arterial Stiffness index:0.017; **Forced expiratory volume in 1-second (FEV1), predicted percentage:0.061**; Average weekly beer plus cider intake:0.067; **hypertension:0.042; essential hypertension:0.035; Worrier/anxious feelings:0.053; Sleep duration:0.053; Fed-up feelings:0.033**; Age at menopause (last menstrual period):0.022; Number of treatments/medications taken:0.021

22	0.014	0.00999	others:0.225; Hematological measurement:0.018; Impedance measures:0.029; Blood pressure:0.083; Non-cancer Illness:0.317; Verbal Interview:0.329	others:0.225; Neutrophill count:0.018; Impedance of arm (left):0.016; Leg fat percentage (right):0.013; Systolic blood pressure, automated reading:0.03; Diastolic blood pressure, automated reading:0.026; Pulse rate, automated reading:0.026; **high cholesterol:0.303**; Alzheimer's disease/dementia:0.014**; Statin:0.237**; simvastatin:0.062; atorvastatin:0.03
99	-0.01	0.01698	others:0.63; Spirometry:0.032; Lifestyle and environment:0.061; Impedance measures:0.067; Verbal Interview:0.087; Eye measurement:0.122	others:0.63; Forced expiratory volume in 1-second (FEV1), predicted percentage:0.032; Average weekly beer plus cider intake:0.061; Impedance of arm (left):0.041; Impedance of arm (right):0.026; Age started wearing glasses or contact lenses:0.071; Seen doctor (GP) for nerves, anxiety, tension or depression:0.016; logMAR in round (left):0.039; logMAR, initial (left):0.027; logMAR, final (left):0.024; Corneal resistance factor (left):0.017; Corneal hysteresis (right):0.015
80	-0.01	0.01698	others:0.652; Hematological measurement:0.017; Non-cancer Illness:0.017; Sex-specific factors:0.021; Lifestyle and environment:0.046; Verbal Interview:0.053; Impedance measures:0.061; Body measurement:0.131	others:0.652; Monocyte percentage:0.017; hypertension:0.017; Birth weight of first child:0.021; Water intake:0.027; Coffee intake:0.019; Worrier/anxious feelings:0.02; gout:0.018; Sleep duration:0.015; Arm predicted mass (left):0.032; Arm fat-free mass (left):0.029; Weight:0.131
69	-0.01	0.02198	others:0.529; Blood pressure:0.038; Impedance measures:0.068; Body measurement:0.089; Hematological measurement:0.132; Lifestyle and environment:0.144	others:0.529; High blood pressure:0.038; Impedance of whole body:0.045; Impedance of arm (right):0.023; Body mass index (BMI):0.062; Weight:0.027; Monocyte percentage:0.11; Eosinophill percentage:0.022; Cereal intake:0.047; Water intake:0.034; Time spend outdoors in summer:0.033; Time spent outdoors in winter:0.029
77	-0.009	0.01798	others:0.748; Urine assays:0.042; Lifestyle and environment:0.043; Verbal Interview:0.049; Brain MRI:0.118	others:0.748; Sodium in urine:0.042; Fresh fruit intake:0.043; Sleep duration:0.028; Number of treatments/medications taken:0.021; Weighted-mean FA in tract superior longitudinal fasciculus (right):0.022; Weighted-mean FA in tract superior longitudinal fasciculus (left):0.019; Mean FA in superior longitudinal fasciculus on FA skeleton (right):0.019; Weighted-mean MD in tract anterior thalamic radiation (right):0.018; Weighted-mean MD in tract anterior thalamic radiation (left):0.015; Weighted-mean L1 in tract anterior thalamic radiation (left):0.013; Mean FA in superior longitudinal fasciculus on FA skeleton (left):0.012
22	0.009	0.01898	others:0.225; Hematological measurement:0.018; Impedance measures:0.029; Blood pressure:0.083; Non-cancer Illness:0.317; Verbal Interview:0.329	others:0.225; Neutrophill count:0.018; Impedance of arm (left):0.016; Leg fat percentage (right):0.013; Systolic blood pressure, automated reading:0.03; Diastolic blood pressure, automated reading:0.026; Pulse rate, automated reading:0.026; high cholesterol:0.303; Alzheimer's disease/dementia:0.014; Statin:0.237; simvastatin:0.062; atorvastatin:0.03
90	-0.009	0.03197	others:0.631; Impedance measures:0.021; Eye	others:0.631; Impedance of arm (right):0.021; logMAR in round (left):0.027; Creatinine (enzymatic) in urine:0.027; Nucleated red blood cell percentage:0.03;

			measurement:0.027; Urine assays:0.027; Blood pressure:0.03; Spirometry:0.034; Non-cancer Illness:0.057; Verbal Interview:0.075; Lifestyle and environment:0.098	Forced expiratory volume in 1-second (FEV1), predicted percentage:0.034; Childhood sunburn occasions:0.034; essential hypertension:0.022; Number of treatments/medications taken:0.05; Irritability:0.025; Fresh fruit intake:0.05; Bread intake:0.048
75	-0.009	0.02797	others:0.509; Verbal Interview:0.071; Non-cancer Illness:0.084; Hematological measurement:0.107; Impedance measures:0.108; Lifestyle and environment:0.12	others:0.509; Number of treatments/medications taken:0.071; Alzheimer's disease/dementia:0.061; essential hypertension:0.023; Monocyte percentage:0.088; Monocyte count:0.019; Whole body water mass:0.041; Whole body fat-free mass:0.04; Basal metabolic rate:0.027; Fresh fruit intake:0.048; Time spend outdoors in summer:0.043; Time spent outdoors in winter:0.029
51	0.011	0.038896	others:0.425; Verbal Interview:0.024; DXA assessment:0.26; Non-cancer Illness:0.291	others:0.425; Wheeze or whistling in the chest in last year:0.024; Femur total BMD (bone mineral density) (left):0.038; Femur total BMD (bone mineral density) T-score (left):0.035; Femur total BMD (bone mineral density) (right):0.033; Femur total BMD (bone mineral density) T-score (right):0.031; Femur troch BMD (bone mineral density) (left):0.028; Femur troch BMD (bone mineral density) (right):0.026; Femur shaft BMD (bone mineral density) (left):0.024; Femur troch BMD (bone mineral density) T-score (left):0.023; Femur troch BMD (bone mineral density) T-score (right):0.021; asthma:0.291
61	-0.009	0.037896	others:0.612; Urine assays:0.016; Lifestyle and environment:0.049; Impedance measures:0.055; Body measurement:0.127; Spirometry:0.14	others:0.612; Sodium in urine:0.016; Water intake:0.036; Time spend outdoors in summer:0.013; Leg fat mass (right):0.027; Arm fat percentage (right):0.015; Trunk fat percentage:0.013; Body mass index (BMI):0.063; Weight:0.035; Waist circumference:0.028; Peak expiratory flow (PEF):0.11; Forced expiratory volume in 1-second (FEV1):0.03
82	-0.012	0.040896	others:0.641; Urine assays:0.027; Body measurement:0.073; Lifestyle and environment:0.078; Impedance measures:0.09; Verbal Interview:0.092	others:0.641; Creatinine (enzymatic) in urine:0.027; Hand grip strength (right):0.038; Hand grip strength (left):0.035; Tea intake:0.063; Water intake:0.015; Impedance of leg (left):0.035; Impedance of leg (right):0.022; Trunk fat mass:0.017; Trunk fat mass:0.017; Sleep duration:0.066; Nervous feelings:0.026
52	0.013	0.022928	others:0.538; Verbal Interview:0.019; Impedance measures:0.027; DXA assessment:0.171; Non-cancer Illness:0.245	others:0.538; Wheeze or whistling in the chest in last year:0.019; Impedance of whole body:0.027; Femur total BMD (bone mineral density) (left):0.027; Femur total BMD (bone mineral density) T-score (left):0.025; Femur total BMD (bone mineral density) (right):0.024; Femur total BMD (bone mineral density) T-score (right):0.022; Femur troch BMD (bone mineral density) (left):0.02; Femur troch BMD (bone mineral density) (right):0.019; Femur shaft BMD (bone mineral density) (left):0.017; Femur troch BMD (bone mineral density) T-score (left):0.016; asthma:0.245
68	-0.01	0.047945	others:0.697; Verbal Interview:0.015; Hematological measurement:0.022; Impedance measures:0.036; Lifestyle and	others:0.697; Long-standing illness, disability or infirmity:0.015; Monocyte percentage:0.022; Leg fat-free mass (left):0.018; Leg predicted mass (left):0.018; Tea intake:0.037; High blood pressure:0.061; Alzheimer's

			environment:0.037; Blood pressure:0.061; Non-cancer Illness:0.132	disease/dementia:0.038; cholelithiasis/gall stones:0.03; hypertension:0.03; cholecystitis:0.019; Childhood sunburn occasions:0.015
84	-0.015	0.004	others:0.648; Urine assays:0.051; Blood pressure:0.07; Verbal Interview:0.077; Lifestyle and environment:0.154	others:0.648; Creatinine (enzymatic) in urine:0.035; Potassium in urine:0.016; High blood pressure:0.07; Sleep duration:0.022; **Fed-up feelings:0.02; Worrier/anxious feelings:0.019; Nervous feelings:0.017; Tea intake:0.072; Time spend outdoors in summer:0.042; Coffee intake:0.022; Time spent outdoors in winter:0.018**
46	-0.011	0.01499	others:0.096; Body measurement:0.045; Non-cancer Illness:0.055; Blood pressure:0.181; Spirometry:0.219; Verbal Interview:0.406	others:0.096; Waist circumference:0.045; diabetes:0.02; Diabetes diagnosed by doctor:0.018; Diabetes:0.017; Diastolic blood pressure, automated reading:0.111; Systolic blood pressure, automated reading:0.07; Forced vital capacity (FVC):0.069; Forced vital capacity (FVC), Best measure:0.066; Forced expiratory volume in 1-second (FEV1), Best measure:0.061; Forced expiratory volume in 1-second (FEV1):0.023; Age when periods started (menarche):0.406
70	-0.008	0.03896	others:0.57; Body measurement:0.033; Urine assays:0.047; Non-cancer Illness:0.053; Impedance measures:0.135; Lifestyle and environment:0.162	others:0.57; Weight:0.033; Sodium in urine:0.047; Childhood sunburn occasions:0.026; Alzheimer's disease/dementia:0.026; Impedance of leg (right):0.037; Impedance of leg (left):0.028; Arm predicted mass (left):0.026; Arm fat-free mass (left):0.023; Arm predicted mass (right):0.021; Cereal intake:0.111; Water intake:0.051

Table S4e.1. Significant correlations between DG components with PS1 1) randomising the sign of each block of 5Mb in the DG loadings and bootsrapping the same blocks.

5057

Ancient East Asians, ancient Steppe populations and ancient Western Eurasians gradient				
DG	**Correlation**	**P-value**	**contribution_category**	**contribution_phenotypes**
38	-0.024	0.001	others:0.279; Non-cancer Illness:0.045; Impedance measures:0.054; **Eye measurement:0.277; Body measurement:0.345**	others:0.279; diabetes:0.024; Diabetes:0.02; Leg fat mass (right):0.028; Whole body fat mass:0.026; Intra-ocular pressure, Goldmann-correlated (right):0.068; Intra-ocular pressure, Goldmann-correlated (left):0.063; Intra-ocular pressure, corneal-compensated (right):0.054; Intra-ocular pressure, corneal-compensated (left):0.044; Corneal hysteresis (right):0.025; Corneal hysteresis (left):0.024; **Hip circumference:0.345**
12	-0.024	0.001	others:0.106; Hematological measurement:0.013; Verbal Interview:0.019; Body measurement:0.215; **Non-cancer Illness:0.271; Blood pressure:0.377**	others:0.106; Monocyte count:0.013; Statin:0.011; Number of treatments/medications taken:0.008; Sitting **height**:0.157; Standing height:0.058; **hypertension**:0.199; essential hypertension:0.064; high cholesterol:0.007; Systolic **blood pressure**, automated reading:0.186; Diastolic blood pressure, automated reading:0.132; High blood pressure:0.059
40	-0.015	0.00799	others:0.165; Verbal Interview:0.024; Body measurement:0.051; Non-cancer Illness:0.055; **Eye measurement:0.336; Hematological measurement:0.369**	others:0.165; levothyroxine sodium:0.024; Hip circumference:0.051; hypothyroidism/myxoedema:0.055; Intra-ocular pressure, Goldmann-correlated (right):0.087; Intra-ocular pressure, Goldmann-correlated (left):0.081; Intra-ocular pressure, corneal-compensated (right):0.07; Intra-ocular pressure, corneal-compensated (left):0.052; Corneal hysteresis (right):0.025; Corneal hysteresis (left):0.021; Lymphocyte count:0.345; Neutrophill count:0.024
28	-0.01	0.001	others:0.247; Spirometry:0.019; Body measurement:0.084; Impedance measures:0.117; Non-cancer Illness:0.193; **skin Cancer**:0.339	others:0.247; Forced expiratory volume in 1-second (FEV1), predicted:0.019; Hip circumference:0.033; Body mass index (BMI):0.029; Trunk predicted mass:0.022; Arm fat percentage (right):0.041; Arm fat percentage (left):0.037; Trunk fat-free mass:0.022; Trunk fat mass:0.018; Childhood sunburn occasions:0.193; skin cancer:0.172; Non-melanoma skin cancer:0.167
23	-0.009	0.028937	others:0.061; Hematological measurement:0.011; Non-cancer Illness:0.014; Cancer:0.02; Local environment:0.894	others:0.061; Immature reticulocyte fraction:0.011; Childhood sunburn occasions:0.014; skin cancer:0.01; Non-melanoma skin cancer:0.01; Nitrogen dioxide air pollution; 2010:0.176; Nitrogen dioxide air pollution; 2007:0.171; Nitrogen dioxide air pollution; 2006:0.167; Nitrogen dioxide air pollution; 2005:0.152; Nitrogen oxides air pollution; 2010:0.113; Particulate matter air pollution (pm2.5); 2010:0.091; Particulate matter air pollution (pm2.5) absorbance; 2010:0.024
24	-0.009	0.042976	others:0.056; Body measurement:0.022; Hematological measurement:0.922	others:0.056; Hand grip strength (left):0.012; Hand grip strength (right):0.011; Immature reticulocyte fraction:0.325; Reticulocyte count:0.234; High light scatter reticulocyte percentage:0.142; Reticulocyte percentage:0.121; High light scatter reticulocyte count:0.051; Neutrophill count:0.028; White blood cell (leukocyte) count:0.01; Red blood cell (erythrocyte) count:0.007; Lymphocyte count:0.004

84	-0.015	0.004	others:0.648; Urine assays:0.051; Blood pressure:0.07; Verbal Interview:0.077; Lifestyle and environment:0.154	others:0.648; Creatinine (enzymatic) in urine:0.035; Potassium in urine:0.016; High blood pressure:0.07; Sleep duration:0.022; **Fed-up feelings:0.02; Worrier/anxious feelings:0.019; Nervous feelings:0.017; Tea intake:0.072; Time spend outdoors in summer:0.042; Coffee intake:0.022; Time spent outdoors in winter:0.018**

Table S4e.2. Significant correlations between DG components with PS2 1) randomising the sign of each block of 5Mb in the DG loadings and bootsrapping the same blocks

1000 GP – FIN vs other European populations gradient.				
DG	**Correlation**	**P-value**	**contribution_category**	**contribution_phenotypes**
46	-0.011	0.01499	others:0.096; Body measurement:0.045; Non-cancer Illness:0.055; Blood pressure:0.181; Spirometry:0.219; Verbal Interview:0.406	others:0.096; Waist circumference:0.045; diabetes:0.02; Diabetes diagnosed by doctor:0.018; Diabetes:0.017; Diastolic blood pressure, automated reading:0.111; Systolic blood pressure, automated reading:0.07; Forced vital capacity (FVC):0.069; Forced vital capacity (FVC), Best measure:0.066; Forced expiratory volume in 1-second (FEV1), Best measure:0.061; Forced expiratory volume in 1-second (FEV1):0.023; Age when periods started (menarche):0.406
27	0.009	0.02498	others:0.093; Hematological measurement:0.907	others:0.093; Mean platelet (thrombocyte) volume:0.408; Neutrophill count:0.11; Red blood cell (erythrocyte) count:0.096; Platelet distribution width:0.052; Haemoglobin concentration:0.05; Red blood cell (erythrocyte) distribution width:0.041; Haematocrit percentage:0.037; Mean corpuscular haemoglobin concentration:0.036; Mean corpuscular haemoglobin:0.034; Monocyte percentage:0.025; Lymphocyte percentage:0.018
20	0.009	0.03996	others:0.189; Verbal Interview:0.016; Non-cancer Illness:0.022; Impedance measures:0.051; Hematological measurement:0.723	others:0.189; Statin:0.016; high cholesterol:0.022; Leg fat percentage (right):0.027; Leg fat percentage (left):0.024; Neutrophill count:0.353; White blood cell (leukocyte) count:0.13; Red blood cell (erythrocyte) count:0.068; Monocyte count:0.064; Monocyte percentage:0.038; Mean platelet (thrombocyte) volume:0.037; Immature reticulocyte fraction:0.033
24	-0.009	0.04296	others:0.056; Body measurement:0.022; Hematological measurement:0.922	others:0.056; Hand grip strength (left):0.012; Hand grip strength (right):0.011; Immature reticulocyte fraction:0.325; Reticulocyte count:0.234; High light scatter reticulocyte percentage:0.142; Reticulocyte percentage:0.121; High light scatter reticulocyte count:0.051; Neutrophill count:0.028; White blood cell (leukocyte) count:0.01; Red blood cell (erythrocyte) count:0.007; Lymphocyte count:0.004

Table S4e.3. Significant correlations between DG components with PS1 of European population in 1000 Genomes Project 1) randomising the sign of each block of 5Mb in the DG loadings and bootsrapping the same blocks.

1000 GP – FIN and Southern Europeans vs Northern populations gradient.

DG	Correlation	P-value	contribution_category	contribution_phenotypes
46	-0.014	0.002	others:0.096; Body measurement:0.045; Non-cancer Illness:0.055; Blood pressure:0.181; Spirometry:0.219; **Verbal Interview**:0.406	others:0.096; Waist circumference:0.045; diabetes:0.02; Diabetes diagnosed by doctor:0.018; Diabetes:0.017; Diastolic blood pressure, automated reading:0.111; Systolic blood pressure, automated reading:0.07; Forced vital capacity (FVC):0.069; Forced vital capacity (FVC), Best measure:0.066; Forced expiratory volume in 1-second (FEV1), Best measure:0.061; Forced expiratory volume in 1-second (FEV1):0.023; Age when periods started (menarche):0.406
45	-0.014	0.003	others:0.099; Non-cancer Illness:0.008; Hematological measurement:0.014; Body measurement:0.082; Blood pressure:0.132; Verbal Interview:0.177; **Spirometry**:0.489	others:0.099; asthma:0.008; Red blood cell (erythrocyte) count:0.014; Waist circumference:0.082; Diastolic blood pressure, automated reading:0.071; Systolic blood pressure, automated reading:0.061; Age when periods started (menarche):0.177; Forced vital capacity (FVC):0.141; Forced vital capacity (FVC), Best measure:0.133; Forced expiratory volume in 1-second (FEV1), Best measure:0.132; Forced expiratory volume in 1-second (FEV1):0.052; Peak expiratory flow (PEF):0.031
96	-0.011	0.01798	others:0.661; Blood pressure:0.019; Eye measurement:0.02; Spirometry:0.025; Sex-specific factors:0.027; Impedance measures:0.034; Non-cancer Illness:0.043; Verbal Interview:0.047; Body measurement:0.123	others:0.661; Pulse rate:0.019; 6mm weak meridian (left):0.02; Forced expiratory volume in 1-second (FEV1), predicted percentage:0.025; Birth weight of first child:0.027; Impedance of arm (left):0.034; aortic dissection:0.022; aortic aneurysm:0.021; Number of treatments/medications taken:0.028; Age at menopause (last menstrual period):0.019; Hand grip strength (right):0.063; Hand grip strength (left):0.06
30	-0.01	0.01099	others:0.341; Eye measurement:0.033; Verbal Interview:0.037; Non-cancer Illness:0.076; Body measurement:0.101; Cancer:0.153; **Impedance measures**:0.258	others:0.341; Spherical power (left):0.033; Neuroticism score:0.037; Childhood sunburn occasions:0.076; Body mass index (BMI):0.069; Trunk predicted mass:0.031; skin cancer:0.077; Non-melanoma skin cancer:0.076; Arm fat percentage (right):0.081; Arm fat percentage (left):0.07; Trunk fat mass:0.053; Trunk fat mass:0.053
52	-0.009	0.03796	others:0.538; Verbal Interview:0.019; Impedance measures:0.027; **DXA** assessment:0.171; Non-cancer Illness:0.245	others:0.538; Wheeze or whistling in the chest in last year:0.019; Impedance of whole body:0.027; Femur total BMD (bone mineral density) (left):0.027; Femur total BMD (bone mineral density) T-score (left):0.025; Femur total BMD (bone mineral density) (right):0.024; Femur total BMD (bone mineral density) T-score (right):0.022; Femur troch BMD (bone mineral density) (left):0.02; Femur troch BMD (bone mineral density) (right):0.019; Femur shaft BMD (bone mineral density) (left):0.017; Femur troch BMD (bone mineral density) T-score (left):0.016; **asthma**:0.245
59	-0.009	0.03796	others:0.455; Urine assays:0.011; Non-cancer Illness:0.023; Brain MRI:0.046; **Lifestyle and environment**:0.218; **Spirometry**:0.247	others:0.455; Sodium in urine:0.011; asthma:0.023; Volume of brain, grey+white:0.016; Volume of grey matter:0.016; Volume of peripheral cortical grey matter:0.014; Water intake:0.131; Tea intake:0.046; Coffee intake:0.041; Peak

				expiratory flow (PEF):0.175; Forced expiratory volume in 1-second (FEV1):0.049; Forced vital capacity (FVC), Best measure:0.022
1000	-0.008	0.04795	others:0.746; Impedance measures:0.014; Hematological measurement:0.019; Urine assays:0.024; Sex-specific factors:0.027; Eye measurement:0.036; Verbal Interview:0.039; Lifestyle and environment:0.095	others:0.746; Impedance of leg (right):0.014; Monocyte percentage:0.019; Sodium in urine:0.024; Birth weight of first child:0.027; Corneal hysteresis (left):0.019; Corneal resistance factor (right):0.017; Age started wearing glasses or contact lenses:0.039; Average weekly beer plus cider intake:0.047; Fresh fruit intake:0.02; Tea intake:0.015; Coffee intake:0.014

Table S4e.4. Significant correlations between DG components with PS2 of European population in 1000 Genomes Project 1) randomising the sign of each block of 5Mb in the DG loadings and bootsrapping the same blocks.

References

1. Tanigawa, Y. *et al.* Components of genetic associations across 2,138 phenotypes in the UK Biobank highlight adipocyte biology. *Nat. Commun.* **10**, 4064 (2019).

2. Luu, K., Bazin, E. & Blum, M. G. B. pcadapt: an R package to perform genome scans for selection based on principal component analysis. *Mol. Ecol. Resour.* **17**, 67–77 (2017).

3. Diamond, J. & Bellwood, P. Farmers and their languages: the first expansions. *Science* **300**, 597–603 (2003).

4. Bowles, S. Cultivation of cereals by the first farmers was not more productive than foraging. *Proceedings of the National Academy of Sciences* vol. 108 4760–4765 (2011).

4f) Polygenic prediction for height, eye colour and hair colour in ancient Danish samples

Anders Rosengren[1,2], Vivek Appadurai[1,2], Andrew Schork[1-4], Andrés Ingason[1-3]

[1]Institute of Biological Psychiatry, Mental Health Services, Copenhagen University Hospital
[2]iPSYCH, The Lundbeck Foundation Initiative for Integrative Psychiatric Research, Aarhus
[3]Lundbeck Foundation GeoGenetics Centre, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark
[4]Neurogenomics Division, The Translational Genomics Research Institute (TGEN), Phoenix

We predicted relative “genetic” height (i.e. expected increase or decrease in height compared to the mean of the contemporary Danish population, based on common genetic variants) as well as eye and hair colour in 100 ancient samples excavated from archeological sites in Denmark. The estimated age of the ancient samples, all sequenced in this study, ranged from roughly 10,500 to 3,000 years (See Figure 4 in main text).

The relative genetic height was calculated from summary statistics of a recent GWAS on adult height in the UK Biobank [1], using only strand-insensitive autosomal SNPs with robustly genome-wide significant allelic effects (P<1e-15) with imputation info >0.8 and minor allele frequency >0.05 in our ancient imputation genotype dataset. We excluded long-range linkage disequilibrium (LD) regions [2] and used the “--clump” function in PLINK [3] to prune for

LD (r2<0.1 within 10Mb window), rendering a total of 310 effect alleles. Per-sample genetic height score was then calculated for the 100 ancient samples as well as a subset of 3,467 Danish ancestry male conscripts from a random population subset of the IPSYCH2012 case-cohort [4] by summing allelic effect multiplied with the effect allele imputed dosage (Appadurai *et al*., in prep.) across the 310 loci. The genetic height score was moderately correlated with height in the subset of 3,467 Danish ancestry conscripts (r^2 = 0.095, P = 3.2e-77), and we rescaled the score to a unit corresponding to 1 cm change in predicted height with the median score in the conscript subset as zero. Thus, the PGS in each ancient sample corresponds to the predicted difference in cm from the average of the present-day Danish population, assuming that the scores are equally predictive in males and females. The genetic height score is however limited in two important ways; firstly, the predictive value of the genetic height score is modest and diminishes with general genetic distance to the population in which the allelic effects were determined (in this case European ancestry British); secondly, the score does not take account of important environmental factors such as health and access to nutrition in childhood.

The genetic predictions of eye and hair colour were done based on the HIrisPlex system [5]. Out of 24 main effect HIrisPlex variants, genotype likelihoods of 18 SNPs were available in the ancient sample 1000G imputation, and imputed effect allele dosages of these were used to derive probabilities for brown, blue and grey/intermediate eye colour and blond, brown, black and red hair colour, following the HIrisPlex formulas [5].

The results of genetic prediction of height, eye and hair colour are shown alongside results of other analyses in the composite figure 4 in the main text, titled “Denmark through time”. When excluding 27 samples with average genomic sequence coverage <0.1x (indicated in shaded colour in figure 4 and with “0” in the Cov QC column of the Table S4f.1 below), the predicted genetic height differed significantly across the three groups defined by the two major population turnover events from Mesolithic Hunter-Gatherer (ML-HG) to Neolithic Early Farmer (NL-EF), and later to Neolithic Steppe Pastoralist (NL-SP), and indicated with thick lines through the panels of figure 4 (ANOVA chi-square test across all three groups, $P=5.1x10^{-8}$). A follow-up pairwise testing found significant differences between all three group pairs (linear regression, P<0.05 in all instances), with the lowest mean (plus/minus standard error) predicted relative genetic height observed in ML-HG (-1.9 ± 0.3 cm), then NL-EF (-0.3 ± 0.5 cm), and highest in NL-SP (1.2 ± 0.4 cm). It should be borne in mind that the population structure of the ancient Danish samples, especially the ML-HG and NL-EF groups, is different from the current day European ancestry British population (in which allelic effects for genetic height were estimated), and although we have used only very

robustly associated effect alleles to calculate the genetic height score, it is likely that it will not have correlated as well with actual height as it does in the current-day European ancestry Danish population (in which the score was rescaled). Therefore, the only conclusion that can be drawn from these results is that the common SNP alleles that contribute most strongly to increased height in current day European ancestry populations, were on average of slightly lower frequency in ML-HG, in similar frequency in NL-EF, and slightly higher frequency in NL-SP.

Among the 18 HIrisPlex SNPs used to predict eye and hair colour, rs12913832 has the strongest overall dark/light pigmentation effect. At the same time rs12913832 has the lowest average maximum genotype probability (GPmax) of the HIrisPlex SNPs across the ancient Danish imputed genotype dataset. To account for this, we applied a second quality filter when comparing predicted eye and hair colour probabilities across groups, by requiring a GPmax>0.6 for rs12913832 and for at least 15 of the other 17 pigmentation SNPs, which removed a further 17 samples in the cross-group comparison (marked with "0" in the Pigm QC column in Table S4f.1 below). In this comparison we did not find a significant difference in probability of brown eye colour (pEye Brown in Table S4f.1 below) across the three groups (ANOVA, P=0.21). On the other hand, the predicted probability of blond hair colour differed significantly across groups (ANOVA, $P=1.1x10^{-9}$), with the mean likelihood (pHair Blond in Table S4f.1 below) increasing over time from ML-HG (0.05 ± 0.01) to NL-EF (0.25 ± 0.06) and NL-SP (0.43 ± 0.07), although the difference between NL-EF and NL-SP was not significant (P=0.07). Although pigmentation traits are polygenic, many of the HIrisPlex system alleles are so-called main effect alleles and therefore it is likely that the increased predicted probabilities for blond hair over time (and corresponding decrease in predicted probabilities for black hair) represent a true change in the prevalence of dark and light hair colour.

Supplementary Table S4f.1. Estimated age, genomic coverage, predicted genetic height difference from current Danish population average, and predicted likelihood of eye and hair colour of 100 ancient Danish samples sequenced in this study

Sample	Age (ybp)	Age group	Genomic coverage	Cov QC	Pigm QC	pHeight (cm)	pEye Blue	pEye Inter	pEye Brown	pHair Blond	pHair Red	pHair Brown	pHair Black
NEO254	10463	ML-HG	0.42	1	0	-1.2	0.09	0.13	0.78	0.04	0	0.19	0.77
NEO13	9507	ML-HG	0.01	0	0	-0.7	0.43	0.19	0.38	0.18	0	0.36	0.46
NEO91	9122	ML-HG	1.18	1	1	-1.8	0.44	0.22	0.34	0.03	0	0.26	0.71
NEO759	9028	ML-HG	2.95	1	1	0.2	0.58	0.16	0.26	0.04	0	0.21	0.75

NEO587	8798	ML-HG	1.14	1	1	-3.5	0.73	0.13	0.14	0.01	0	0.18	0.81
NEO123	8182	ML-HG	0.29	1	1	-3	0.09	0.13	0.78	0.11	0	0.26	0.63
NEO19	8163	ML-HG	3.26	1	1	-3.5	0.53	0.27	0.2	0	0	0.28	0.72
NEO122	8146	ML-HG	0.56	1	0	-0.4	0.12	0.22	0.66	0	0	0.15	0.85
NEO600	7817	ML-HG	0.10	0	1	-0.7	0.01	0.09	0.9	0.02	0	0.28	0.7
NEO683	7529	ML-HG	1.82	1	1	-1.8	0.26	0.25	0.49	0.01	0	0.22	0.77
NEO932	7499	ML-HG	2.76	1	1	-4.6	0.64	0.19	0.17	0.01	0	0.15	0.84
NEO589	7478	ML-HG	7.41	1	1	-1	0.46	0.21	0.33	0.02	0	0.24	0.74
NEO748	7129	ML-HG	0.08	0	0	-0.2	0.32	0.17	0.51	0.05	0	0.27	0.68
NEO814	7125	ML-HG	0.06	0	0	-2.6	0.08	0.16	0.76	0.03	0	0.24	0.73
NEO749	7070	ML-HG	1.91	1	1	-1.3	0.57	0.19	0.24	0.03	0	0.26	0.71
NEO791	7048	ML-HG	2.49	1	1	-0.1	0.84	0.11	0.05	0.05	0	0.31	0.64
NEO586	7031	ML-HG	0.20	1	1	-1.4	0.51	0.21	0.28	0.02	0	0.28	0.7
NEO746	6991	ML-HG	0.14	1	0	-3.1	0.16	0.15	0.69	0.11	0	0.29	0.6
NEO583	6981	ML-HG	0.18	1	1	-2.5	0.44	0.15	0.41	0.2	0	0.28	0.52
NEO822	6978	ML-HG	0.06	0	0	0.7	0.47	0.18	0.35	0.13	0	0.32	0.55
NEO930	6888	ML-HG	0.05	0	0	-2.1	0.4	0.18	0.42	0.07	0	0.26	0.67
NEO733	6824	ML-HG	1.32	1	1	-1.1	0.55	0.24	0.21	0.03	0	0.34	0.63
NEO732	6815	ML-HG	0.13	1	1	-4.5	0.34	0.24	0.42	0.03	0	0.36	0.61
NEO745	6790	ML-HG	0.45	1	1	-3	0.03	0.11	0.86	0.01	0	0.16	0.83
NEO856	6777	ML-HG	0.56	1	1	-2.3	0.39	0.22	0.39	0.02	0	0.19	0.79
NEO747	6729	ML-HG	0.25	1	0	-2.1	0.12	0.25	0.63	0	0	0.25	0.75
NEO568	6586	ML-HG	1.98	1	1	-4.4	0.57	0.19	0.24	0.01	0	0.26	0.73
NEO1	6585	ML-HG	0.02	0	0	0.8	0.09	0.15	0.76	0.1	0	0.27	0.63
NEO941	6372	ML-HG	0.14	1	1	-5	0.19	0.21	0.6	0.03	0	0.29	0.68
NEO570	6369	ML-HG	2.86	1	1	-0.6	0.72	0.15	0.13	0.01	0	0.19	0.8

NEO751	6343	ML-HG	0.30	1	1	-0.1	0.62	0.17	0.21	0.11	0	0.33	0.56
NEO852	6308	ML-HG	0.19	1	1	1	0.44	0.24	0.32	0.02	0	0.32	0.66
NEO855	6302	ML-HG	1.38	1	1	-0.4	0.76	0.13	0.11	0.02	0	0.27	0.71
NEO569	6142	ML-HG	0.67	1	1	-2.2	0.64	0.17	0.19	0.01	0	0.2	0.79
NEO598	6075	ML-HG	0.73	1	1	-0.5	0.09	0.15	0.76	0	0	0.17	0.83
NEO853	6047	ML-HG	1.96	1	1	-1.8	0.88	0.08	0.04	0.04	0	0.3	0.66
NEO3	5965	ML-HG	0.03	0	0	-1	0.15	0.16	0.69	0.14	0	0.28	0.58
NEO960	5926	ML-HG	0.15	1	1	-3.9	0.75	0.11	0.14	0.32	0	0.32	0.36
NEO645	5870	ML-HG	0.21	1	1	-0.4	0.85	0.09	0.06	0.09	0	0.33	0.58
NEO962	5786	ML-HG	0.04	0	0	-1.9	0.07	0.11	0.82	0.2	0	0.27	0.53
NEO601	5753	NL-EF	0.08	0	0	1.5	0	0.02	0.98	0.01	0	0.21	0.78
NEO790	5662	NL-EF	0.69	1	0	-4.2	0.28	0.16	0.56	0.25	0	0.29	0.46
NEO891	5661	NL-EF	0.60	1	1	1.1	0.69	0.14	0.17	0.11	0	0.35	0.54
NEO571	5534	NL-EF	0.06	0	0	-1.5	0.11	0.15	0.74	0.29	0	0.28	0.43
NEO23	5533	NL-EF	3.34	1	1	-3.4	0.19	0.27	0.54	0.17	0	0.35	0.48
NEO753	5531	NL-EF	0.16	1	1	-0.4	0	0.03	0.97	0.02	0	0.15	0.83
NEO942	5491	NL-EF	0.89	1	1	0.4	0.7	0.12	0.18	0.08	0	0.33	0.59
NEO29	5489	NL-EF	0.53	1	1	3.9	0.7	0.15	0.15	0.45	0	0.31	0.24
NEO564	5468	NL-EF	0.08	0	0	-2.6	0.35	0.18	0.47	0.13	0	0.32	0.55
NEO41	5462	NL-EF	0.02	0	0	1.4	0.05	0.13	0.82	0.13	0	0.32	0.55
NEO28	5459	NL-EF	0.92	1	0	2.7	0.43	0.2	0.37	0.22	0	0.3	0.48
NEO886	5457	NL-EF	0.27	1	1	-1.9	0.7	0.13	0.17	0.66	0	0.19	0.15
NEO866	5456	NL-EF	1.52	1	1	-0.5	0.84	0.09	0.07	0.72	0	0.19	0.09
NEO595	5452	NL-EF	0.22	1	0	-2.3	0.53	0.15	0.32	0.41	0	0.28	0.31
NEO757	5452	NL-EF	0.13	1	0	-0.3	0.05	0.11	0.84	0.03	0	0.27	0.7
NEO896	5446	NL-EF	0.12	1	0	-1.8	0.01	0.05	0.94	0.03	0	0.17	0.8

NEO945	5445	NL-EF	1.38	1	1	0.3	0.01	0.05	0.94	0.03	0	0.22	0.75
NEO888	5383	NL-EF	0.06	0	0	1.5	0.09	0.13	0.78	0.18	0	0.28	0.54
NEO933	5337	NL-EF	0.52	1	1	2.4	0.65	0.15	0.2	0.36	0	0.31	0.33
NEO744	5333	NL-EF	0.22	1	0	3.6	0.03	0.12	0.85	0.22	0	0.32	0.46
NEO795	5333	NL-EF	0.03	0	0	-1	0.13	0.16	0.71	0.2	0	0.33	0.47
NEO702	5263	NL-EF	0.15	1	1	-0.7	0.52	0.24	0.24	0.02	0	0.31	0.67
NEO7	5242	NL-EF	0.01	0	0	-1.1	0	0.04	0.96	0.03	0	0.23	0.74
NEO597	5210	NL-EF	0.18	1	0	0.5	0.1	0.14	0.76	0.31	0	0.29	0.4
NEO935	5187	NL-EF	5.03	1	1	-1.4	0.09	0.15	0.76	0.44	0	0.27	0.29
NEO865	5179	NL-EF	0.09	0	0	-3.2	0.01	0.06	0.93	0.06	0	0.23	0.71
NEO594	5174	NL-EF	0.05	0	0	-1.8	0.47	0.19	0.34	0.31	0	0.33	0.36
NEO961	5137	NL-EF	0.02	0	0	1.1	0.24	0.19	0.57	0.22	0	0.33	0.45
NEO599	5134	NL-EF	0.19	1	0	-2.6	0.02	0.11	0.86	0.08	0	0.24	0.68
NEO602	5134	NL-EF	0.09	0	0	-1.4	0.25	0.17	0.58	0.47	0	0.25	0.28
NEO566	5130	NL-EF	0.02	0	0	-3.2	0.44	0.17	0.39	0.38	0	0.28	0.34
NEO33	5128	NL-EF	0.05	0	0	-3.1	0.07	0.15	0.78	0.02	0	0.2	0.78
NEO898	5080	NL-EF	3.8	1	1	-4	0.51	0.13	0.36	0.21	0	0.3	0.49
NEO43	5067	NL-EF	0.11	1	0	1.9	0.05	0.14	0.81	0.17	0	0.36	0.47
NEO25	4956	NL-EF	0.36	1	1	1.1	0.09	0.16	0.75	0.08	0	0.33	0.59
NEO925	4947	NL-EF	0.29	1	0	-2.8	0	0.03	0.97	0.05	0	0.25	0.7
NEO943	4614	NL-EF	1.75	1	1	2.1	0.03	0.07	0.9	0.2	0	0.26	0.54
NEO580	4611	NL-EF	0.01	0	0	-1.3	0.03	0.09	0.88	0.09	0	0.28	0.63
NEO792	4493	NL-SP	0.25	1	1	-0.4	0.04	0.09	0.87	0.14	0	0.27	0.59
NEO876	4338	NL-SP	0.05	0	0	-2.4	0.01	0.08	0.91	0.16	0	0.31	0.53
NEO870	4240	NL-SP	0.58	1	1	2.2	0.23	0.26	0.51	0.37	0	0.33	0.3
NEO92	4188	NL-SP	0.61	1	1	-0.9	0.82	0.11	0.07	0.48	0	0.25	0.27

NEO737	4106	NL-SP	0.24	1	1	1	0.77	0.15	0.08	0.27	0	0.41	0.32
NEO738	4103	NL-SP	1.21	1	1	-0.3	0.92	0.05	0.03	0.73	0	0.17	0.1
NEO861	4102	NL-SP	0.38	1	1	0.5	0.79	0.11	0.1	0.74	0	0.17	0.09
NEO878	4026	NL-SP	0.28	1	1	-2.3	0.45	0.15	0.4	0.37	0	0.33	0.3
NEO872	3979	NL-SP	0.08	0	0	-0.1	0.62	0.15	0.23	0.44	0	0.3	0.26
NEO735	3972	NL-SP	0.67	1	1	2.5	0.78	0.11	0.11	0.76	0.01	0.16	0.07
NEO875	3970	NL-SP	0.18	1	0	1.6	0.07	0.11	0.82	0.18	0	0.27	0.55
NEO739	3965	NL-SP	1.88	1	1	0.9	0.01	0.04	0.95	0.02	0	0.15	0.83
NEO934	3809	NL-SP	0.08	0	0	2	0.19	0.15	0.66	0.29	0	0.31	0.4
NEO93	3735	NL-SP	1.98	1	1	-0.3	0.26	0.25	0.49	0.1	0	0.39	0.51
NEO860	3697	NL-SP	0.20	1	0	1.2	0.55	0.2	0.25	0.66	0	0.21	0.13
NEO857	3637	NL-SP	0.04	0	0	-1.9	0.09	0.17	0.74	0.26	0	0.31	0.43
NEO752	3589	NL-SP	2.09	1	1	1.8	0.64	0.12	0.24	0.31	0	0.31	0.38
NEO815	3471	NL-SP	0.11	1	0	4.3	0.48	0.24	0.28	0.33	0.04	0.42	0.21
NEO563	3350	NL-SP	0.81	1	1	3.6	0.69	0.14	0.17	0.59	0.09	0.19	0.13
NEO590	3290	NL-SP	1.01	1	1	2.9	0.85	0.07	0.08	0.84	0.01	0.11	0.04
NEO951	3242	NL-SP	0.45	1	0	1.7	0.01	0.06	0.93	0.2	0	0.34	0.46
NEO946	3094	NL-SP	1.24	1	1	2.2	0.8	0.1	0.1	0.28	0.01	0.37	0.34

References

1. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

2. Price, A. L. *et al.* Long-range LD can confound genome scans in admixed populations. *American journal of human genetics* vol. 83 132–5; author reply 135–9 (2008).

3. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

4. Pedersen, C. B. *et al.* The iPSYCH2012 case-cohort sample: new directions for unravelling genetic and environmental architectures of severe mental disorders. *Mol. Psychiatry* **23**, 6–14 (2018).

5. Walsh, S. *et al.* The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. *Forensic Sci. Int. Genet.* **7**, 98–115 (2013).

4g) Calling chr17q21.31 KANSL1 Duplications in Ancient Genomes

Alma S. Halgren[1], Andrés Ingason[2,3], and Peter H. Sudmant[1]

[1]Department of Integrative Biology, University of California, Berkeley
[2]Institute of Biological Psychiatry, Mental Health Services, Copenhagen University Hospital
[3]Lundbeck Foundation GeoGenetics Centre, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark

To call the distinct H1 and H2 specific KANSL1 duplications at chr17q21.31 (H1D and H2D respectively), we first removed samples that were too noisy to be accurately genotyped. To do so, we calculated the standard deviation of genome-wide copy number (after removing the top and bottom fifth percentiles of copy number to exclude outliers). We chose standard deviation cutoffs based on a visual inspection of the data – 0.49 for the fastq samples and 0.804 for the bam samples. While bam and fastq samples exhibited genome-wide differences in coverage and noise resulting in these different cutoffs, the selected samples exhibit similar signatures at the KANSL1 locus. Together this resulted in a total of 1143 samples, 427 fastq samples and 716 bam samples. We then set a copy number cutoff of 10 at the KANSL1 locus (copy number signal above 10 is likely noise) and calculated the average copy number in the H1D and H2D coordinate blocks (Fig **S4g.1**A). We filtered for samples with an average copy number in both the H1D and H2D blocks between 1.5 and 5.25 (values outside of these bounds are likely noise). The samples clustered into duplication genotypes (colored below) with the exception of a handful of samples on the border of genotype groups for which we could not accurately assign a genotype (marked

with ‘x’). Figure **S4g.1B** shows the SNP inversion calls for these samples, which align with the duplication calls.

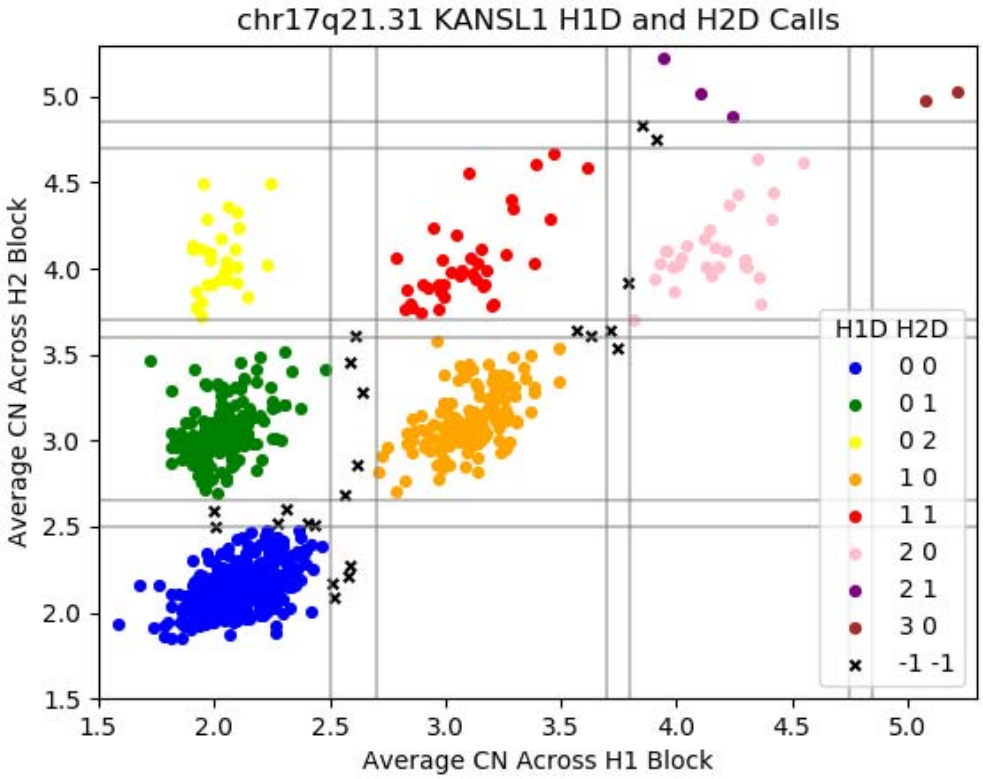


Figure S4g.1A. *KANSL1* H1D and H2D calls grouped by genotype. The samples marked with an ‘x’ are considered ambiguous as they are between groups.

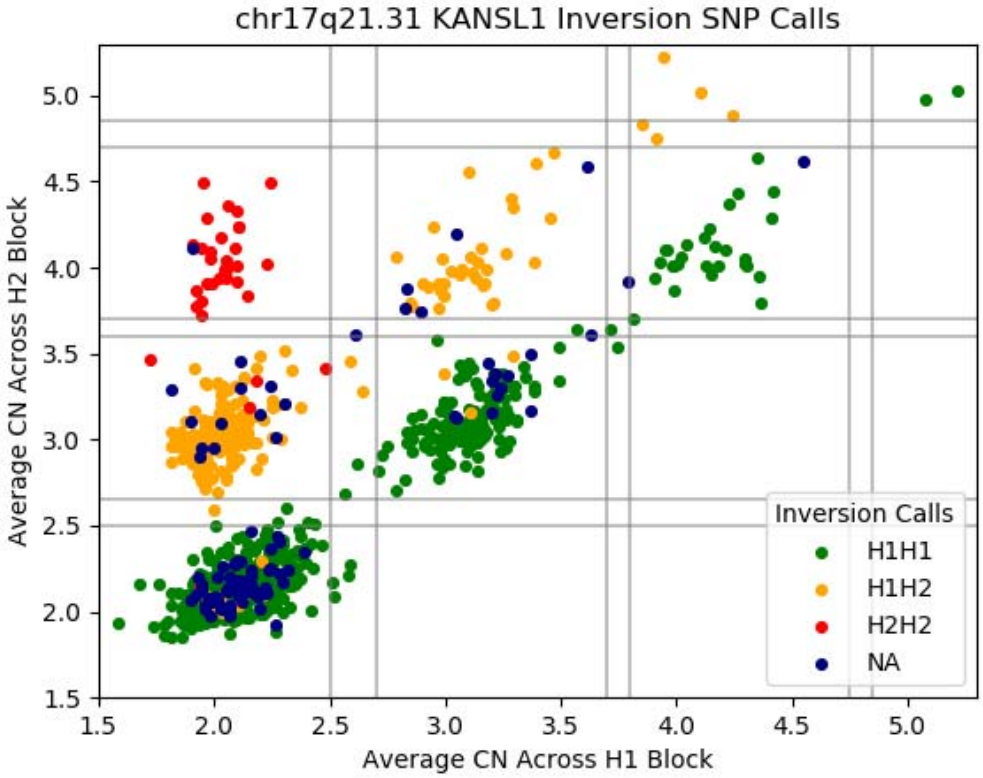


Figure S4g.1B. *KANSL1* H1 and H2 inversion calls mapped atop H1D and H2D average copy number values. The inversion and duplication calls align.

4h) Calculating ancestral contributions to modern complex phenotypes

Will Barrie[1] and Dan Lawson[2]

[1]Zoology Department, University of Cambridge, UK.

[2]School of Mathematics and Integrative Epidemiology Unit, University of Bristol, UK.

Introduction

Most studies that look at polygenic risk scores in ancient populations use genotypes of ancient individuals, combined with effect sizes from modern GWAS studies, to reconstruct risk scores for ancient individuals [1]. This involves exporting effect sizes across space and

time, which is known to dramatically reduce the accuracy of the estimates [2]. Additionally, these scores are usually impossible to verify (except with specific phenotypes such as height where calibration is possible [3,4], and don't necessarily measure what an ancient population contributed to phenotypic diversity in a modern population(s), especially when there has been selection or bottleneck events in between.

Here, we aim to use local ancestry information resulting from painting the UK Biobank (UKB) (Supplementary Note S3h) to estimate ancestral contributions to modern complex phenotypes, by calculating polygenic risk scores for each ancestry based on local painting results. This is a well-powered approach due to the large modern sample size, and is a more direct measure of the variants that a given ancestry contributed to the "white British" genetic landscape. Thus we can draw conclusions about the differing contributions of each ancestry to modern genetic risk, whether due to drift or selection. We use bootstrapping to test whether some ancestries are significantly and systematically over-represented for a phenotype, indicating selection. Additionally, we look at the ancestral haplotypic background of a high effect variant, ApoE4, which is implicated in Alzheimer's Disease [5,6].

Methods

We used effect size estimates from the UK Biobank Neale lab GWAS [7], and used 1,703 non-overlapping and approximately independent linkage disequilibrium (LD) blocks [8]. For each block, we restricted the SNPs to those with a p-value less than the genome-wide significance threshold (5e-8), and from these chose the SNP with the lowest p-value. We then used these SNPs to calculate polygenic risk scores for each ancestry, using ancestry-specific 'effect allele frequencies' derived from the painting.

In order to calculate the effect allele frequency for a given ancestry f_{anc,i} we used the formula:

$$f_{\{anc,i\}} = \frac{\sum_j^M \quad Painting\ certainty_{\{j,effect\}}}{\sum_j^M \quad Painting\ certainty_{\{j,effect\}} + \sum_j^M \quad Painting\ certainty_{\{j,alt\}}}$$

Where there are M individuals, and $\sum_j^M \quad Painting\ certainty_{\{j,effect\}}$ is the sum of the painting probabilities for that ancestry of all effect alleles. This calculates an effect allele frequency for an ancestry which is weighted by the painting probabilities: if a haplotype with the effect allele was painted with low probability for that ancestry, it will contribute little to the

calculation, and vice versa. One benefit of this approach is that because it only matters how effect alleles are painted relative to alternate alleles for an ancestry group, and differences in genome-wide painting averages between ancestries will not cause bias.

To calculate an ancestry-specific PRS we used an additive model, including a transformation as in Berg & Coop [9] and in line with (Supplementary Note S4c). We derived standard deviations for each score by running a block bootstrap (1000 iterations) on (1) loci and (2) individuals. We calculated polygenic risk scores for 39 traits shown to be significantly over-dispersed across ancient populations beyond what would be expected under a null model of genetic drift (Supplementary Note S4c). For computational reasons, we used a random batch of 48,000 painted individuals to calculate the effect allele frequencies, which is sufficiently large to approximate the frequencies even for ancestries that are painted less.

Our calculations were limited to the 549,323 SNPs used in the painting of the UKB (Supplementary Note S3h). This is expected to reduce predictive power compared to using the full set of imputed SNPs in the UKB, but only slightly [10]. There was a ~15% decrease in the number of SNPs included per phenotype in the PRS calculation compared with the imputed data.

To test the ancestral background of a single variant, APOE4, we calculated the average painting score for each ancestry at all sites on the chromosome of haplotypes containing the effect allele. This makes it clear when there is an excess of a particular ancestry at the site of interest.

Results

Our results tell us about the ancestral contribution to modern phenotypes in the white British population (Figure S4h.1, Figure S4h.2), and we stress we are not making claims about the phenotypes of ancient populations.

We find that Yamnaya, CHG and EHG ancestral contributions (which together form a ‘steppe’ component) have relatively high scores for height, whereas Farmers and WHG ancestral contributions have relatively low scores. This accords with most previous studies [3,11,12] but not all [13]. EHG and Yamnaya both score highly for body mass and basal metabolic rate.

Hair and skin pigmentation show significant differences between the ancestral contributions, with risk scores for skin colour for the three hunter-gatherer ancestries being higher (i.e. darker) than Farmer and Steppe (as in Ju and Mathieson [14]). On the other hand, traits related to malignant neoplasms of skin show higher scores for the Farmer ancestral contribution; while Farmer and Yamnaya ancestral contributions have higher scores for blonde and light brown hair, with the hunter-gatherer ancestries showing higher scores for dark brown. CHG is the only ancestral contribution which stands out as having a high risk score for black hair.

Intriguingly, the WHG ancestral component has strikingly high scores for traits related to cholesterol, blood pressure and diabetes, both when bootstrapping individuals and loci [cf. 13]. In terms of psychiatric traits, the Farmer component scores highest for anxiety, guilty feelings, and irritability.

Our two bootstrapping methods mean slightly different things. Individuals in the UKB are related through shared genealogies, and so by bootstrapping over non-independent individuals (Figure S4h.1) we are testing the consistency of the signal within the population. From this bootstrapping exercise we can conclude whether a difference in allele frequencies in ancient populations contributed to phenotypic variation today. Unsurprisingly, with a large enough sample size most phenotypes will show differences in ancestral contributions for this, usually due to drift or founder effects. However, this goes further than just reporting risk scores for ancient populations, because we are looking directly at coalescent tracts in the British population. We can conclude that “ancestry X contributes higher genetic risk for phenotype Y in the test population”. On the other hand, because we have used independent LD blocks to select SNPs to include in the PRS calculation, the requirement for independence is met when we bootstrap with loci (Figure S4h.2). A positive result here is therefore much stronger, showing a systematic over/under-representation of an ancestry at loci affecting a given trait, beyond what is expected given the correlation among individuals. This points towards selection as an explanation.

The effect/risk allele (rs429358, n=127,760) of ApoE4 is preferentially painted as WHG/EHG, with a clear depletion of other ancestries (especially Farmer) at this locus compared to the genome-wide average (Figure S4h.3). This indicates that this allele was contributed at least in part by hunter-gatherer ancestry into modern (British) populations, above what we would expect by chance.

Discussion

The methods here directly link genetic contributions from pre-defined ancestries to complex phenotypes in modern people. For most traits, each ancestry contributed differently to the modern genetic landscape, with some conveying enhanced or reduced risk either due to drift (including population bottlenecks/founder events) or selection. Because gradients exist in these ancestries across the British Isles and further afield (Supplementary Note S3h), these differing risk scores indicate how geographically heterogeneous ancestry distributions may contribute to differing genetic risk profiles, in addition to other factors such as geography, socio-economic status etc.

A caveat for all studies involving polygenic risk calculation is that they rely on effect size estimates from an original GWAS which may be affected by population stratification in the GWAS panel, even when it has apparently been controlled for. This seems to be less of a problem in the UKB than in previous GWAS studies [15], but should be kept in mind. One benefit of our approach is that there is no requirement to export these risk scores across time and space: we are using effect sizes estimated from the modern population to calculate ancestral contributions to the same modern population.

ApoE4 is an isoform of the APOE gene, resulting from linkage disequilibrium between two SNPs, rs429358 and rs7412 [16], and associated with increased risk for metabolic, vascular and neurodegenerative diseases in adulthood [17]. It may provide some enhanced cognitive ability in children and young adults [18] and other health and immunity benefits, particularly in highly infected environments [e.g. 19]. There are several lines of evidence suggesting a link between the evolution of diet and the ApoE isoforms: ε2 and ε3 alleles are associated with lower levels of blood cholesterol [20,21], while ε4 is associated with higher levels, leading some to speculate that the derived ε3 allele is 'meat-adaptive' [22,23]. In a study of South Americans, there was a five-fold increase in the ApoE4 allele in hunter-gatherers versus horticulturalists [24], potentially because the immune benefits outweighed the advantages of low blood cholesterol [25]. Generally, ε4 prevalence is higher in indigenous foraging groups such as the Pygmies, Khoi San, Papuans and some Native Amercians, while ε3 is most frequent in populations with a long-established agricultural economy [26]. Finally, ApoE4 is implicated in higher blood vitamin D levels [27].

The ε4 variant has been shown to be ancestral in humans [28]. There is a linear increasing trend in ε4 prevalence from South to North in Europe, with Sardinians showing the lowest prevalence [29–31], while there is a more than two-fold increase in Nordic versus Mediterranean

countries [32]. Sardinians are an unusual population, having the highest level of neolithic farmer ancestry of all modern European populations [33]. In this light, differences in genome-wide ancestry proportions between northern (high WHG/EHG, low Farmer) and southern Europe (high Farmer, low WHG/EHG) (Supplementary Note S3h) may explain at least part of the differences in frequency of the ε4 variant and subsequent AD genetic risk.

Figures/tables

Figure S4h.1. Ancestry-specific polygenic risk scores with 95% confidence intervals derived from bootstrapping individuals for phenotypes shown to be significantly over-dispersed between ancient populations. Confidence intervals were calculated by re-running PRS calculation on random batches of 48,000 individuals, with replacement (1000 iterations), while keeping all other annotations intact. Here we show 2 x standard deviation error bars, expected to represent ~95%

confidence interval under a normal distribution. Bootstrapping individuals tests the extent to which ancestry X contributed higher genetic risk for phenotype Y in a given population, either due to drift or selection.

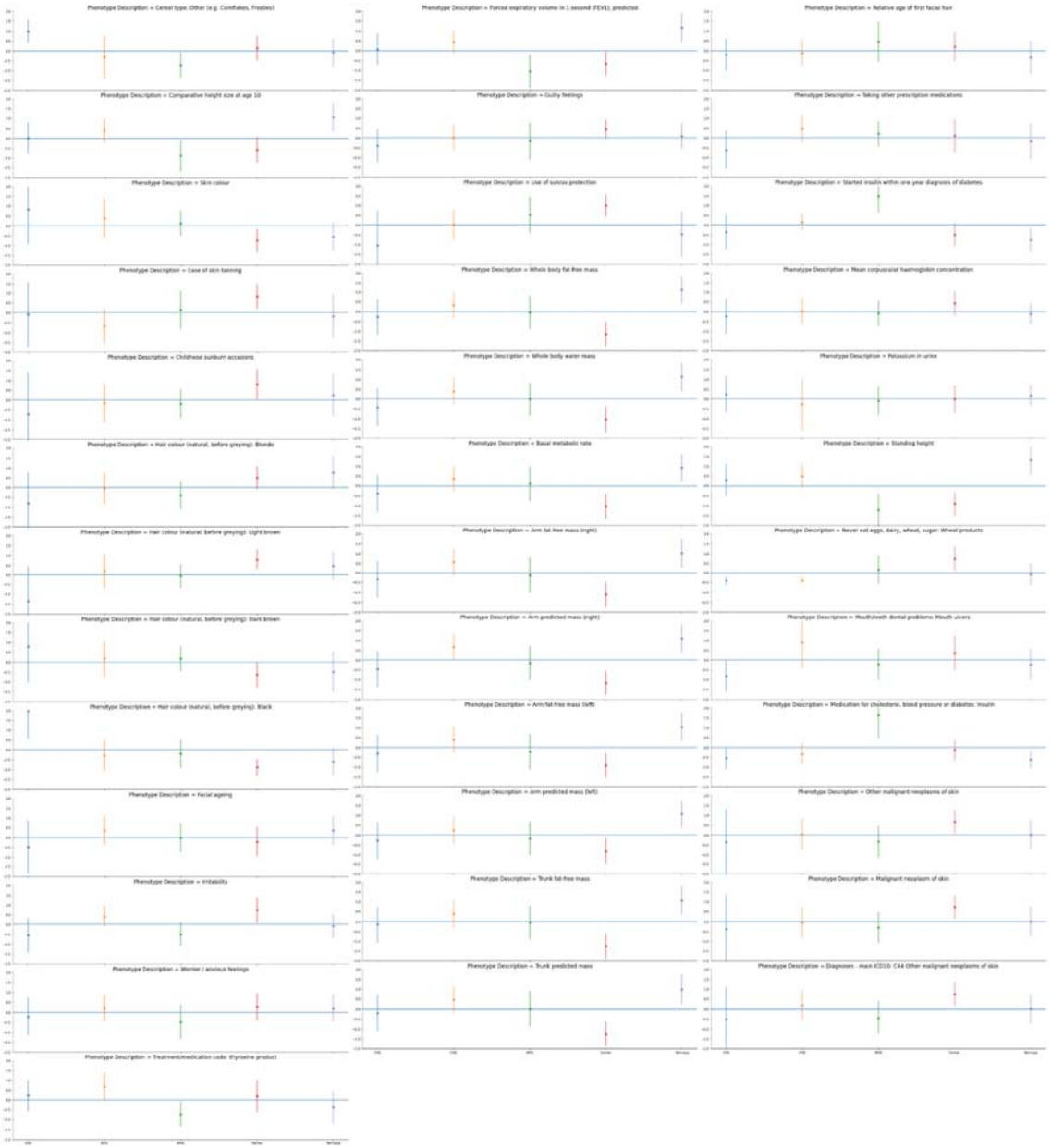


Figure S4h.2. Ancestry-specific polygenic risk scores with 95% confidence intervals derived from bootstrapping loci for phenotypes shown to be significantly over-dispersed between ancient populations. Confidence intervals were calculated by bootstrapping independent loci from separate LD blocks (1000 iterations), while keeping all other annotations intact. Here we show 2 x standard deviation error bars, expected to represent ~95% confidence interval under a normal distribution. Bootstrapping loci tests whether there is a systematic bias towards an ancestry for a given phenotype across all significant SNPs, possibly indicating selection.

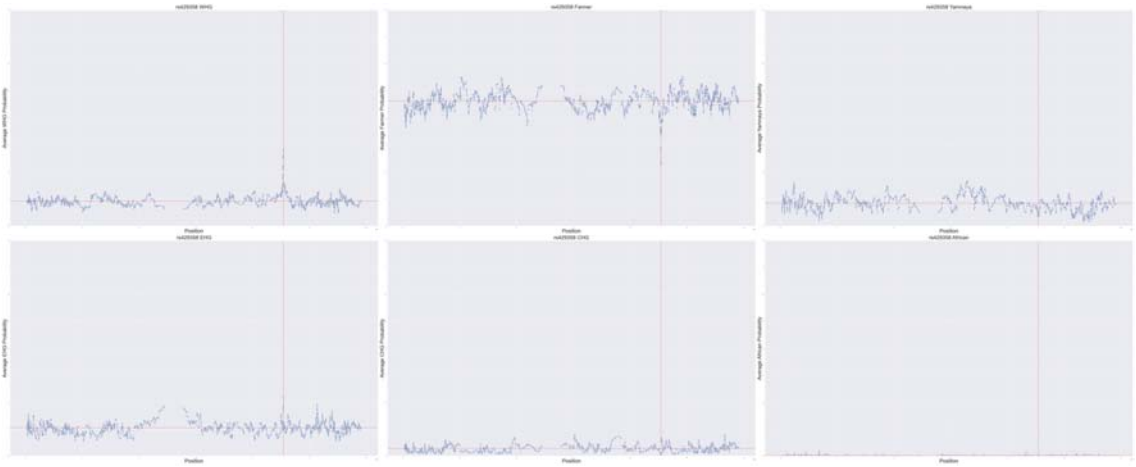


Figure S4h.3. Average painting score for each ancestry at all sites on chromosome 19 of haplotypes containing the effect allele for ApoE4 (rs429358, n=127,760). Vertical red line indicates the position of the SNP of interest; horizontal red line indicates the average painting score for that ancestry for haplotypes containing the effect allele across the entirety of chromosome 19. There is a clear excess of WHG/EHG ancestry and a depletion of Farmer ancestry at this locus.

References

1. Irving-Pease, E. K., Muktupavela, R., Dannemann, M. & Racimo, F. Quantitative Human Paleogenetics: What can Ancient DNA Tell us About Complex Trait Evolution? *Front. Genet.* **12**, 703541 (2021).

2. Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. Commun.* **10**, 3328 (2019).

3. Cox, S. L., Ruff, C. B., Maier, R. M. & Mathieson, I. Genetic contributions to variation in human stature in prehistoric Europe. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 21484–21492 (2019).

4. Cox, S. L. *et al.* Predicting skeletal stature using ancient DNA. *American Journal of Biological Anthropology* **177**, 162–174 (2022).

5. Corder, E. H. *et al.* Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261**, 921–923 (1993).

6. Strittmatter, W. J. *et al.* Binding of human apolipoprotein E to synthetic amyloid beta peptide: isoform-specific effects and implications for late-onset Alzheimer disease. *Proc. Natl. Acad. Sci. U. S. A.* **90**, 8098–8102 (1993).

7. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

8. Berisa, T. & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32**, 283–285 (2016).

9. Berg, J. J. & Coop, G. A population genetic signal of polygenic adaptation. *PLoS Genet.* **10**, e1004412 (2014).

10. Choi, S. W. & O'Reilly, P. F. PRSice-2: Polygenic Risk Score software for biobank-scale data. *Gigascience* **8**, (2019).

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

12. Martiniano, R. *et al.* The population genomics of archaeological transition in west Iberia: Investigation of ancient substructure using imputation and haplotype-based methods. *PLoS Genet.* **13**, e1006852 (2017).

13. Marnetto, D. *et al.* Ancestral contributions to contemporary European complex traits. *bioRxiv* 2021.08.03.454888 (2021) doi:10.1101/2021.08.03.454888.

14. Ju, D. & Mathieson, I. The evolution of skin pigmentation-associated variation in West Eurasia. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).

15. Berg, J. J. *et al.* Reduced signal for polygenic adaptation of height in UK Biobank. *Elife* **8**, (2019).

16. Rall, S. C., Jr, Weisgraber, K. H. & Mahley, R. W. Human apolipoprotein E. The complete amino acid sequence. *J. Biol. Chem.* **257**, 4171–4178 (1982).

17. de-Almada, B. V. P. *et al.* Protective effect of the APOE-e3 allele in Alzheimer's disease. *Braz. J. Med. Biol. Res.* **45**, 8–12 (2012).

18. Tuminello, E. R. & Han, S. D. The apolipoprotein e antagonistic pleiotropy hypothesis: review and recommendations. *Int. J. Alzheimers. Dis.* **2011**, 726197 (2011).

19. Oriá, R. B., Patrick, P. D., Blackman, J. A., Lima, A. A. M. & Guerrant, R. L. Role of apolipoprotein E4 in protecting children against early childhood diarrhea outcomes and implications for later development. *Med. Hypotheses* **68**, 1099–1107

(2007).

20. Petkeviciene, J. *et al.* Associations between apolipoprotein E genotype, diet, body mass index, and serum lipids in Lithuanian adult population. *PLoS One* **7**, e41525 (2012).

21. Carvalho-Wells, A. L. *et al.* Interactions between age and apoE genotype on fasting and postprandial triglycerides levels. *Atherosclerosis* **212**, 481–487 (2010).

22. Finch, C. E. & Stanford, C. B. Meat-adaptive genes and the evolution of slower aging in humans. *Q. Rev. Biol.* **79**, 3–50 (2004).

23. Allen, J. S., Bruss, J. & Damasio, H. The aging brain: the cognitive reserve hypothesis and hominid evolution. *Am. J. Hum. Biol.* **17**, 673–689 (2005).

24. Reales, G. *et al.* A tale of agriculturalists and hunter-gatherers: Exploring the thrifty genotype hypothesis in native South Americans. *Am. J. Phys. Anthropol.* **163**, 591–601 (2017).

25. Trumble, B. C. *et al.* Apolipoprotein E4 is associated with improved cognitive function in Amazonian forager-horticulturalists with a high parasite burden. *FASEB J.* **31**, 1508–1515 (2017).

26. Corbo, R. M. & Scacchi, R. Apolipoprotein E (APOE) allele distribution in the world. Is APOE*4 a 'thrifty' allele? *Ann. Hum. Genet.* **63**, 301–310 (1999).

27. Huebbe, P. *et al.* APOE ε4 is associated with higher vitamin D levels in targeted replacement mice and humans. *FASEB J.* **25**, 3262–3270 (2011).

28. Fullerton, S. M. *et al.* Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. *Am. J. Hum. Genet.* **67**, 881–900 (2000).

29. Corbo, R. M., Scacchi, R., Mureddu, L., Mulas, G. & Alfano, G. Apolipoprotein E polymorphism in Italy investigated in native plasma by a simple polyacrylamide gel isoelectric focusing technique. Comparison with frequency data of other European populations. *Ann. Hum. Genet.* **59**, 197–209 (1995).

30. Lucotte, G., Loirat, F. & Hazout, S. Pattern of gradient of apolipoprotein E

allele *4 frequencies in western Europe. *Hum. Biol.* **69**, 253–262 (1997).

31. Adler, G. *et al.* Bosnian study of APOE distribution (BOSAD): a comparison with other European populations. *Ann. Hum. Biol.* **44**, 568–573 (2017).

32. Trumble, B. C. & Finch, C. E. THE EXPOSOME IN HUMAN EVOLUTION: FROM DUST TO DIESEL. *Q. Rev. Biol.* **94**, 333–394 (2019).

33. Chiang, C. W. K. *et al.* Genomic history of the Sardinian population. *Nat. Genet.* **50**, 1426–1434 (2018).

4i) Pathogenic structural variants in ancient vs. modern-day humans

Alma S. Halgren[1], Andrés Ingason[2,3], and Peter H. Sudmant[1]

[1]Department of Integrative Biology, University of California, Berkeley
[2]Institute of Biological Psychiatry, Mental Health Services, Copenhagen University Hospital
[3]Lundbeck Foundation GeoGenetics Centre, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark

Rare, recurrent copy-number variants (CNVs) are known to cause neurodevelopmental disorders and are associated with a range of psychiatric and physical traits with variable expressivity and incomplete penetrance [1,2]. We examined 50 regions susceptible to recurrent CNV known to be the most prevalent drivers of human developmental pathologies [3] in 1442 ancient Eurasians and 1093 modern human populations (for comparison) to understand the prevalence of pathogenic structural variants over time.

This analysis examines 1442 ancient humans from primarily West Eurasia and Central Asia as well as 1093 publicly-available high-coverage modern human genomes encompassing 136 populations worldwide (from the Human Genome Diversity Project [4] and the Simons Genome Diversity Project [5]). In modern humans and 690 ancient individuals with fastq files available, paired-end Illumina reads were mapped to the human reference genome GRCh38 with BWA-MEM [6]. In the remaining 984 samples, only BAM files that had been mapped to hg19 were available. Of note, in these 984 samples, the filtering of duplicate reads resulted in the absence of signal over segmental duplications. Nonetheless, we were able to

characterise structural variants intersecting unique sequences in these samples. The large putatively pathogenic loci which we focused on in this analysis generally consist of unique sequences flanked by segmental duplications. 232 samples were removed due to low coverage and genotype yield out of 1674 total ancient samples, leaving 1442 samples for the final analysis (601 from the fastq hg38 dataset and 841 from the BAM hg19 dataset). In all samples, average read depth in 1kb sliding genomic windows was extracted from the subsequent BAM files with pysamstats [7]. All alternate haplotypes were removed prior to mapping. To approximate copy number from read depth, we masked tandem repeats (Tandem Repeat Finder [8]), corrected read-depth estimates for underlying GC content (similar method to Sudmant et al. 2013 [9]), and normalised by median read depth per individual. We implemented a Gaussian Hidden Markov Model [10] to call structural variants from read depth.

We identified CNVs in ancient individuals at ten loci using digital Comparative Genomic Hybridization [11] (Table S4i.1; Figures S4i.1-S4i.20). Although most of the observed CNVs (including duplications at 15q11.2 and *CHRNA7,* and CNVs spanning parts of the TAR locus and 22q11.2 distal) have not been unambiguously associated with disease in large studies, the identified CNVs include deletions and duplications that have been associated with developmental delay, dysmorphic features, and neuropsychiatric abnormalities such as autism (most notably at 1q21.1, 3q29, 16p12.1 and the DiGeorge/VCFS locus, but also deletions at 15q11.2 and duplications at 16p13.11). However, phenotypes and risk associated with these structural variants vary widely, and recent population-based studies [12,13] suggest that they may be more common in the general population than previously thought [1,2]. Overall, the carrier frequency in ancient samples is similar to that reported in the UK Biobank (1.25% vs 1.6% at 15q11.2 and *CHRNA7* combined, and 0.8% vs 1.1% across the remaining loci combined) [12]. These results suggest that large, recurrent CNVs that can lead to several pathologies were present in ancient populations at similar frequencies as modern populations.

Region	Ancient Deletions	SGDP & HGDP Deletions	UK Biobank Deletions	Ancient Duplications	SGDP & HGDP Duplications	UK Biobank Duplications
1q21.1	0 (0%)	0 (0%)	113 (0.027%)	1 (0.069%)	0 (0%)	177 (0.042%)
3q29	1 (0.069%)	0 (0%)	9 (0.002%)	1 (0.069%)	0 (0%)	5 (0.001%)
15q11.2	4 (0.28%)	2 (0.18%)	1664 (0.39%)	10 (0.69%)	9 (0.82%)	2041 (0.48%)
15q11q13 (BP3-BP4)	1 (0.069%)	0 (0%)	16 (0.004%)	0 (0%)	0 (0%)	53 (0.013%)
15q13.3 (*CHRNA7*)	0 (0%)	1 (0.09%)	10 (0.002%)	4 (0.28%)	8 (0.73%)	3031 (0.72%)
16p12.1	1 (0.069%)	0 (0%)	246 (0.058%)	1 (0.069%)	0 (0%)	202 (0.048%)
16p13.11	1 (0.069%)	0 (0%)	131 (0.031%)	4 (0.28%)	0 (0%)	828 (0.2%)
22q11.2 (distal)*	4 (0.28%)	0 (0%)	N/A	13 (0.90%)	6 (0.55%)	N/A
DiGeorge-VCFS	0 (0%)	0 (0%)	10 (0.0024%)	1 (0.069%)	0 (0%)	280 (0.066%)
TAR*	1 (0.069%)	0 (0%)	N/A	2 (0.14%)	0 (0%)	N/A

Table S4i.1. Table of 10 pathogenic loci (out of 50 examined) with structural variants identified in the ancient dataset. For each dataset, we report the prevalence of each SV as both the number of individuals identified as well as the percentage in each population (ancient dataset: 1442 samples; modern human Simons Genome Diversity Project (SGDP) [5] and Human Genome Diversity Project (HGDP) [4] dataset: 1093 samples; UK Biobank dataset: 421268 samples).
*Ancient SVs do not span entire locus, and therefore we cannot compare the UK Biobank frequencies for these loci (which span the entirety of the locus) to the ancient frequencies.

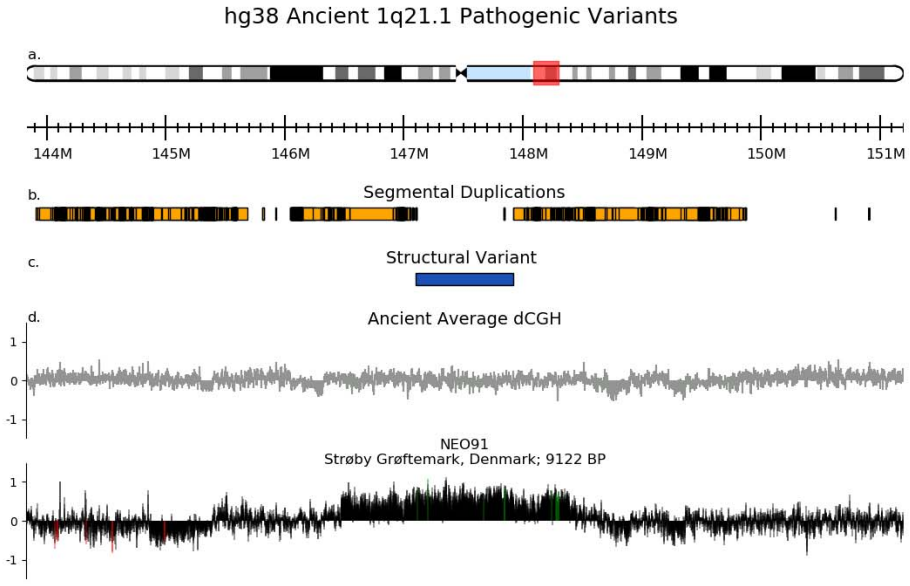


Figure S4i.1.. Microduplication of an ancient individual at the 1q21.1 locus from the hg38 fastq dataset. This 2 Mb microduplication is known to be associated with increased risk of neurological and psychiatric problems, delayed development, Tetralogy of Fallot, and micro/macrocephaly [14–16]. a. Chromosome 1 G-bands and axis bar (in Mbp). b. Segmental Duplications of >1000 bases of non-RepeatMasked [17] sequence from the UCSC Genome Browser [18]. c. Indices of the structural variant from the literature. d. The average dCGH (in

gray) is computed as follows: first, a non-noisy individual for this locus is selected; then, for every other individual, the $\log_2$ ratio of its copy number values over the non-noisy individuals' copy number values is calculated; finally, the average of these ratios is depicted. The $\log_2$ ratio of NEO91's copy number values over those of the non-noisy individual is shown below (green = the ratio is at least 1.5 the standard deviation above the average dCGH; red = 1.5 std. dev. below). While there is little deviation from the "norm" copy number at this locus on average, the dCGH of NEO91 is significantly greater than the average and indicates a duplication.

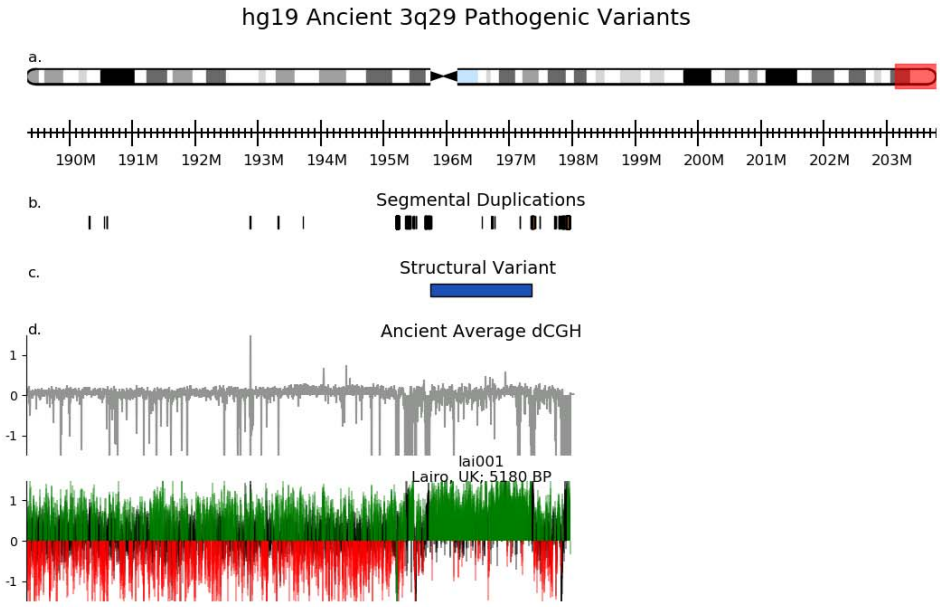


Figure S4i.2. Microduplication of an ancient individual at the 3q29 locus from the hg19 BAM dataset. At 3q29, microdeletions and microduplications are associated with speech and developmental delay, cleft palate, microcephaly, and increased risk for psychiatric disorders [19,20]. No modern individuals in the SGDP or HGDP datasets present with a microdeletion or microduplication at 3q29.

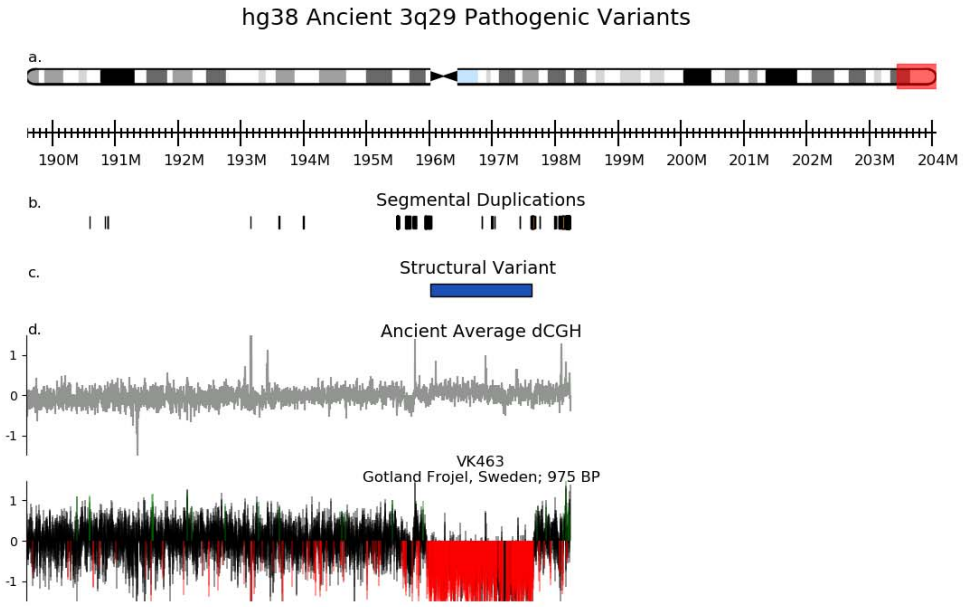


Figure S4i.3. Microdeletion of an ancient individual at the 3q29 locus from the hg38 fastq dataset.

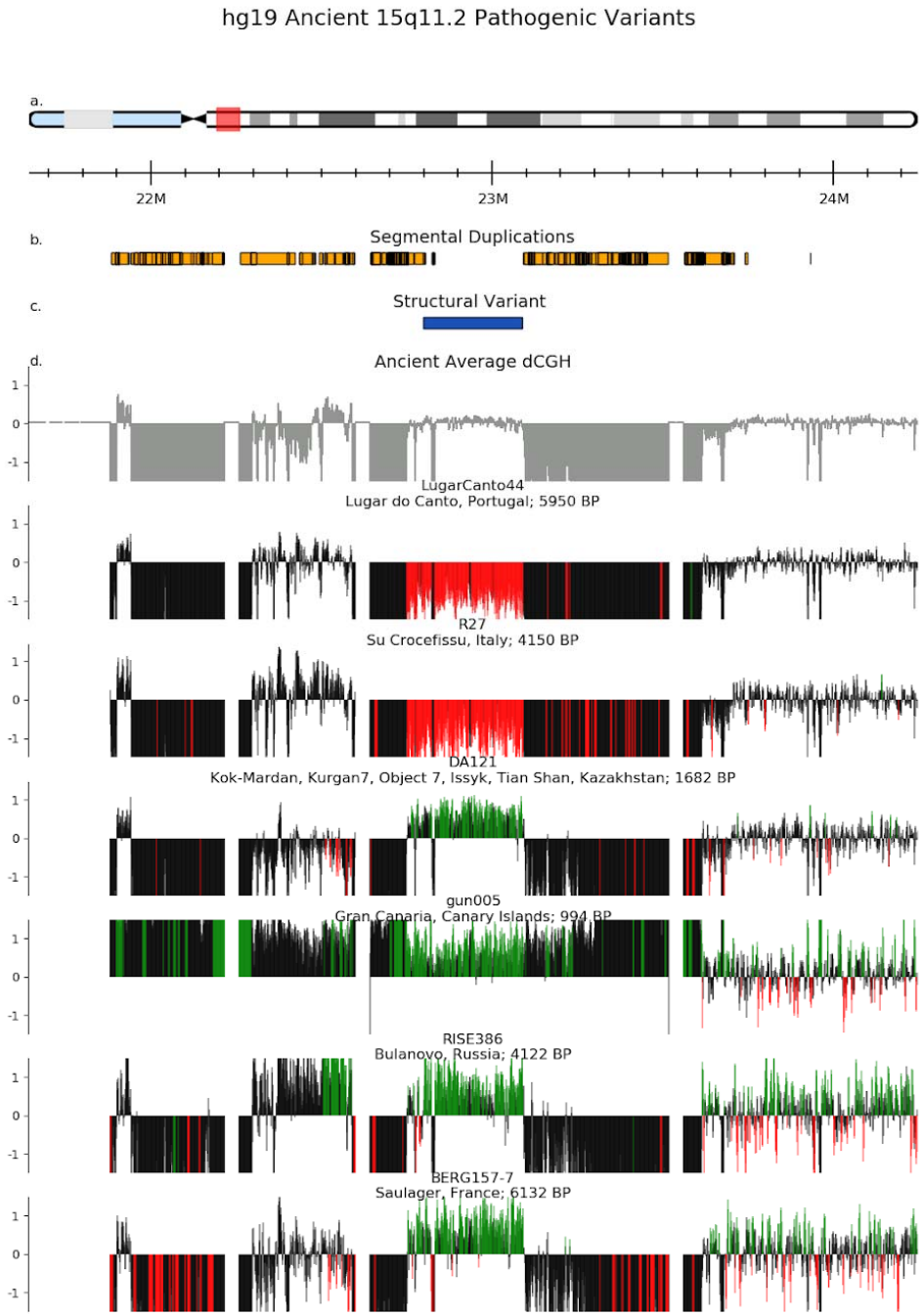


Figure S4i.4. Two microdeletions and four microduplications are present at the 15q11.2 locus from the hg19 BAM dataset. Although the microdeletion is considered to confer modest risk of schizophrenia, epilepsy, learning problems, and ADHD, the microduplication is largely considered to be benign [21–24]. The blocks flanking the structural variant with copy number 0 are where duplicated reads have been filtered from the dataset.

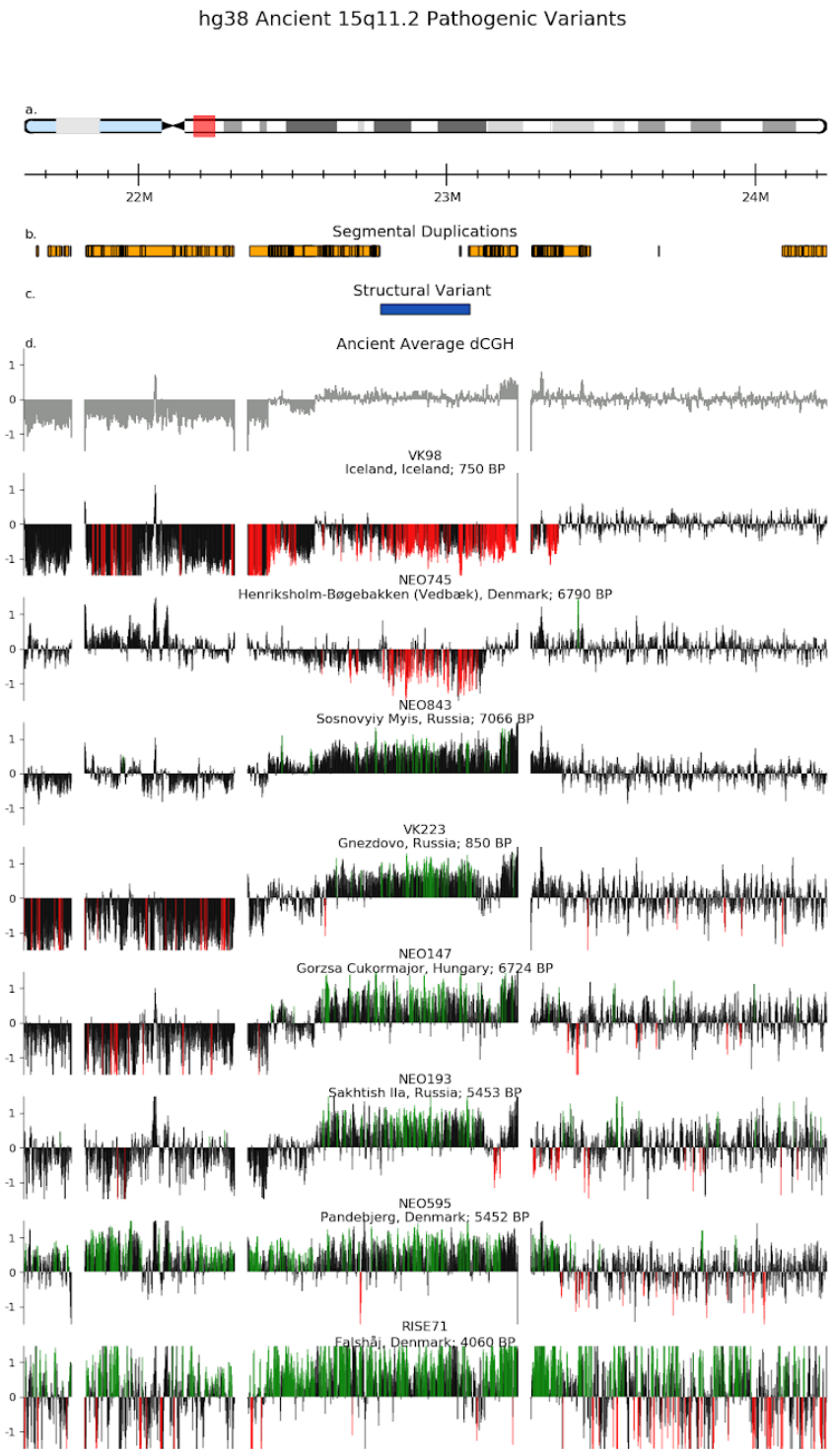


Figure S4i.5. Two microdeletions and six microduplications are present at the 15q11.2 locus from the hg38 fastq dataset.

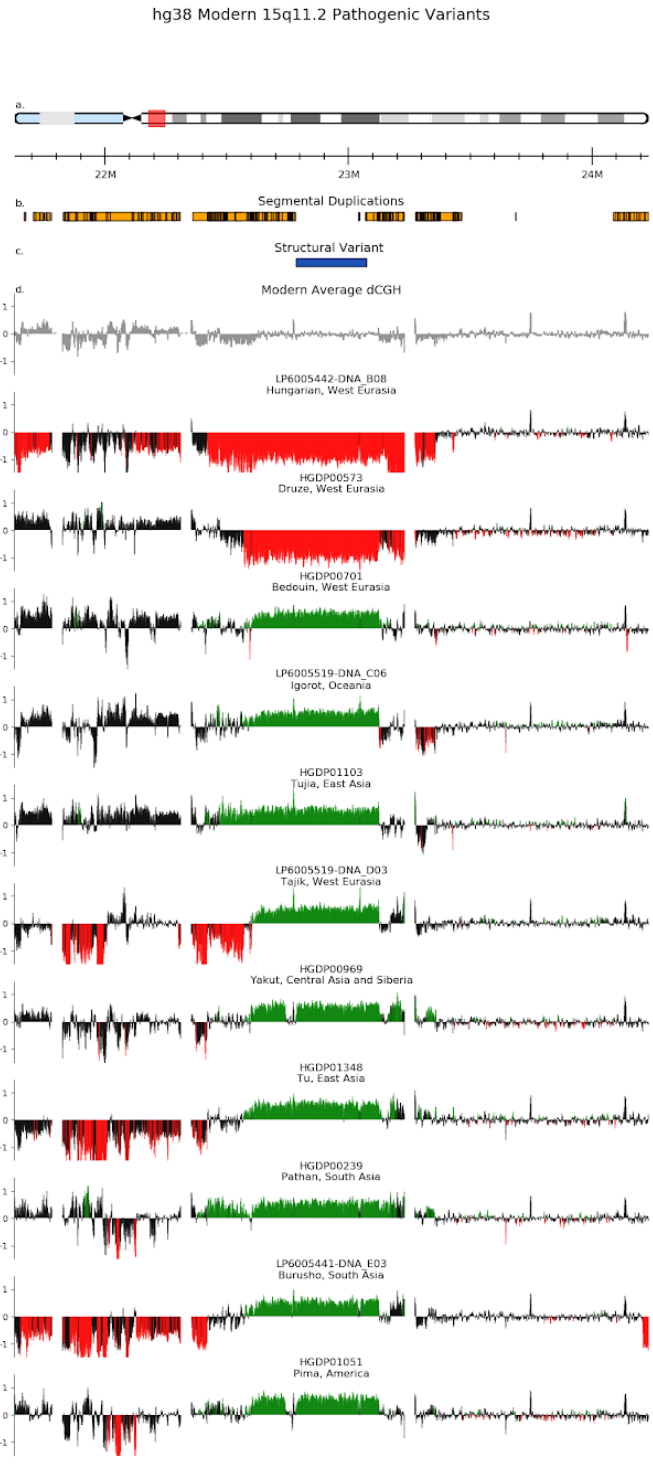


Figure S4i.6. Two microdeletions and nine microduplications are present at the 15q11.2 locus from the modern human dataset. There are perhaps two duplications present – a long (HGDP01103 and HGDP00239) and a short (the other six humans).

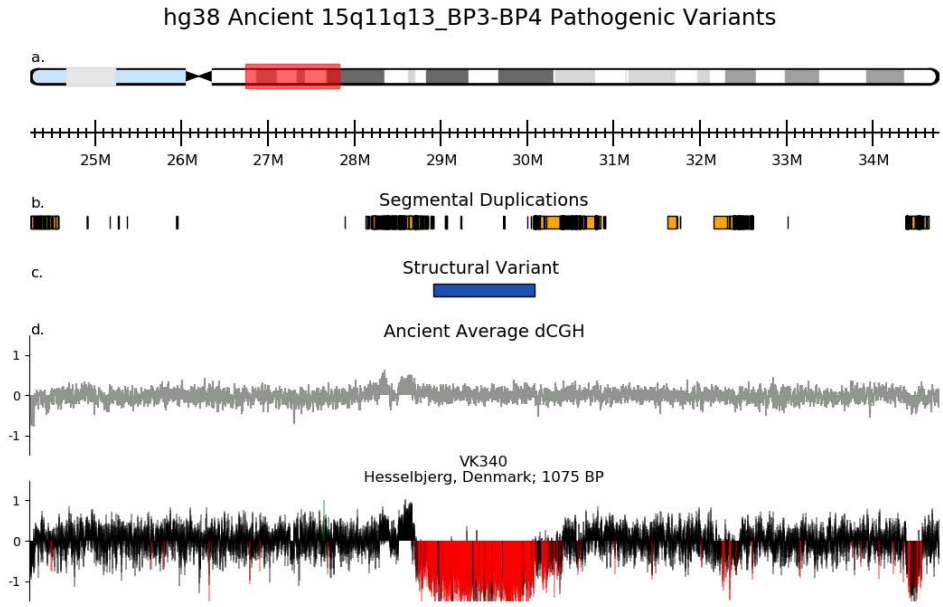


Figure S4i.7. Microdeletion of an ancient individual at the 15q11q13 (breakpoints BP3-BP4) locus from the hg38 fastq dataset. Although the BP3-BP4 microdeletion at 15q11q13 has not been formally associated with disease, carriers have been reported with short stature, microcephaly, hypotonia, and facial dysmorphia [25]. None of the studied modern individuals have the deletion.

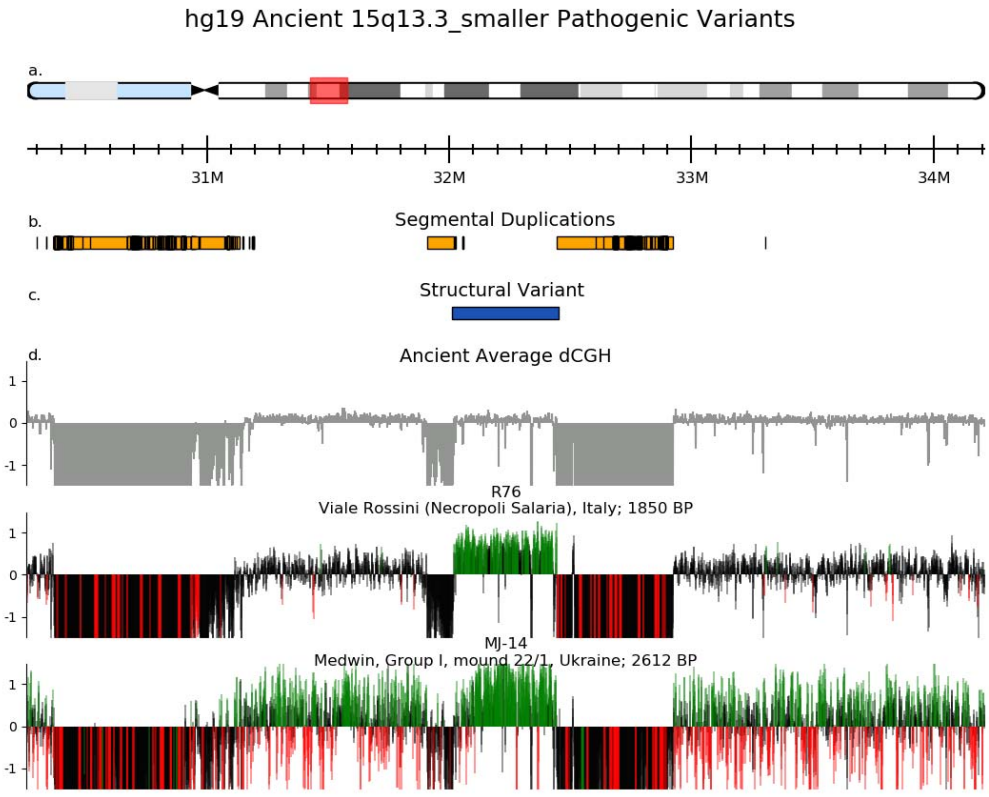


Figure S4i.8. Two microduplications are present at the 15q13.3 (*CHRNA7*) locus from the hg19 BAM dataset. The *CHRNA7* microdeletion has been associated with developmental delay and psychiatric disorders, but it is less clear which (if any) phenotypes are associated with the *CHRNA7* microduplication and there is evidence that *CHRNA7* duplication carriers have just as good cognitive function as others [26].

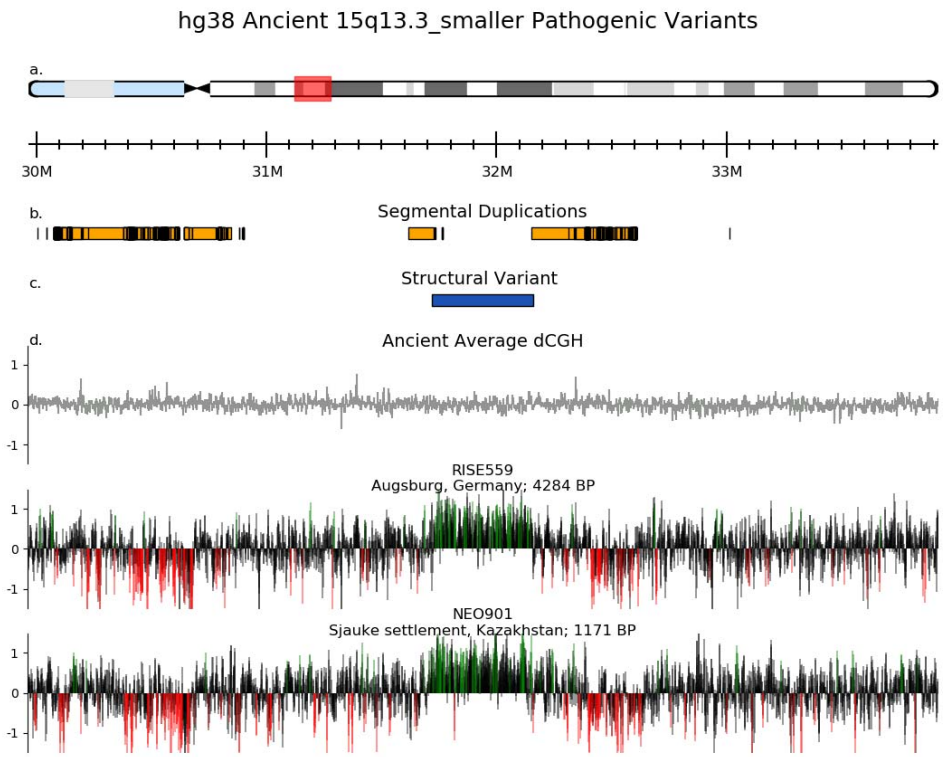


Figure S4i.9. Two microduplications are present at the 15q13.3 (*CHRNA7*) locus from the hg38 fastq dataset.

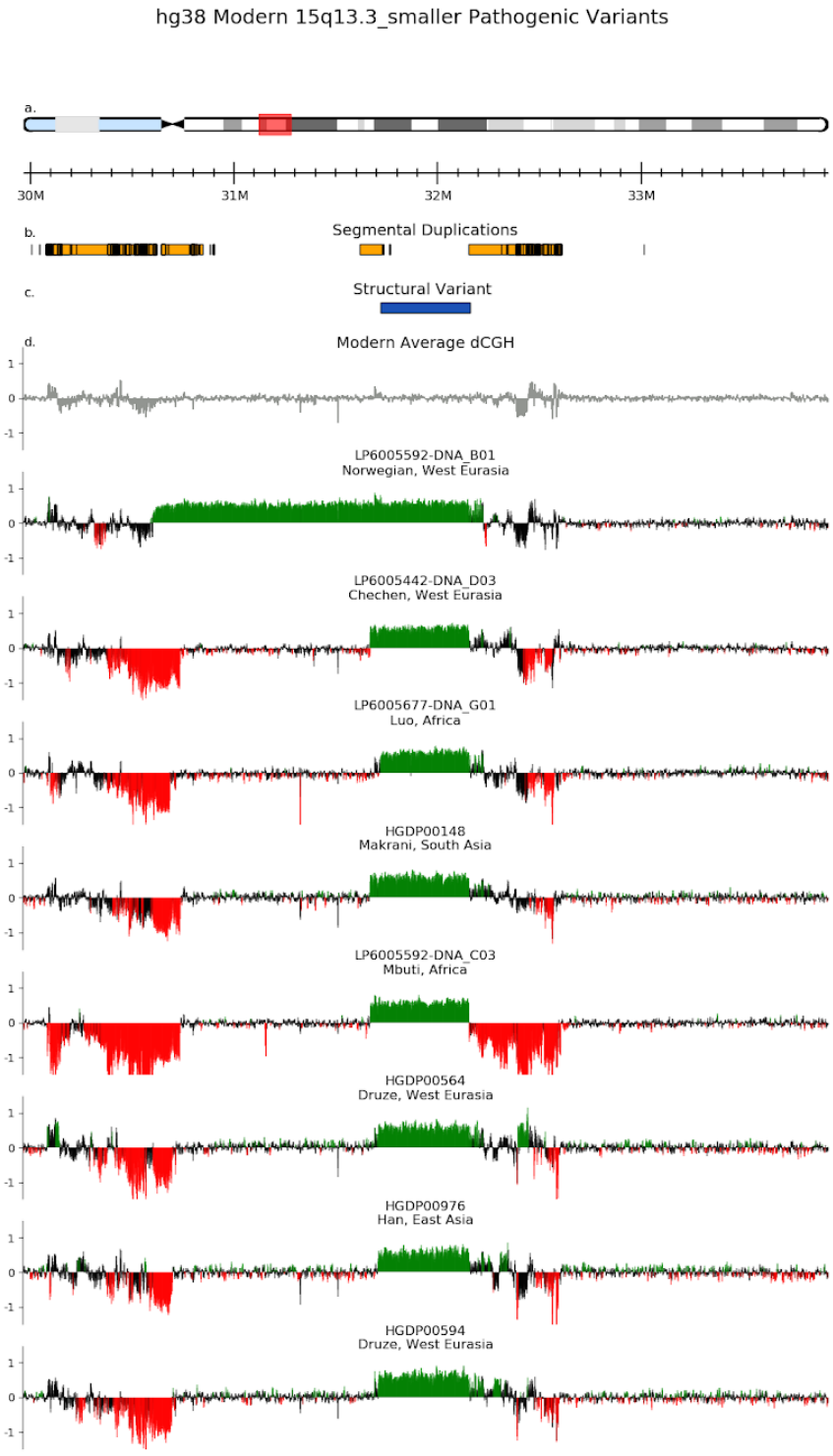


Figure S4i.10. Two different microduplications are present at the 15q13.3 (*CHRNA7*) locus from the modern human dataset.

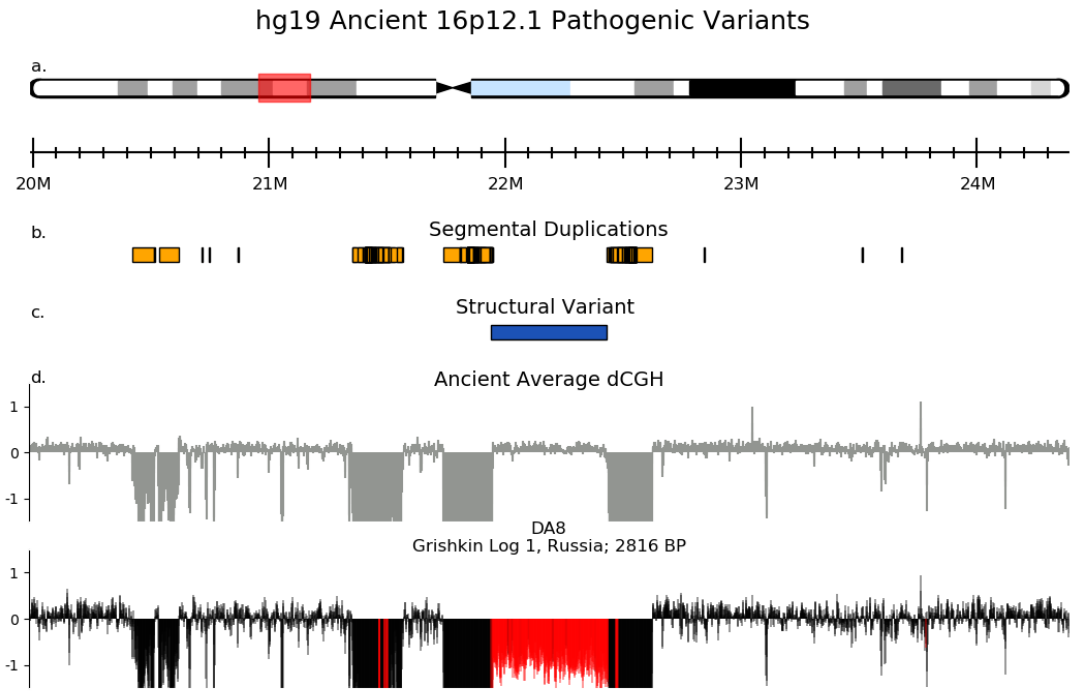


Figure S4i.11. One ancient individual from the hg19 BAM dataset has a deletion at the 16p12.1 locus. At this locus, microdeletions and microduplications are associated with developmental delay, cognitive impairment, growth impairment, cardiac malformations, epilepsy, and psychiatric and behavioural problems [27,28].

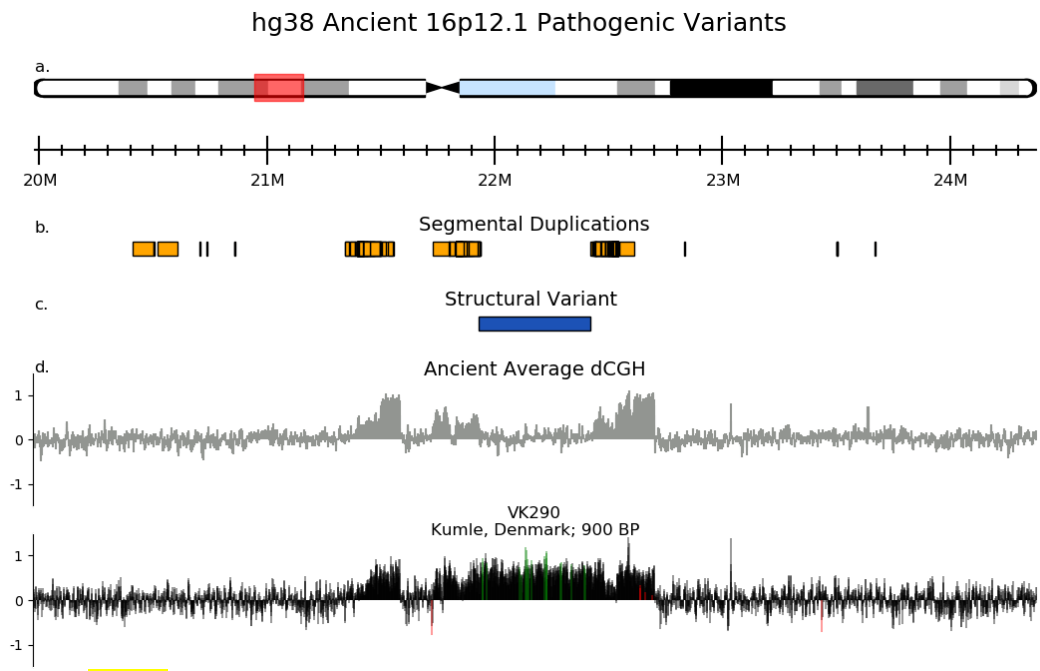


Figure S4i.12. One ancient individual from the hg38 fastq dataset has a duplication at the 16p12.1 locus.

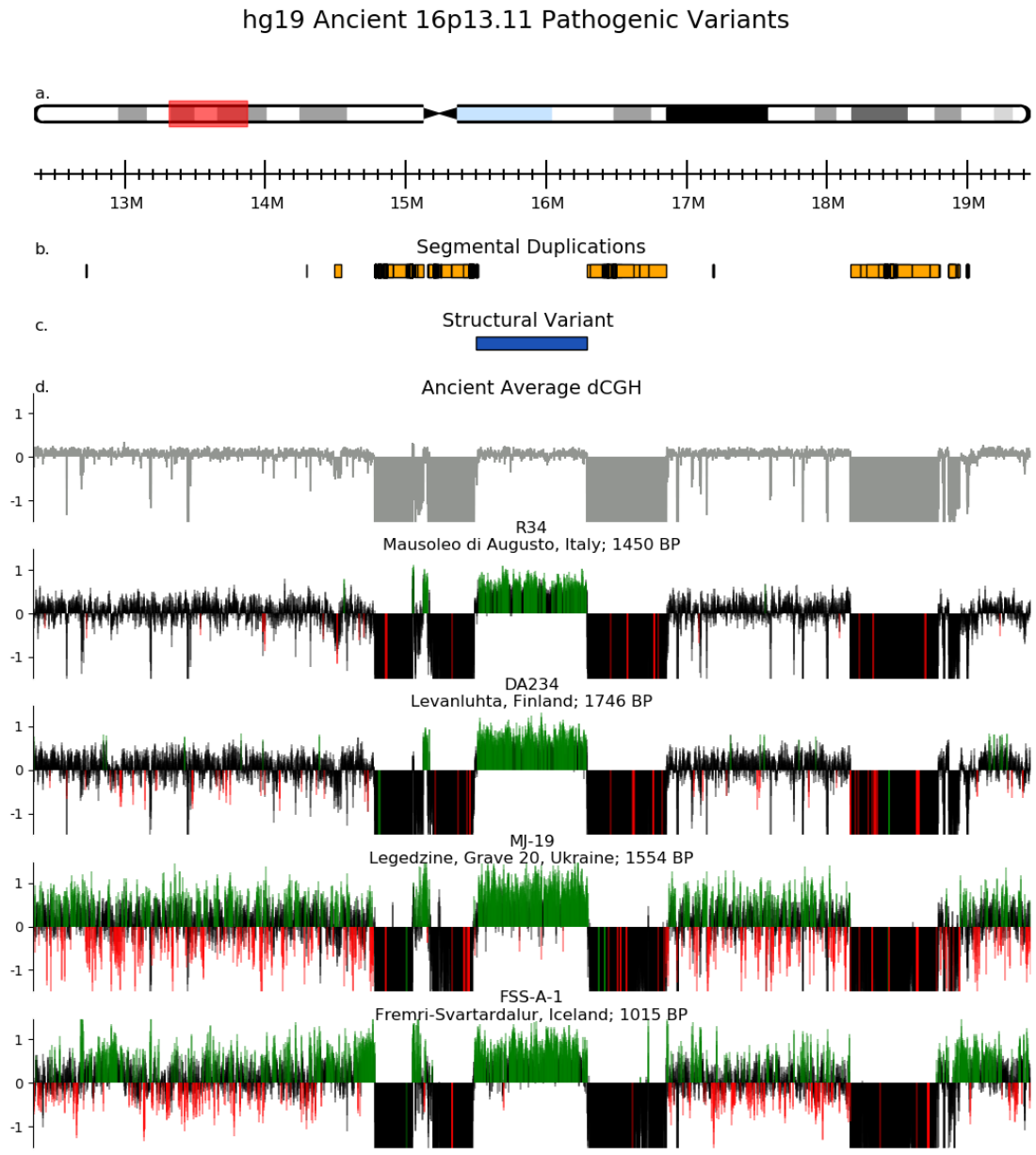


Figure S4i.13. Four microduplications are present at the 16q13.11 locus from the hg19 BAM dataset. At the 16p13.11 locus, microduplications and microdeletions have been associated with several phenotypes including behavioral abnormalities, developmental delay, congenital heart defects, and skeletal abnormalities [29,30].

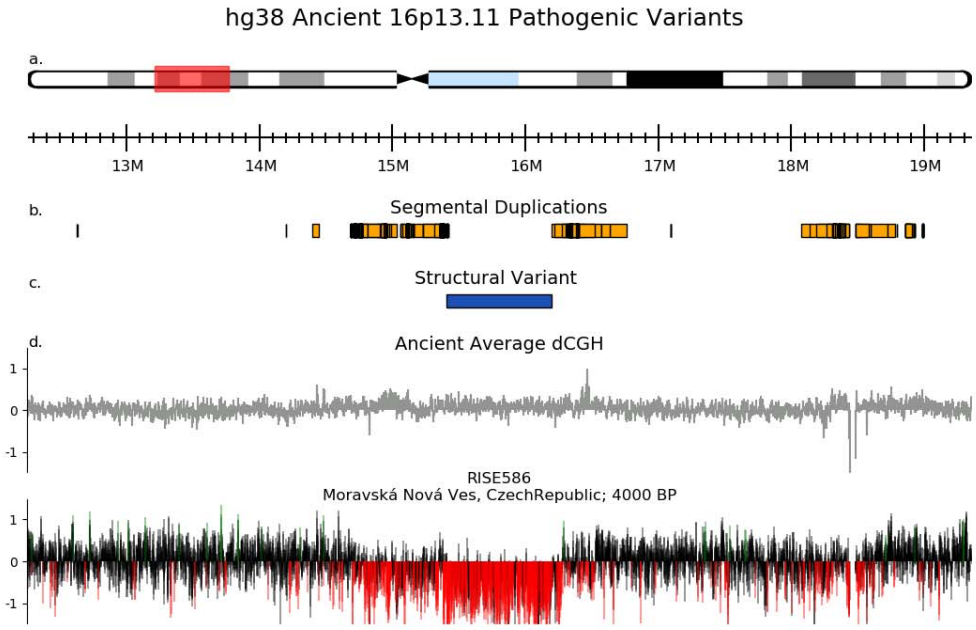
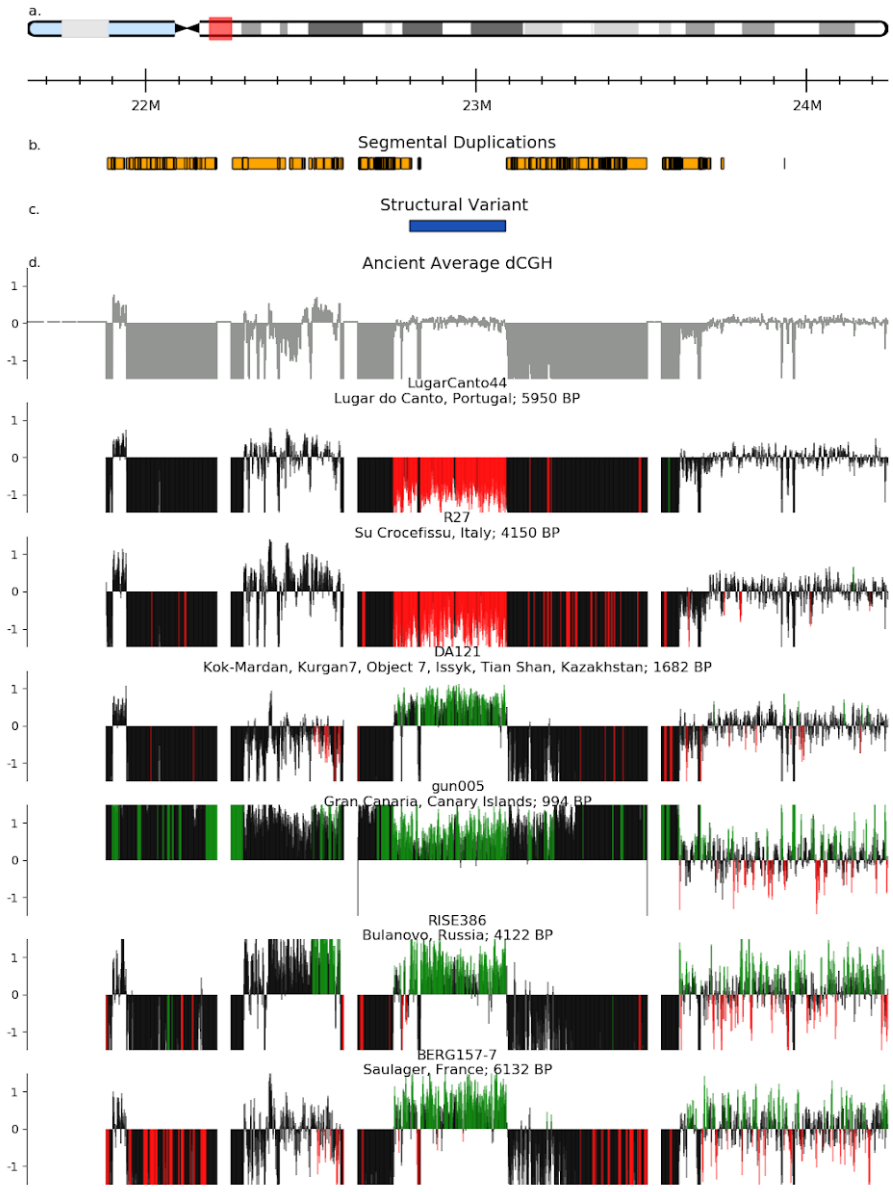


Figure S4i.14. One microdeletion is present at the 16q13.11 locus from the hg38 fastq dataset.

hg19 Ancient 15q11.2 Pathogenic Variants



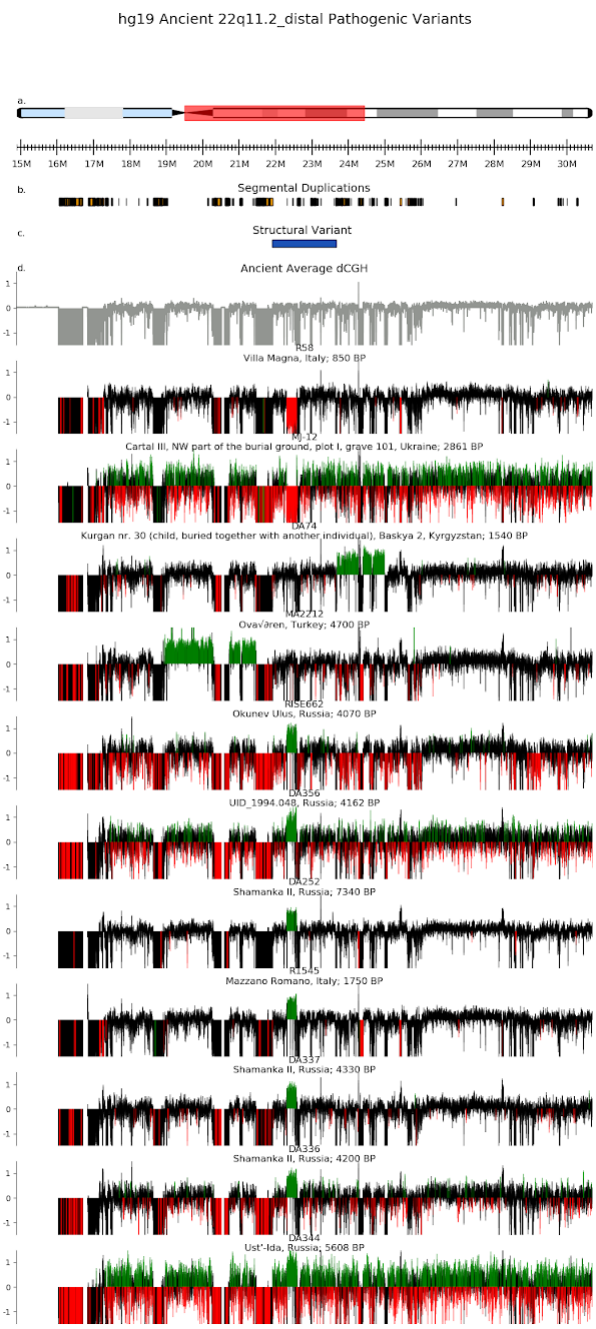


Figure S4i.15. Two microdeletions and nine microduplications are present at the 22q11.2 (distal) locus from the hg19 BAM dataset. At the 22q11.2 locus, the *TOP3B* microdeletion has been reported to be associated with autism, learning disabilities, and dysmorphic features [31,32], and *TOP3B* knockout mice exhibit behaviour similar to psychiatric disorders and cognitive impairment [33], but risk estimates from large-scale population-based studies are lacking. All hg19 ancient individuals have this *TOP3B* structural variant with identical breakpoints as a 12-year-old patient with autism, cognitive impairment, and dysmorphic features [34] (Figure S4), while two other ancient individuals have other duplication breakpoints (DA74 and MA2212).

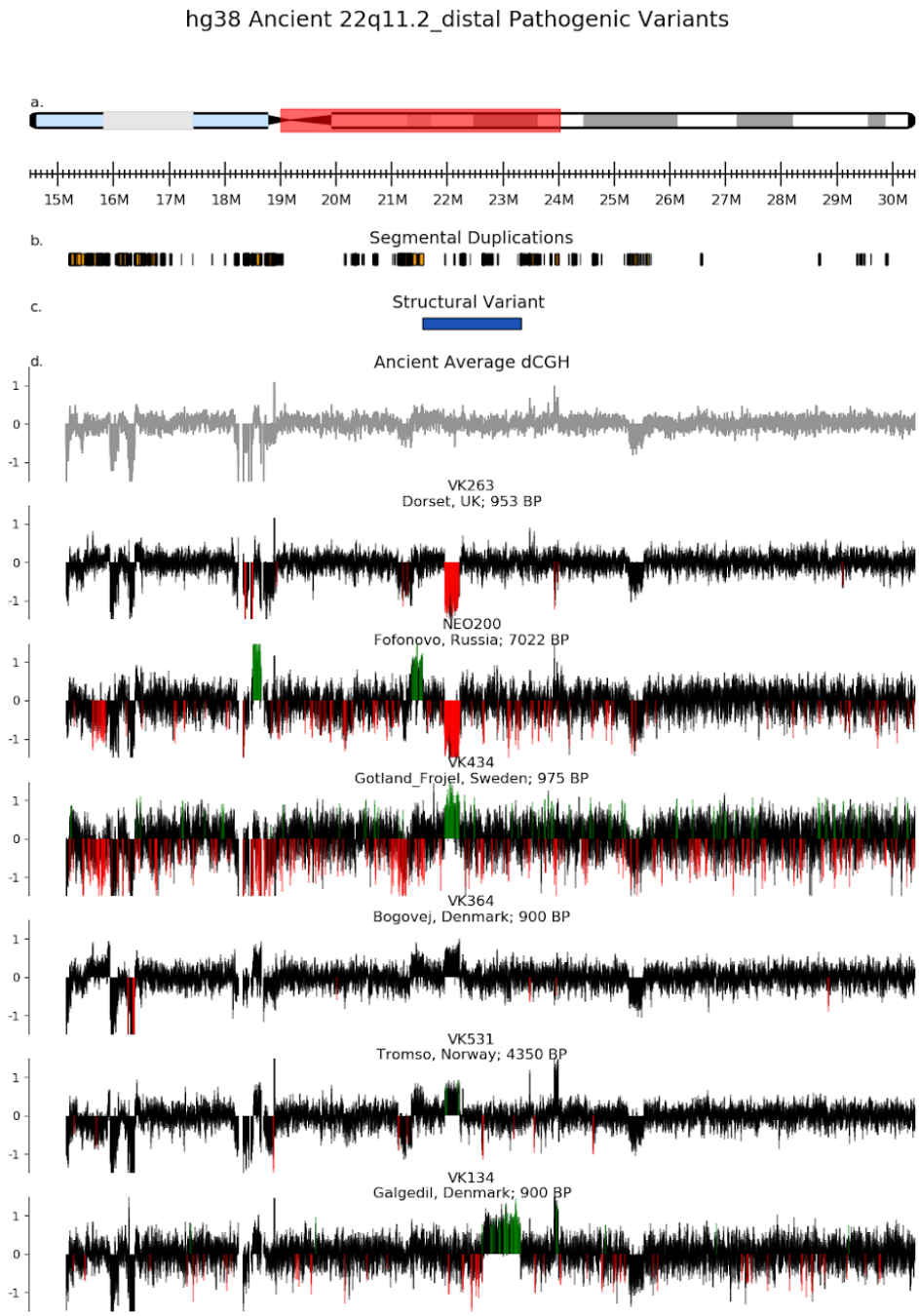


Figure S4i.16. Two microdeletions and four microduplications are present at the 22q11.2 (distal) locus from the hg19 BAM dataset.

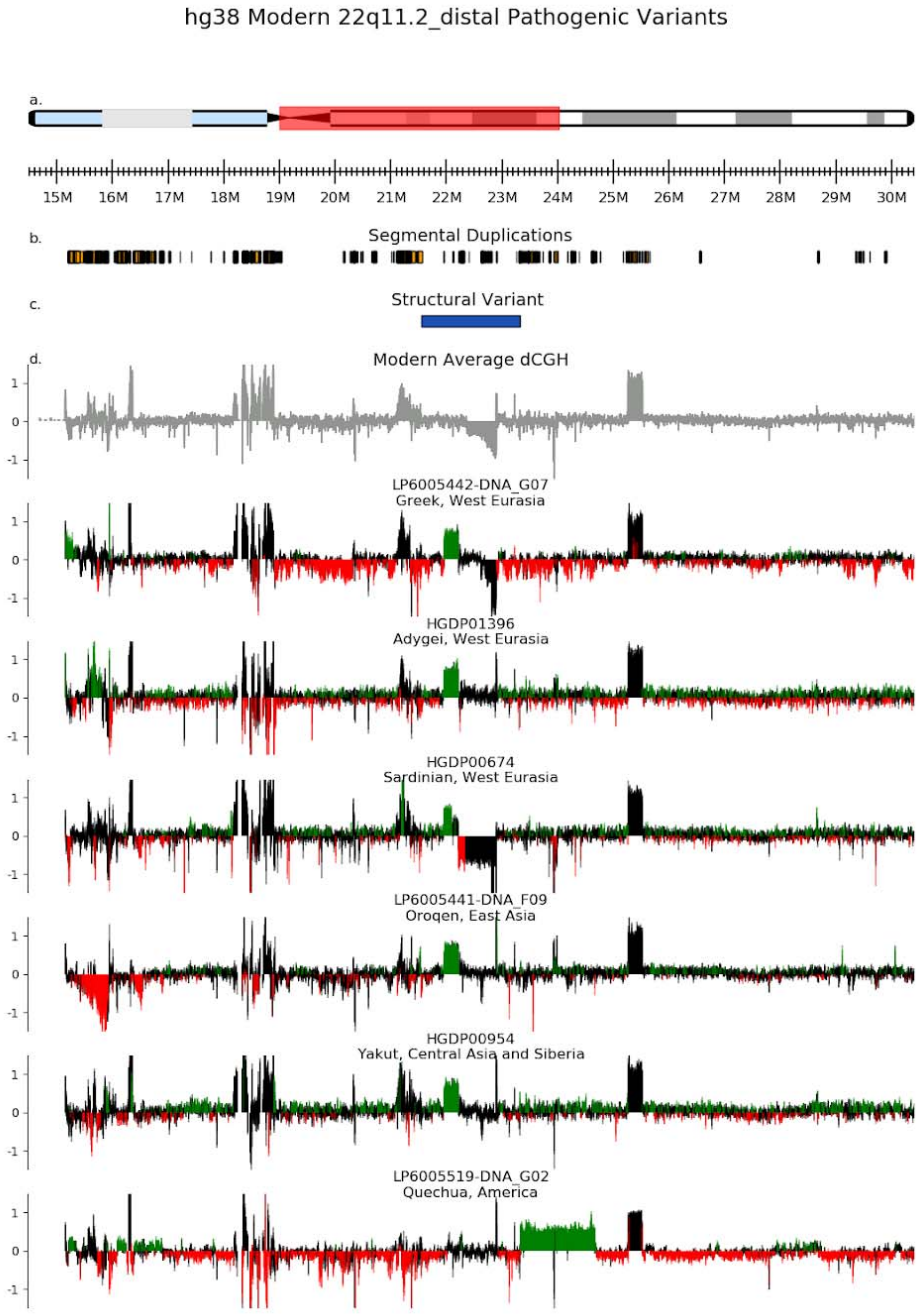
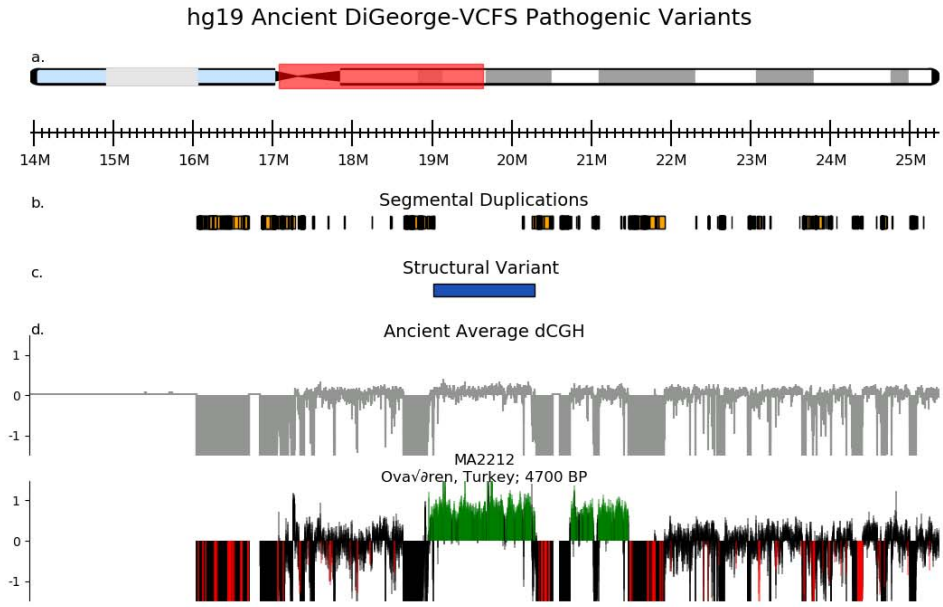


Figure S4i.17. Six microduplications (five of which directly overlap with *TOP3B*) are present at the 22q11.2 (distal) locus from the modern human dataset.



5743
5744 **Figure S4i.18. One ancient individual has a duplication at the DiGeorge-VCFS locus**
5745 **from the hg19 BAM dataset.** Duplications at this locus are associated with schizophrenia
5746 as well as cardiovascular, parathyroid, thymic, and craniofacial abnormalities [35].
5747
5748
5749
5750

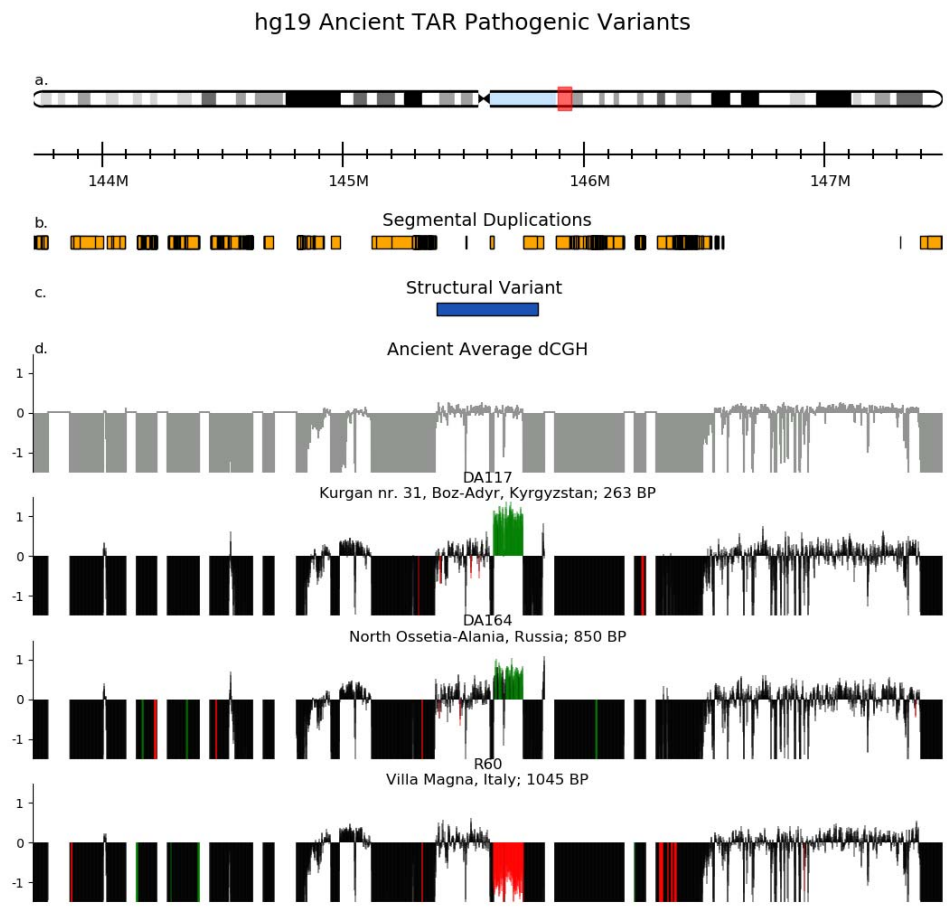


Figure S4i.19. Ancient TAR pathogenic variants. TAR (thrombocytopenia with absent radius) syndrome is a rare genetic disorder featuring the absence of the radius bone in the forearm [36]; while no ancient individuals have a structural variant across the entire breakpoint of the syndrome, one ancient individual has a deletion across part of the locus and two ancient individuals have a duplication across the same breakpoints. The possible pathological implication of these CNVs is therefore unknown.

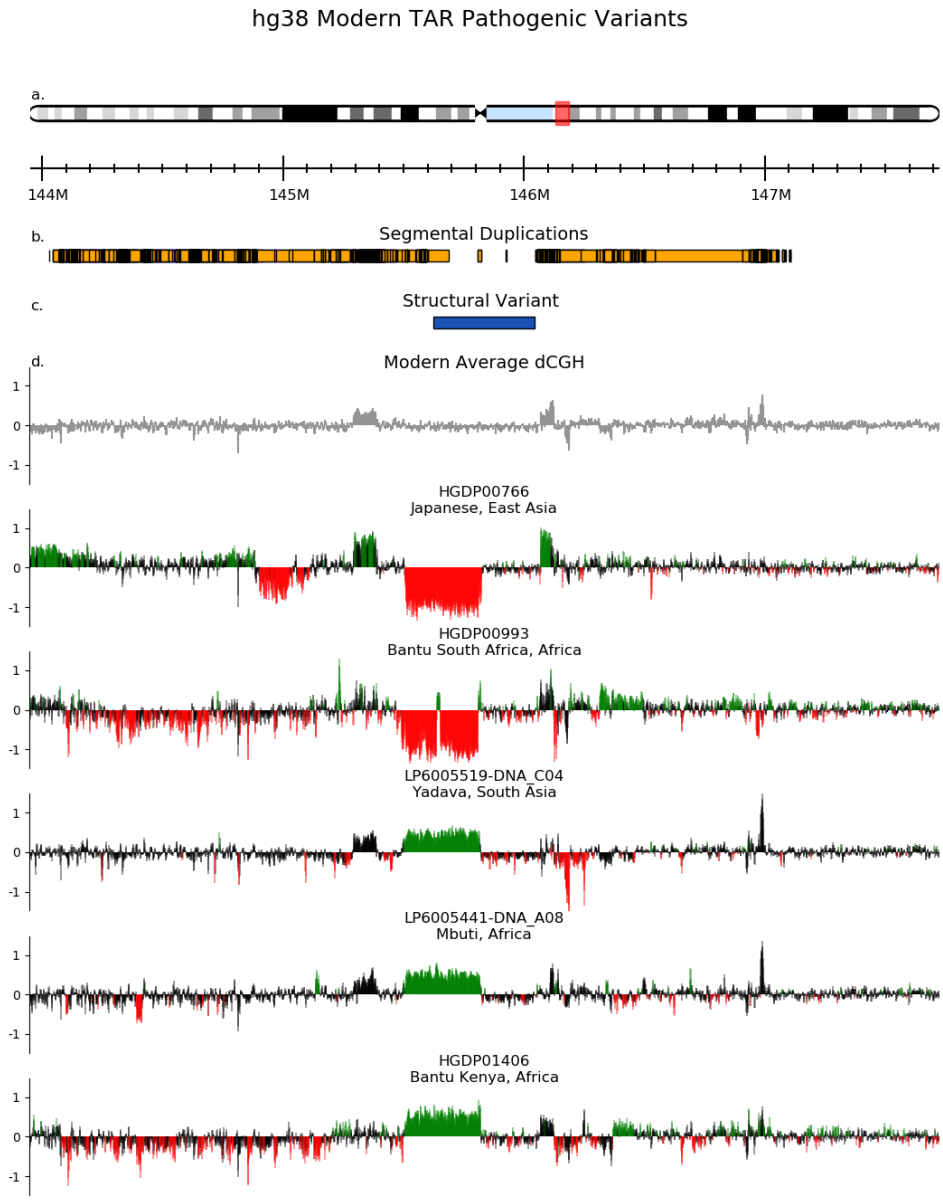


Figure S4i.20. Two deletions and three duplications are present in the modern human dataset, but with different breakpoints than the ancient human structural variants.

References

1. Girirajan, S., Campbell, C. D. & Eichler, E. E. Human copy number variation and complex genetic disease. *Annu. Rev. Genet.* **45**, 203–226 (2011).

2. Weise, A. *et al.* Microdeletion and microduplication syndromes. *J. Histochem. Cytochem.* **60**, 346–358 (2012).

3. Girirajan, S. *et al.* Phenotypic heterogeneity of genomic disorders and rare copy-

number variants. *N. Engl. J. Med.* **367**, 1321–1331 (2012).

4. Bergström, A. *et al.* Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**, (2020).

5. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).

6. Li, H. Burrows-Wheeler Aligner. http://bio-bwa.sourceforge.net/.

7. Miles, A. *pysamstats: A fast Python and command-line utility for extracting simple statistics against genome positions based on sequence alignments from a SAM or BAM file*. (Github).

8. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).

9. Sudmant, P. H. *et al.* Evolution and diversity of copy number variation in the great ape lineage. *Genome Res.* **23**, 1373–1382 (2013).

10. Leffler, E. M. *et al.* Resistance to malaria through structural variation of red blood cell invasion receptors. *Science* **356**, (2017).

11. Sudmant, P. H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330**, 641–646 (2010).

12. Crawford, K. *et al.* Medical consequences of pathogenic CNVs in adults: analysis of the UK Biobank. *J. Med. Genet.* **56**, 131–138 (2019).

13. Olsen, L. *et al.* Prevalence of rearrangements in the 22q11. 2 region and population-based risk of neuropsychiatric and developmental disorders in a Danish population: a case-cohort study. *The Lancet Psychiatry* **5**, 573–580 (2018).

14. Dolcetti, A. *et al.* 1q21.1 Microduplication expression in adults. *Genet. Med.* **15**, 282–289 (2013).

15. Mefford, H. C. *et al.* Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. *N. Engl. J. Med.* **359**, 1685–1699 (2008).

16. Brunetti-Pierri, N. *et al.* Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and

behavioral abnormalities. *Nat. Genet.* **40**, 1466–1471 (2008).

17. Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).

18. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

19. Coyan, A. G. & Dyer, L. M. 3q29 microduplication syndrome: Clinical and molecular description of eleven new cases. *Eur. J. Med. Genet.* **63**, 104083 (2020).

20. Glassford, M. R., Rosenfeld, J. A., Freedman, A. A., Zwick, M. E. & Mulle, J. G. Novel features of 3q29 deletion syndrome: Results from the 3q29 registry. *Am. J. Med. Genet. A* **170**, 999–1006 (2016).

21. Stefansson, H. *et al.* CNVs conferring risk of autism or schizophrenia affect cognition in controls. *Nature* **505**, 361–366 (2014).

22. Kirov, G. *et al.* The penetrance of copy number variations for schizophrenia and developmental delay. *Biol. Psychiatry* **75**, 378–385 (2014).

23. Gudmundsson, O. O. *et al.* Attention-deficit hyperactivity disorder shares copy number variant risk with schizophrenia and autism spectrum disorder. *Transl. Psychiatry* **9**, 258 (2019).

24. Jønch, A. E. *et al.* Estimating the effect size of the 15Q11.2 BP1–BP2 deletion and its contribution to neurodevelopmental symptoms: recommendations for practice. *J. Med. Genet.* **56**, 701–710 (2019).

25. Rosenfeld, J. A. *et al.* Deletions flanked by breakpoints 3 and 4 on 15q13 may contribute to abnormal phenotypes. *Eur. J. Hum. Genet.* **19**, 547–554 (2011).

26. Kendall, K. M. *et al.* Cognitive performance and functional outcomes of carriers of pathogenic copy number variants: analysis of the UK Biobank. *Br. J. Psychiatry* **214**, 297–304 (2019).

27. Girirajan, S., Pizzo, L., Moeschler, J. & Rosenfeld, J. *16p12.2 Recurrent Deletion*. (University of Washington, 1993).

28. D'Angelo, D. *et al.* Defining the Effect of the 16p11.2 Duplication on

Cognition, Behavior, and Medical Comorbidities. *JAMA Psychiatry* **73**, 20–30 (2016).

29. Nagamani, S. C. S. *et al.* Phenotypic manifestations of copy number variation in chromosome 16p13.11. *Eur. J. Hum. Genet.* **19**, 280–286 (2011).

30. Allach El Khattabi, L. *et al.* 16p13.11 microduplication in 45 new patients: refined clinical significance and genotype-phenotype correlations. *J. Med. Genet.* **57**, 301–307 (2020).

31. Stoll, G. *et al.* Deletion of TOP3β, a component of FMRP-containing mRNPs, contributes to neurodevelopmental disorders. *Nat. Neurosci.* **16**, 1228–1237 (2013).

32. Ahmad, M. *et al.* Topoisomerase 3β is the major topoisomerase for mRNAs and linked to neurodevelopment and mental dysfunction. *Nucleic Acids Res.* **45**, 2704–2713 (2017).

33. Joo, Y. *et al.* Topoisomerase 3β knockout mice show transcriptional and behavioural impairments associated with neurogenesis and synaptic plasticity. *Nat. Commun.* **11**, 3143 (2020).

34. Kaufman, C. S., Genovese, A. & Butler, M. G. Deletion of TOP3B Is Associated with Cognitive Impairment and Facial Dysmorphism. *Cytogenet. Genome Res.* **150**, 106–111 (2016).

35. Agatsuma, S. & Hiroi, N. Chromosome 22q11 and schizophrenia. *Nihon Shinkei Seishin Yakurigaku Zasshi* **25**, 79–84 (2005).

36. Boussion, S. *et al.* TAR syndrome: Clinical and molecular characterization of a cohort of 26 patients and description of novel noncoding variants of RBM8A. *Hum. Mutat.* **41**, 1220–1225 (2020).