Population Structure of UK Biobank and Ancient Eurasians Reveals Adaptation at Genes Influencing Blood Pressure

Kevin J. Galinsky,[1,2,*] Po-Ru Loh,[2,3] Swapan Mallick,[2,4] Nick J. Patterson,[2] and Alkes L. Price[1,2,3,*]

Analyzing genetic differences between closely related populations can be a powerful way to detect recent adaptation. The very large sample size of the UK Biobank is ideal for using population differentiation to detect selection and enables an analysis of the UK population structure at fine resolution. In this study, analyses of 113,851 UK Biobank samples showed that population structure in the UK is dominated by five principal components (PCs) spanning six clusters: Northern Ireland, Scotland, northern England, southern England, and two Welsh clusters. Analyses of ancient Eurasians revealed that populations in the northern UK have higher levels of Steppe ancestry and that UK population structure cannot be explained as a simple mixture of Celts and Saxons. A scan for unusual population differentiation along the top PCs identified a genome-wide-significant signal of selection at the coding variant rs601338 in *FUT2* ($p = 9.16 \times 10^{-9}$). In addition, by combining evidence of unusual differentiation within the UK with evidence from ancient Eurasians, we identified genome-wide-significant ($p = 5 \times 10^{-8}$) signals of recent selection at two additional loci: *CYP1A2-CSK* and *F12*. We detected strong associations between diastolic blood pressure in the UK Biobank and both the variants with selection signals at *CYP1A2-CSK* ($p = 1.10 \times 10^{-19}$) and the variants with ancient Eurasian selection signals at the *ATXN2-SH2B3* locus ($p = 8.00 \times 10^{-33}$), implicating recent adaptation related to blood pressure.

Introduction

Detecting signals of selection can provide biological insights into adaptations that have shaped human history.[1–4] Searching for genetic variants that are unusually differentiated between populations is a powerful way to detect recent selection on standing variation;[5] this approach has been applied for detecting signals of selection linked to lactase persistance[6,7] (MIM: 223100), fatty-acid decomposition[8] (MIM: 612795), hypoxia response[9–11] (MIM: 609070), malaria resistance[12–14] (MIM: 61162), and other traits and diseases.[15–18]

Leveraging population differentiation to detect selection is particularly powerful in analyses of closely related subpopulations with large sample sizes.[19] Here, we analyzed 113,851 samples of UK ancestry from the UK Biobank (see Web Resources) in conjunction with recently published datasets on People of the British Isles (PoBI)[20] and ancient DNA[21–24] to draw inferences about population structure and recent selection. We employed a recently developed selection statistic that detects unusual population differentiation along continuous principal components (PCs) instead of between discrete subpopulations[25] and combined our results with independent results from ancient Eurasians.[23] We detected three interesting signals of selection and were able to show that genetic variants at these and previously reported[23] signals of selection are strongly associated with diastolic blood pressure (DBP [MIM: 145500]) in UK Biobank samples.

Material and Methods

UK Biobank Dataset

The UK Biobank phase 1 data release contains 847,131 SNPs and 152,729 samples. We removed SNPs that were multi-allelic, had a genotyping rate less than 99%, or had a minor allele frequency (MAF) less than 1%. We also removed samples with non-British ancestry (including admixed samples) and samples with a genotyping rate less than 98%. This left 510,665 SNPs and 118,650 samples, a dataset that we call "QC*." Using PLINK2[26] (see Web Resources), we removed SNPs not in Hardy-Weinberg equilibrium ($p < 10^{-6}$), and we pruned SNPs to have a linkage disequilibrium (LD) of $r^2 < 0.2$. We then generated a genetic relationship matrix (GRM) and removed one of each pair of samples with relatedness greater than 0.05. This dataset, which we call "LD," contained 210,113 SNPs and 113,851 samples. We combined the full set of SNPs from the QC* dataset and the set of unrelated samples from the LD dataset to produce the final "QC" dataset.

PoBI and POPRES Datasets

The UK PoBI dataset consisted of 2,039 samples from the 4,371 samples collected as part of the PoBI project.[20] These 2,039 samples were a subset of 2,886 samples that had been genotyped on the Illumina Human 1.2M Duo genotyping chip, and after 2,510 passed quality-control procedures, 2,039 showed that all four grandparents had been born within 80 km of each other. This dataset allows us to examine the population genetics of the UK prior to the migrations of the late 19th and early 20th centuries. We also examined 2,988 European POPRES samples from the LOLIPOP and CoLaus collections.[27] These samples were genotyped on the Affymetrix GeneChip 500K Array. The POPRES

[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA; [2]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA; [3]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA; [4]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA
*Correspondence: galinsky@fas.harvard.edu (K.J.G.), aprice@hsph.harvard.edu (A.L.P.)
http://dx.doi.org/10.1016/j.ajhg.2016.09.014.

dataset allowed us to compare the UK Biobank population structure with that of continental Europe.

Ancient-DNA Datasets

Ancient DNA was gathered from several regions. Nine Steppe samples were collected from the Yamna oblast in Russia,[22] seven western European hunter-gatherers were collected from Loschbour,[21] 26 Neolithic farmer samples were collected from the Anatolian region,[22] and ten Saxon samples were collected from three sites in the UK.[24] DNA was extracted from bone tissue, PCR amplified, and then purified by a hybrid capture approach.[22–24] The resulting DNA was sequenced on an Illumina MiSeq, HiSeq, or NextSeq platform. Sequenced reads were aligned to the human genome (GRCh37) with the Burrows-Wheeler Aligner, and called SNPs were intersected with the SNPs found on the Human Origins Array.[28]

PC Analysis

We ran PC analysis (PCA) on the UK Biobank LD dataset by using the FastPCA software in EIGENSOFT[25] (see Web Resources). We identified several artifactual PCs dominated by regions of long-range LD (Figure S1). Removing loci with significant or suggestive selection signals (Table S1) along with their flanking 1 Mb regions from the LD dataset and rerunning PCA eliminated these artifactual PCs (Figure S2). We refer to the resulting dataset with 202,486 SNPs and 113,851 samples as the "PC" dataset. This dataset allowed us to better capture axes of variation that correspond to population structure rather than artifacts due to LD.

PC Projection

We projected PoBI[20] (642,288 SNPs and 2,039 samples from 30 populations), POPRES[27] (453,442 SNPs and 4,079 samples from 60 populations), and ancient-DNA[22,23] (159,588 SNPs and 52 samples from 4 populations) samples onto the UK Biobank PCs via PC projection.[29] The SNPs in the UK Biobank QC dataset were intersected with those in the projected dataset, and A/T and C/G SNPs were removed because of strand ambiguity (75,254, 37,593, and 24,467 SNPs for PoBI, POPRES, and ancient DNA, respectively). The intersected set of SNPs was stringently LD pruned for $r^2 < 0.05$ with PLINK2[26] (see Web Resources), leaving 27,769, 20,914, and 15,722 SNPs for PoBI, POPRES, and ancient DNA, respectively. SNP weights were computed for the intersected set of SNPs, and these weights were then used for projecting the new samples onto the UK Biobank PCs.[29]

Analysis of Population Clusters

After running PCA, we clustered individuals by *k*-means clustering into six clusters on five PCs. We labeled clusters by comparing the centroids of each cluster with the centroids of the projected PoBI populations, as well as by visual inspection. We then analyzed these clusters by running TreeMix[30,31] with default settings (see Web Resources) in order to assess the hierarchical population structure between the clusters.

Pairwise Discrete Subpopulation-Based Selection Statistic

Although we focused primarily on the PCA-based selection statistic from Galinsky et al.[25] (see below), we also applied the discrete subpopulation-based selection statistic from Bhatia et al.,[19] which we briefly review here. Suppose two populations with genetic distance F_{ST} are descended from a single ancestral population. The allele frequencies at a particular SNP in these two populations (p_1 and p_2) follow a normal distribution such that $p_1, p_2 \sim N(p_a, F_{ST}\, p_a(1-p_a))$, where p_a is the ancestral allele frequency of this SNP. As a result, the difference in allele frequency ($D = p_1 - p_2$) also follows a normal distribution with mean 0 and variance $2F_{ST}\, p_a(1-p_a)$. Thus, the sample allele-frequency difference $\widehat{D} = \widehat{p}_1 - \widehat{p}_2 \sim N(0, p_a(1-p_a)(2F_{ST} + N_1^{-1} + N_2^{-1}))$, where N_1 and N_2 are the number of observed haplotypes from each population. By estimating $\widehat{F_{ST}}$ and $\widehat{p}_a = (\widehat{p}_1 + \widehat{p}_2)/2$, we can assess the statistical significance of unusually large values of $\widehat{D}$. We applied this statistic to pairs of population clusters.

PCA-Based Selection Statistic

We applied the PCA-based selection statistic from Galinsky et al.,[25] which we briefly review here. PCA is equivalent to the singular-value decomposition ($\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$), where $\boldsymbol{X}$ is the normalized genomic matrix, $\boldsymbol{U}$ is the matrix of left singular vectors, $\boldsymbol{V}$ is the matrix of right singular vectors, and $\boldsymbol{\Sigma}$ is a diagonal matrix of singular values. The singular values are related to the eigenvalues of the GRM by the relationship $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^2/M$, where M is the number of SNPs used for computing the GRM $\boldsymbol{X}^T\boldsymbol{X}/M$. The matrix $\boldsymbol{U}$ has the properties $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$ and $\boldsymbol{U} = \boldsymbol{X}\boldsymbol{V}\boldsymbol{\Sigma}^{-1}$. By the central limit theorem, the elements of $\boldsymbol{U}$ follow a normal distribution, and after rescaling by M, they follow a chi-square (1 degree of freedom [df]) distribution. In other words, the statistic $M(\boldsymbol{X}_i\boldsymbol{V}_k)^2/\Sigma_k^2 = (\boldsymbol{X}_i\boldsymbol{V}_k)/\boldsymbol{\Lambda}_k$ for the i^{th} SNP at the k^{th} PC follows a chi-square (1 df) distribution.[25] One benefit of this statistic is that the PCs can be generated on one set of SNPs (here, we used the PC dataset described earlier in order to capture axes of variation related to true population structure), and the selection statistic can be calculated on another set of SNPs (we used the QC dataset in order to maximize the set of SNPs evaluated for signals of selection).

We clustered signals of selection by considering all SNPs for which the p value with respect to at least one PC was less than an initial threshold (which we set at 10^{-6}) and by clustering together SNPs within 1 Mb of each other. SNPs with signals on different PCs but in close proximity were clustered together because loci often have signals of selection on multiple PCs. We defined genome-wide-significant loci on the basis of clusters that contained at least one SNP with a p value smaller than the genome-wide significance threshold. Because we analyzed 5 PCs and 510,665 SNPs, the genome-wide significance threshold was $0.05/(5 \times 510{,}665) = 1.96 \times 10^{-8}$. We defined suggestive loci on the basis of clusters with at least two SNPs crossing the initial threshold (but none crossing the genome-wide significance threshold).

Combined Selection Statistic

We intersected the chi-square (4 df) ancient Eurasian selection statistics for 1,004,613 SNPs from Mathieson et al.[23] with the PC-based chi-square (1 df) UK Biobank selection statistics for 510,665 QC SNPs, producing a list of 115,066 SNPs. For each SNP and each PC, we added the ancient Eurasian selection statistics to the UK Biobank selection statistics for that PC, producing chi-square (5 df) statistics, which we corrected by using genomic control.

Association Tests

Association analyses were performed with PLINK2[26] with the top five PCs as covariates and the "–linear" or "–logistic" flags.

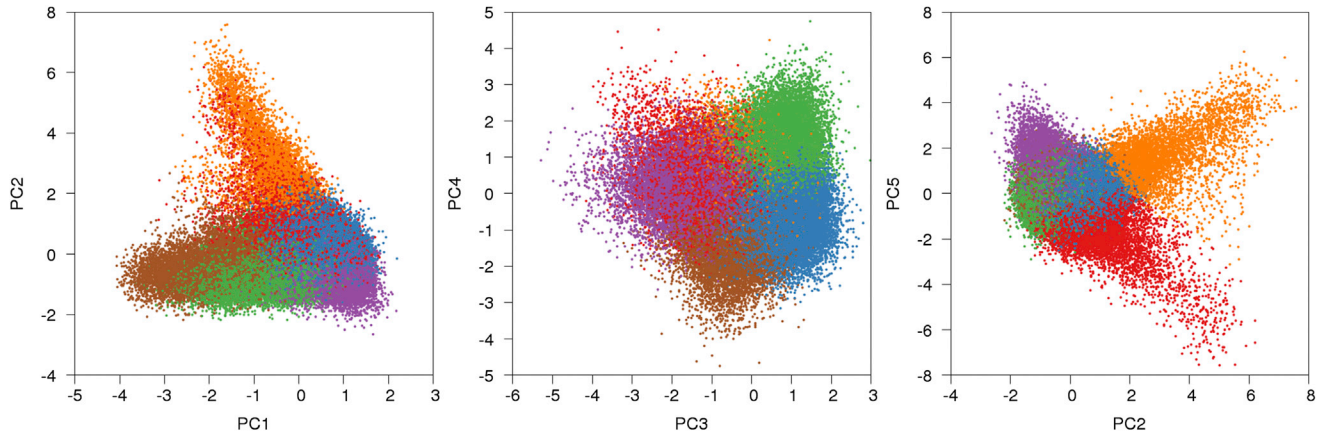

Figure 1. Results of PCA with *k*-Means Clustering
The top five PCs in UK Biobank data are displayed. Samples were clustered with these PCs into six clusters by *k*-means clustering (see Table 1). PC5 is plotted against PC2 because PC5 primarily separated the orange and red clusters, which were separated from the other clusters by PC2.

Results

Population Structure in the UK Biobank

We restricted our analyses of population structure to 113,851 UK Biobank samples of UK ancestry and 202,486 SNPs after quality-control filtering and LD pruning (see Material and Methods). We ran PCA on this data by using our FastPCA implementation[25] (see Web Resources). By visually examining plots of the top ten PCs (Figure S2) and observing that the eigenvalues for the top five PCs were above background levels, we determined that the top five PCs represent geographic population structure (Figure 1). PC1–PC4 were also strongly correlated with birth coordinate (Table S2). The eigenvalue for PC1 was 20.99, which corresponds to the eigenvalue that would be expected at this sample size for two discrete subpopulations of equal size with an F_{ST} of 1.76×10^{-4} (Table S2).

We ran *k*-means clustering on these five PCs to partition the samples into six clusters, given that K PCs can differentiate $K+1$ populations (Figure 1, Table 1, and Figure S3). To identify the populations underlying the six clusters, we projected the PoBI dataset,[20] comprising 2,039 samples from 30 regions of the UK, onto the UK Biobank PCs (Figure 2 and Figure S4). The individuals in the PoBI study were from rural areas of the UK and had all four grandparents born within 80 km of each other, allowing a glimpse into the genetics of the UK before the increase in mobility in the 20th century. We selected representative PoBI sample regions that best aligned with the six UK Biobank clusters by comparing centroids of each projected population region with those from the UK Biobank clusters via visual inspection (see Material and Methods and Table 1). The largest cluster represented southern and eastern England, three clusters represented different regions in the northern UK (northern England, Northern Ireland, and Scotland), and two clusters represented North and South Wales. The PCs separated the six UK clusters along two general geographical axes: a north-south axis and a Welsh-specific axis. PC1 and PC3 both separated individuals on north-south axes of variation, with southern England on one end and one of the northern UK clusters on the other. PC2 separated the Welsh clusters from the rest of the UK. PC4 separated the Scotland cluster from the Northern Ireland cluster. PC5 separated the North Wales and South Wales (also known as Pembrokeshire) clusters from each other. To confirm these clusterings, we ran TreeMix[30,31] on our UK Biobank clusters (Figure S5), as well as on the UK Biobank clusters and PoBI populations (Figure S6), and found that the Celtic subpopulations were grouped separately from the Saxon-related subpopulations; surprisingly, the North and South Wales clusters were separated by TreeMix, possibly because the North Wales cluster potentially contains Saxon-related samples (we note the low F_{ST} values between North Wales and southern England; see F_{ST} values in Table S6). Overall, our results were generally similar to those from the PoBI study;[20] for example, both analyses identified a Welsh axis of differentiation (PC2) and split North and South Wales (PC5). We also observed some differences due to the different

Table 1. Correspondence between UK Biobank Clusters and PoBI Populations

Color	Count	Cluster Name	PoBI Populations
Purple	19,452	northern England	Yorkshire, Lancashire
Blue	41,494	southern England	Hampshire, Devon, Norfolk
Brown	12,895	Northern Ireland	Northern Ireland
Green	21,215	Scotland	Argyll and Bute, Banff and Buchan, Orkney
Red	14,190	North Wales	North Wales
Orange	4,605	South Wales (Pembrokeshire)	northern and southern Pembrokeshire

We report the PoBI population that most closely corresponds to each UK Biobank cluster (see main text).

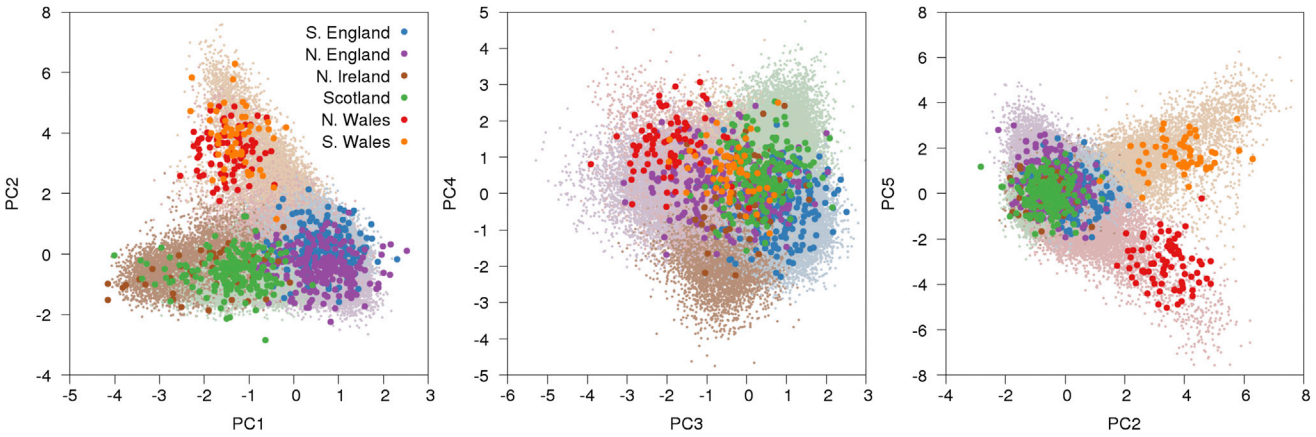

Figure 2. Results of PCA with Projection of PoBI Samples
The top five PCs in UK Biobank data are displayed with PoBI samples projected onto them. PoBI populations that visually best matched the clusters from *k*-means clustering were used for assigning names to the six clusters (Table 1).

sampling schemes; in particular, the UK Biobank dataset contained many more Irish and Scottish samples (driving variation along PC1) and fewer Orkney samples, which affected the clustering of PoBI samples.

We next analyzed UK Biobank population structure in conjunction with ancient DNA samples. Modern European populations are currently thought to have descended from three ancestral populations: Steppe, Mesolithic Europeans, and Neolithic farmers.[21,22] We projected ancient samples from these three populations as well as ancient Saxon samples[24] onto the UK Biobank PCs (see Figure 3, Figure S7, and Material and Methods). These populations were primarily differentiated along PC1 and PC3, indicating higher levels of Steppe ancestry in northern UK populations. Additionally, the lack of any correlation between ancient samples and PC2 suggests that Welsh populations are not differentially admixed with any ancient population in our dataset and most likely underwent Welsh-specific genetic drift. We confirmed these findings by projecting pan-European POPRES[27] samples onto the UK Biobank PCs (see Material and Methods and Figure S8). We note that the Irish and Scottish POPRES populations were projected on top of their corresponding UK Biobank population clusters. Of the continental European populations, Russians (who have the most Steppe ancestry) were projected farther in the Steppe direction along PC1 and PC3 than Spanish and Italians (who have the least Steppe ancestry[22]). Additionally, none of the continental European populations projected onto the same regions as the Welsh on PC2 and PC5. We ran TreeMix on the UK Biobank clusters and ancient populations (Figure S9) as well as the UK Biobank clusters and POPRES populations (Figure S10); although using the ancient populations to infer population structure was challenging, the UK Biobank samples were most closely grouped with the Scottish and Irish POPRES samples.

In addition to the impact of ancient Eurasian populations migrating to the UK, we know that the genetics of the UK has been strongly affected by Anglo-Saxon migrations since the Iron Age,[24] such that the Angles arrived in eastern England and the Saxons arrived in southern England. The Anglo-Saxons interbred with the native Celts,

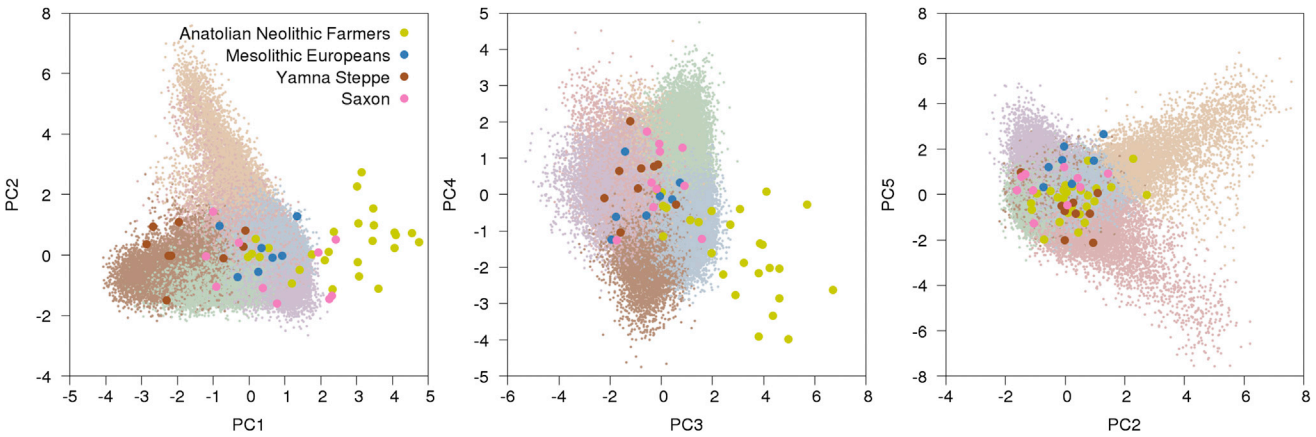

Figure 3. Results of PCA with Projection of Ancient Samples
The top five PCs in UK Biobank data are displayed with ancient samples projected onto them.

Table 2. Results of f_4 Statistics in Ancient and Modern British Samples

		Pop2		
Grouping	**Pop1**	**Hampshire**	**Devon**	**Norfolk**
Ancient	Saxon	2.543	3.732	5.118
Scotland	Argyll and Bute	3.323	6.223	9.560
North Wales	North Wales	1.918	5.239	8.490
South Wales	northern Pembrokeshire	1.759	4.430	7.124

We report f_4 statistics of the form f_4 (Steppe, Neolithic farmer; Pop1, Pop2), representing a Z score whose positive values indicate more Steppe ancestry in Pop1 than in Pop2. Samples for Pop1 were either modern Celtic (Scotland and Wales) or ancient Saxon. Samples for Pop2 were modern Anglo-Saxon (southern and eastern England).

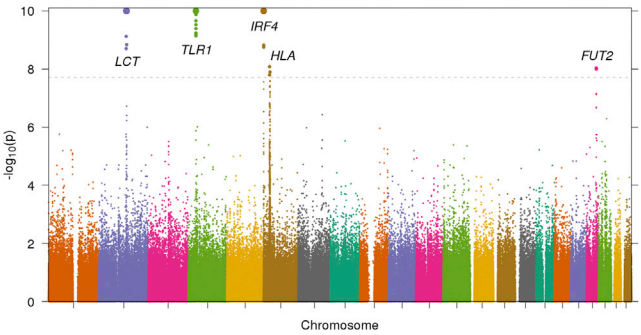

Figure 4. Selection Statistics for UK Biobank along PC1
A Manhattan plot with $-\log_{10}$(p) values is displayed. Values above the significance threshold (dotted line, $p = 1.96 \times 10^{-8}$, $\alpha = 0.05$ after correction for 5 PCs and 510,665 SNPs) are displayed as larger points and are labeled with the locus they correspond to (see Table 3). $-\log_{10}$(p) values larger than 10 are truncated at 10 for easier visualization and are displayed as even larger points.

which explains much of the genetic landscape in the UK. We analyzed a variety of samples from predominantly Celtic (Scotland and Wales) and Anglo-Saxon (southern and eastern England) populations from modern Britain in conjunction with the PoBI samples[20] and ten ancient Saxon samples from eastern England[24] in order to assess the relative amounts of Steppe ancestry. We computed f_4 statistics[28] of the form

$$f_4(\text{Steppe}, \text{Neolithic farmer}; \text{Pop1}, \text{Pop2}),$$

where the Steppe and Neolithic farmer populations are from Lazaridis et al.[21] and Haak et al.,[22] population 1 (Pop1) is either a modern Celtic (Scotland or Wales) or ancient Saxon population, and population 2 (Pop2) is a modern Anglo-Saxon (southern and eastern England) population (Table 2 and Table S3). This statistic is sensitive to Steppe ancestry, such that positive values indicate more Steppe ancestry in Pop1 than in Pop2. We consistently obtained significantly positive f_4 statistics, implying that both the modern Celtic samples and the ancient Saxon samples have more Steppe ancestry than the modern Anglo-Saxon samples from southern and eastern England. This indicates that southern and eastern England are not exclusively a genetic mix of Celts and Saxons, which each have more Steppe ancestry. There are a variety of possible explanations, but one is that the present genetic structure of Britain, although subtle, is quite old and that southern England in Roman times already had less Steppe ancestry than Wales and Scotland.

Signals of Natural Selection

We searched for signals of selection by using a recently developed selection statistic that detects unusual population differentiation along continuous PCs.[25] Notably, this statistic is able to detect selection signals at genome-wide significance. We analyzed the top five UK Biobank PCs (which were computed with LD-pruned SNPs) and computed selection statistics at 510,665 SNPs, reflecting the set of SNPs after quality control but before LD pruning (see Material and Methods). The Manhattan plot for PC1 is reported in Figure 4, and additional plots are provided in Figure S11. We detected genome-wide-significant signals of selection at *FUT2* (MIM: 182100) and at several loci with widely known signals of selection (Table 3). Loci with suggestive signals of selection ($p < 10^{-6}$) are reported in Table S4. *FUT2* has also previously been reported as a target of natural selection;[36,37] those results focused on frequency differences between highly diverged continental populations, whereas our results implicate much more recent selection because UK Biobank populations diverged much more recently than continental populations. *FUT2* encodes fucosyltransferase 2, an enzyme that affects the Lewis blood group. The SNP with the most significant p value, rs601338, is a coding variant where the variant rs601338*G encodes the secretor allele and the rs601338*A variant encodes the nonsecretor allele, which protects against the Norwalk norovirus.[38,39] This SNP also affects the progression of HIV infection[40] (MIM: 609423) and is associated with vitamin B_{12} levels,[41] Crohn disease[42]

Table 3. Top Signals of Selection for UK Biobank along PC1–PC5

Locus	Chromosome	Position (Mb)	PC	Top SNP	p Value
LCT[6]	2	134.9–137.2	1	rs7570971	3.96×10^{-15}
TLR1[32]	4	38.8–38.9	1	rs4833095	7.96×10^{-15}
			2		1.27×10^{-8}
			3		7.89×10^{-9}
			4		1.54×10^{-11}
IRF4[33,34]	6	0.4–0.5	1	rs62389423	2.31×10^{-43}
HLA[35]	6	31.1–32.9	1	rs9366778	8.45×10^{-9}
FUT2	19	49.2–49.2	1	rs601338	9.16×10^{-9}

We report the top signal of natural selection for each locus reaching genome-wide significance (1.96×10^{-8}) along any of the top five PCs. Neighboring SNPs < 1 Mb apart with genome-wide-significant signals were grouped together into a single locus.

Table 4. Top Signals of Selection for Combined Selection Statistics

Locus	Chromosome	Position (Mb)	PC	Top SNP	Combined p Value	UK Biobank p Value	Ancient Eurasian p Value
F12	5	33.9–34.0	4	rs2545801	2.05×10^{-10}	3.99×10^{-5}	5.35×10^{-8}
CYP1A2-CSK	15	75.0–75.1	2	rs1378942	4.65×10^{-8}	1.05×10^{-2}	1.08×10^{-7}

Restricting to loci that were not genome-wide significant in either the UK Biobank or the ancient Eurasian selection statistics, we report the top selection statistic for each locus reaching genome-wide significance. Neighboring SNPs < 1 Mb apart with genome-wide significant signals were grouped together into a single locus.

(MIM: 266600), celiac disease (MIM: 212750), and inflammatory bowel disease,[43] possibly because of changes in energy metabolism in the gut microbiome.[44] rs601338*A is more common in northern UK samples (Table S5). The GERA[45] and PoBI[20] datasets do not include rs601338 but exhibited similar allele-frequency patterns at rs492602 and rs676388 (Table S5), two linked *FUT2* SNPs whose allele frequencies vary on a north-south axis in UK Biobank data. All three SNPs had genome-wide-significant signals of selection in the UK Biobank, and rs601338 and rs492602 were also genome-wide significant when we analyzed the six UK Biobank clusters described above with a selection test based on unusual differentiation between pairs of discrete subpopulations (Table S6). On the other hand, rs492602 and rs676388 were only suggestively significant ($p < 1.00 \times 10^{-6}$) in selection tests using the GERA dataset (Table S7), emphasizing the advantage of analyzing more closely related subpopulations in very large sample sizes in the UK Biobank dataset.

To detect additional signals of selection, we combined our PC-based selection statistics from the UK Biobank data with a previously described selection statistic that detects unusual allele-frequency differences after the admixture of ancient Eurasian populations by identifying SNPs whose allele frequencies are inconsistent with admixture proportions inferred from genome-wide data.[23] For each of PC1–PC5 in the UK Biobank, we summed our chi-square (1 df) selection statistics for that PC with the chi-square (4 df) selection statistics from Mathieson et al.[23] to produce chi-square (5 df) statistics combining these independent signals (see Material and Methods). We confirmed the independence of the two selection statistics by examining the correlations between the two selection statistics and checking that the combined statistics were not substantially inflated, obtaining λ_{GC} values of 1.04–1.06 (Table S8; see Figure S12 for a probability-probability plot). In order to produce maximally conservative statistics, we corrected our combined statistics with these λ_{GC} values. We looked for signals that were genome-wide significant in the combined selection statistic but not in either of the constituent UK Biobank or ancient Eurasian selection statistics. Results are reported in Table 4.

We detected genome-wide-significant signals of selection at the *F12* (MIM: 610619) and *CYP1A2* (MIM: 124060)-*CSK* (MIM: 124095) loci. We are not currently aware of previous evidence of selection at *F12*. *F12* codes for coagulation factor XII, a protein involved in blood clotting.[46] The SNP at the *F12* locus, rs2545801, was suggestively significant in the ancient Eurasian analysis ($p = 5.35 \times 10^{-8}$), and combining it with the UK Biobank selection statistic on PC2 produced a genome-wide-significant signal. This SNP has been associated with activated partial thromboplastin time, a measure of blood-clotting speed where shorter time is a risk factor for strokes.[47] An additional significant SNP at *F12*, rs2731672, affects expression of *F12* in the liver[48] and is associated with plasma levels of factor XII.[49] The *CYP1A2-CSK* locus has previously been reported as a target of natural selection in comparisons of inter-continental allele and haplotype frequencies,[50,51] but our results implicate much more recent selection. The two detected SNPs at this locus are in strong LD ($r^2 = 0.858$). The top SNP, rs1378942, is in an intron in *CSK*. This SNP has a greatly varying allele frequency across continents[51] and is associated with blood pressure[52,53] and systemic sclerosis[54] (an autoimmune disease affecting connective tissue; MIM: 181750). The second SNP, rs2472304 in *CYP1A2*, is associated with esophageal cancer[55] (MIM: 133239) and caffeine consumption[56] and might mediate the protective effect of caffeine on Parkinson disease[57] (MIM: 168600).

Using the top five PCs as covariates, we tested SNPs with genome-wide-significant signals of selection in the constituent UK Biobank or ancient Eurasian scans or the combined scan for association with 15 phenotypes in the UK Biobank dataset (see Table S9 and Material and Methods). The top SNP at *F12* (rs2545801) was associated with height ($p = 4.8 \times 10^{-11}$), and the top SNP at *CYP1A2-CSK* (rs1378942) was associated with DBP ($p = 3.6 \times 10^{-19}$) and hypertension ($p = 4.8 \times 10^{-9}$), consistent with previous findings.[58] We detected additional associations with DBP ($p = 8.0 \times 10^{-33}$) and hypertension ($p = 1.3 \times 10^{-9}$) at the *ATXN2* (MIM: 601517)-*SH2B3* (MIM: 605093) locus, which was reported as under selection in the ancient Eurasian scan. The top SNP in *ATXN2-SH2B3*, rs3184504, is known to be associated with blood pressure.[59] We note that PC1 and PC3 were strongly associated with height in the UK Biobank dataset, and PC3 and PC4 were associated with DBP (Table S10). *GRK4*[60] (MIM: 137026), *AGT*[60] (MIM: 106150), and *ATP1A1*[14] (MIM: 182310) have also been reported to be under selection and to be associated with DBP or hypertension. None of the SNPs in *GRK4* or *ATP1A1* were found to be under selection or associated with DBP or hypertension in our analyses. The *AGT* SNP rs699 was associated with DBP ($p = 7.2 \times 10^{-10}$) and nominally associated with hypertension ($p = 4.8 \times 10^{-4}$), although it did not produce a significant signal of selection in our analyses.

Discussion

In this study, we used PCA to analyze the population structure of a large UK cohort (N = 113,851). We detected five PCs representing geographic population structure that partitioned this cohort into six subpopulation clusters. Projecting ancient samples onto these PCs revealed greater Steppe ancestry in northern UK samples. No ancient samples were found to vary along the Welsh-specific axis, suggesting that the Welsh populations differ from the rest of the UK as a result of drift rather than different levels of admixture. We also determined that UK population structure cannot be explained as a simple mixture of Celts and Saxons.

We leveraged the subtle population structure and large sample size of the UK Biobank dataset to detect signals of natural selection. We determined that the rs601338*A allele of *FUT2* is more common in northern UK samples, suggesting that pathogens might have exerted selective pressure in those populations. Combining a selection statistic that detects selection via population differentiation within the UK with a separate statistic that detects selection since ancient population admixture in Europe, we were able to detect selection at two additional loci, *F12* and *CYP1A2-CSK*. We additionally found associations with DBP at *CYP1A2-CSK* and at the *ATXN2-SH2B3* locus, implicated in a previous selection scan.

We conclude by noting three limitations in our work. First, we employed PCA, a widely used method for analyzing population structure,[25,29,61] but haplotype-based methods such as fineSTRUCTURE could be more powerful;[20,62,63] recent advances in computationally efficient phasing[64,65] increase the prospects for applying such methods to biobank-scale data. Second, we employed methods designed to detect selection at individual loci but did not employ methods to detect polygenic selection;[66–70] our observation that top PCs were correlated with height and DBP in the UK Biobank dataset, which could potentially be consistent with the action of polygenic selection on these traits, motivates further analyses of possible polygenic selection. Finally, the PC-based test for selection that we employed assumes that allele frequencies vary linearly along a PC. The spatial ancestry analysis (SPA) method[71–73] allows for a logistic relationship between allele frequency and ancestry and is not constrained by this limitation. However, the advantage of the PC-based test for selection over SPA is that it provides an assessment of statistical significance (p values), allowing for the detection of genome-wide-significant signals, a key consideration in genome scans for selection.

Supplemental Data

Supplemental Data include 12 figures and 10 tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2016.09.014.

Acknowledgments

We thank Iain Mathieson and David Reich for helpful discussions and Stephan Schiffels for technical assistance with Saxon samples. This research was conducted with the UK Biobank Resource (application numbers 10438 and 14292) and was funded by NIH grant R01 HG006399.

Received: June 16, 2016
Accepted: September 21, 2016
Published: October 20, 2016

Web Resources

EIGENSOFT v.6.1.3, http://www.hsph.harvard.edu/alkes-price/software
Genome Reference Consortium Human Build 37 (GRCh37), https://www.ncbi.nlm.nih.gov/assembly/2758/
OMIM, http://www.omim.org
PLINK2, https://www.cog-genomics.org/plink2
TreeMix, https://bitbucket.org/nygcresearch/treemix/wiki/Home
UK Biobank, http://www.ukbiobank.ac.uk

References

1. Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z.P., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J., et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. Nature *419*, 832–837.
2. Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C., and Clark, A.G. (2007). Recent and ongoing selection in the human genome. Nat. Rev. Genet. *8*, 857–868.
3. Novembre, J., and Di Rienzo, A. (2009). Spatial patterns of variation due to natural selection in humans. Nat. Rev. Genet. *10*, 745–755.
4. Scheinfeldt, L.B., and Tishkoff, S.A. (2013). Recent human adaptation: genomic approaches, interpretation and insights. Nat. Rev. Genet. *14*, 692–702.
5. Shriver, M.D., Kennedy, G.C., Parra, E.J., Lawson, H.A., Sonpar, V., Huang, J., Akey, J.M., and Jones, K.W. (2004). The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. Hum. Genomics *1*, 274–286.
6. Bersaglieri, T., Sabeti, P.C., Patterson, N., Vanderploeg, T., Schaffner, S.F., Drake, J.A., Rhodes, M., Reich, D.E., and Hirschhorn, J.N. (2004). Genetic signatures of strong recent positive selection at the lactase gene. Am. J. Hum. Genet. *74*, 1111–1120.
7. Tishkoff, S.A., Reed, F.A., Ranciaro, A., Voight, B.F., Babbitt, C.C., Silverman, J.S., Powell, K., Mortensen, H.M., Hirbo, J.B., Osman, M., et al. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. Nat. Genet. *39*, 31–40.
8. Fumagalli, M., Moltke, I., Grarup, N., Racimo, F., Bjerregaard, P., Jørgensen, M.E., Korneliussen, T.S., Gerbault, P., Skotte, L., Linneberg, A., et al. (2015). Greenlandic Inuit show genetic signatures of diet and climate adaptation. Science *349*, 1343–1347.
9. Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z.X.P., Pool, J.E., Xu, X., Jiang, H., Vinckenbosch, N., Korneliussen, T.S., et al. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. Science *329*, 75–78.

10. Bigham, A., Bauchet, M., Pinto, D., Mao, X., Akey, J.M., Mei, R., Scherer, S.W., Julian, C.G., Wilson, M.J., López Herráez, D., et al. (2010). Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. PLoS Genet. *6*, e1001116.
11. Lorenzo, F.R., Huff, C., Myllymäki, M., Olenchock, B., Swierczek, S., Tashi, T., Gordeuk, V., Wuren, T., Ri-Li, G., McClain, D.A., et al. (2014). A genetic mechanism for Tibetan high-altitude adaptation. Nat. Genet. *46*, 951–956.
12. Hamblin, M.T., and Di Rienzo, A. (2000). Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am. J. Hum. Genet. *66*, 1669–1679.
13. Ayodo, G., Price, A.L., Keinan, A., Ajwang, A., Otieno, M.F., Orago, A.S.S., Patterson, N., and Reich, D. (2007). Combining evidence of natural selection with association analysis increases power to detect malaria-resistance variants. Am. J. Hum. Genet. *81*, 234–242.
14. Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., Karthikeyan, S., Iles, L., Pollard, M.O., Choudhury, A., et al. (2015). The African Genome Variation Project shapes medical genetics in Africa. Nature *517*, 327–332.
15. Lamason, R.L., Mohideen, M.-A.P.K., Mest, J.R., Wong, A.C., Norton, H.L., Aros, M.C., Jurynec, M.J., Mao, X., Humphreville, V.R., Humbert, J.E., et al. (2005). SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science *310*, 1782–1786.
16. Perry, G.H., Dominy, N.J., Claw, K.G., Lee, A.S., Fiegler, H., Redon, R., Werner, J., Villanea, F.A., Mountain, J.L., Misra, R., et al. (2007). Diet and the evolution of human amylase gene copy number variation. Nat. Genet. *39*, 1256–1260.
17. Hancock, A.M., Witonsky, D.B., Alkorta-Aranburu, G., Beall, C.M., Gebremedhin, A., Sukernik, R., Utermann, G., Pritchard, J.K., Coop, G., and Di Rienzo, A. (2011). Adaptations to climate-mediated selective pressures in humans. PLoS Genet. *7*, e1001375.
18. Ko, W.-Y., Rajan, P., Gomez, F., Scheinfeldt, L., An, P., Winkler, C.A., Froment, A., Nyambo, T.B., Omar, S.A., Wambebe, C., et al. (2013). Identifying Darwinian selection acting on different human APOL1 variants among diverse African populations. Am. J. Hum. Genet. *93*, 54–66.
19. Bhatia, G., Patterson, N., Pasaniuc, B., Zaitlen, N., Genovese, G., Pollack, S., Mallick, S., Myers, S., Tandon, A., Spencer, C., et al. (2011). Genome-wide comparison of African-ancestry populations from CARe and other cohorts reveals signals of natural selection. Am. J. Hum. Genet. *89*, 368–381.
20. Leslie, S., Winney, B., Hellenthal, G., Davison, D., Boumertit, A., Day, T., Hutnik, K., Royrvik, E.C., Cunliffe, B., Lawson, D.J., et al.; Wellcome Trust Case Control Consortium 2; International Multiple Sclerosis Genetics Consortium (2015). The fine-scale genetic structure of the British population. Nature *519*, 309–314.
21. Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P.H., Schraiber, J.G., Castellano, S., Lipson, M., et al. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature *513*, 409–413.
22. Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G., Nordenfelt, S., Harney, E., Stewardson, K., et al. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. Nature *522*, 207–211.
23. Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S.A., Harney, E., Stewardson, K., Fernandes, D., Novak, M., et al. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. Nature *528*, 499–503.
24. Schiffels, S., Haak, W., Paajanen, P., Llamas, B., Popescu, E., Loe, L., Clarke, R., Lyons, A., Mortimer, R., Sayer, D., et al. (2016). Iron Age and Anglo-Saxon genomes from East England reveal British migration history. Nat. Commun. *7*, 10408.
25. Galinsky, K.J., Bhatia, G., Loh, P.-R., Georgiev, S., Mukherjee, S., Patterson, N.J., and Price, A.L. (2016). Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. Am. J. Hum. Genet. *98*, 456–472.
26. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 7.
27. Nelson, M.R., Bryc, K., King, K.S., Indap, A., Boyko, A.R., Novembre, J., Briley, L.P., Maruyama, Y., Waterworth, D.M., Waeber, G., et al. (2008). The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. Am. J. Hum. Genet. *83*, 347–358.
28. Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2012). Ancient admixture in human history. Genetics *192*, 1065–1093.
29. Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet. *2*, e190.
30. Pickrell, J.K., and Pritchard, J.K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. *8*, e1002967.
31. Pickrell, J.K., Patterson, N., Barbieri, C., Berthold, F., Gerlach, L., Güldemann, T., Kure, B., Mpoloka, S.W., Nakagawa, H., Naumann, C., et al. (2012). The genetic prehistory of southern Africa. Nat. Commun. *3*, 1143.
32. Heffelfinger, C., Pakstis, A.J., Speed, W.C., Clark, A.P., Haigh, E., Fang, R., Furtado, M.R., Kidd, K.K., and Snyder, M.P. (2014). Haplotype structure and positive selection at TLR1. Eur. J. Hum. Genet. *22*, 551–557.
33. Burton, P.R., Clayton, D.G., Cardon, L.R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D.P., McCarthy, M.I., Ouwehand, W.H., Samani, N.J., et al.; Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661–678.
34. Pickrell, J.K., Coop, G., Novembre, J., Kudaravalli, S., Li, J.Z., Absher, D., Srinivasan, B.S., Barsh, G.S., Myers, R.M., Feldman, M.W., and Pritchard, J.K. (2009). Signals of recent positive selection in a worldwide sample of human populations. Genome Res. *19*, 826–837.
35. de Bakker, P.I.W., McVean, G., Sabeti, P.C., Miretti, M.M., Green, T., Marchini, J., Ke, X., Monsuur, A.J., Whittaker, P., Delgado, M., et al. (2006). A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat. Genet. *38*, 1166–1172.
36. Ferrer-Admetlla, A., Sikora, M., Laayouni, H., Esteve, A., Roubinet, F., Blancher, A., Calafell, F., Bertranpetit, J., and Casals, F. (2009). A natural history of FUT2 polymorphism in humans. Mol. Biol. Evol. *26*, 1993–2003.
37. Fumagalli, M., Cagliani, R., Pozzoli, U., Riva, S., Comi, G.P., Menozzi, G., Bresolin, N., and Sironi, M. (2009). Widespread balancing selection and pathogen-driven selection at blood group antigen genes. Genome Res. *19*, 199–212.

38. Thorven, M., Grahn, A., Hedlund, K.-O., Johansson, H., Wahlfrid, C., Larson, G., and Svensson, L. (2005). A homozygous nonsense mutation (428G–>A) in the human secretor (FUT2) gene provides resistance to symptomatic norovirus (GGII) infections. J. Virol. *79*, 15351–15355.
39. Carlsson, B., Kindberg, E., Buesa, J., Rydell, G.E., Lidón, M.F., Montava, R., Abu Mallouh, R., Grahn, A., Rodríguez-Díaz, J., Bellido, J., et al. (2009). The G428A nonsense mutation in FUT2 provides strong but not absolute protection against symptomatic GII.4 Norovirus infection. PLoS ONE *4*, e5593.
40. Kindberg, E., Hejdeman, B., Bratt, G., Wahren, B., Lindblom, B., Hinkula, J., and Svensson, L. (2006). A nonsense mutation (428G–>A) in the fucosyltransferase FUT2 gene affects the progression of HIV-1 infection. AIDS *20*, 685–689.
41. Hazra, A., Kraft, P., Selhub, J., Giovannucci, E.L., Thomas, G., Hoover, R.N., Chanock, S.J., and Hunter, D.J. (2008). Common variants of FUT2 are associated with plasma vitamin B12 levels. Nat. Genet. *40*, 1160–1162.
42. McGovern, D.P.B., Jones, M.R., Taylor, K.D., Marciante, K., Yan, X., Dubinsky, M., Ippoliti, A., Vasiliauskas, E., Berel, D., Derkowski, C., et al.; International IBD Genetics Consortium (2010). Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn's disease. Hum. Mol. Genet. *19*, 3468–3476.
43. Parmar, A.S., Alakulppi, N., Paavola-Sakki, P., Kurppa, K., Halme, L., Färkkilä, M., Turunen, U., Lappalainen, M., Kontula, K., Kaukinen, K., et al. (2012). Association study of FUT2 (rs601338) with celiac disease and inflammatory bowel disease in the Finnish population. Tissue Antigens *80*, 488–493.
44. Tong, M., McHardy, I., Ruegger, P., Goudarzi, M., Kashyap, P.C., Haritunians, T., Li, X., Graeber, T.G., Schwager, E., Huttenhower, C., et al. (2014). Reprograming of gut microbiome energy metabolism by the FUT2 Crohn's disease risk polymorphism. ISME J. *8*, 2193–2206.
45. Banda, Y., Kvale, M.N., Hoffmann, T.J., Hesselson, S.E., Ranatunga, D., Tang, H., Sabatti, C., Croen, L.A., Dispensa, B.P., Henderson, M., et al. (2015). Characterizing Race/Ethnicity and Genetic Ancestry for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. Genetics *200*, 1285–1295.
46. Renné, T., Schmaier, A.H., Nickel, K.F., Blombäck, M., and Maas, C. (2012). In vivo roles of factor XII. Blood *120*, 4296–4303.
47. Tang, W., Schwienbacher, C., Lopez, L.M., Ben-Shlomo, Y., Oudot-Mellakh, T., Johnson, A.D., Samani, N.J., Basu, S., Gögele, M., Davies, G., et al. (2012). Genetic associations for activated partial thromboplastin time and prothrombin time, their gene expression profiles, and risk of coronary artery disease. Am. J. Hum. Genet. *91*, 152–162.
48. Innocenti, F., Cooper, G.M., Stanaway, I.B., Gamazon, E.R., Smith, J.D., Mirkov, S., Ramirez, J., Liu, W., Lin, Y.S., Moloney, C., et al. (2011). Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. PLoS Genet. *7*, e1002078.
49. Guerrero, J.A., Rivera, J., Quiroga, T., Martinez-Perez, A., Antón, A.I., Martínez, C., Panes, O., Vicente, V., Mezzano, D., Soria, J.-M., and Corral, J. (2011). Novel loci involved in platelet function and platelet count identified by a genome-wide study performed in children. Haematologica *96*, 1335–1343.
50. Wooding, S.P., Watkins, W.S., Bamshad, M.J., Dunn, D.M., Weiss, R.B., and Jorde, L.B. (2002). DNA sequence variation in a 3.7-kb noncoding sequence 5′ of the CYP1A2 gene: implications for human population history and natural selection. Am. J. Hum. Genet. *71*, 528–542.
51. Ding, K., and Kullo, I.J. (2011). Geographic differences in allele frequencies of susceptibility SNPs for cardiovascular disease. BMC Med. Genet. *12*, 55.
52. Newton-Cheh, C., Johnson, T., Gateva, V., Tobin, M.D., Bochud, M., Coin, L., Najjar, S.S., Zhao, J.H., Heath, S.C., Eyheramendy, S., et al.; Wellcome Trust Case Control Consortium (2009). Genome-wide association study identifies eight loci associated with blood pressure. Nat. Genet. *41*, 666–676.
53. Tabara, Y., Kohara, K., Kita, Y., Hirawa, N., Katsuya, T., Ohkubo, T., Hiura, Y., Tajima, A., Morisaki, T., Miyata, T., et al.; Global Blood Pressure Genetics Consortium (2010). Common variants in the ATP2B1 gene are associated with susceptibility to hypertension: the Japanese Millennium Genome Project. Hypertension *56*, 973–980.
54. Martin, J.-E., Broen, J.C., Carmona, F.D., Teruel, M., Simeon, C.P., Vonk, M.C., van 't Slot, R., Rodriguez-Rodriguez, L., Vicente, E., Fonollosa, V., et al.; Spanish Scleroderma Group (2012). Identification of CSK as a systemic sclerosis genetic risk factor through Genome Wide Association Study follow-up. Hum. Mol. Genet. *21*, 2825–2835.
55. Xie, Q., Ratnasinghe, L.D., Hong, H., Perkins, R., Tang, Z.-Z., Hu, N., Taylor, P.R., and Tong, W. (2005). Decision forest analysis of 61 single nucleotide polymorphisms in a case-control study of esophageal cancer; a novel method. BMC Bioinformatics *6* (*Suppl 2*), S4.
56. Cornelis, M.C., Monda, K.L., Yu, K., Paynter, N., Azzato, E.M., Bennett, S.N., Berndt, S.I., Boerwinkle, E., Chanock, S., Chatterjee, N., et al. (2011). Genome-wide meta-analysis identifies regions on 7p21 (AHR) and 15q24 (CYP1A2) as determinants of habitual caffeine consumption. PLoS Genet. *7*, e1002033.
57. Popat, R.A., Van Den Eeden, S.K., Tanner, C.M., Kamel, F., Umbach, D.M., Marder, K., Mayeux, R., Ritz, B., Ross, G.W., Petrovitch, H., et al. (2011). Coffee, ADORA2A, and CYP1A2: the caffeine connection in Parkinson's disease. Eur. J. Neurol. *18*, 756–765.
58. Hong, K.-W., Jin, H.-S., Lim, J.-E., Kim, S., Go, M.J., and Oh, B. (2010). Recapitulation of two genomewide association studies on blood pressure and essential hypertension in the Korean population. J. Hum. Genet. *55*, 336–341.
59. Ehret, G.B., Munroe, P.B., Rice, K.M., Bochud, M., Johnson, A.D., Chasman, D.I., Smith, A.V., Tobin, M.D., Verwoert, G.C., Hwang, S.J., et al.; International Consortium for Blood Pressure Genome-Wide Association Studies; CARDIoGRAM consortium; CKDGen Consortium; KidneyGen Consortium; EchoGen consortium; CHARGE-HF consortium (2011). Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature *478*, 103–109.
60. Sabeti, P.C., Schaffner, S.F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T.S., Altshuler, D., and Lander, E.S. (2006). Positive natural selection in the human lineage. Science *312*, 1614–1620.
61. Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., Nelson, M.R., et al. (2008). Genes mirror geography within Europe. Nature *456*, 98–101.
62. Lawson, D.J., Hellenthal, G., Myers, S., and Falush, D. (2012). Inference of population structure using dense haplotype data. PLoS Genet. *8*, e1002453.
63. Walter, K., Min, J.L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J.R., Xu, C., Futema, M., Lawson, D.,

et al.; UK10K Consortium (2015). The UK10K project identifies rare variants in health and disease. Nature *526*, 82–90.
64. Loh, P.R., Palamara, P.F., and Price, A.L. (2016). Fast and accurate long-range phasing in a UK Biobank cohort. Nat. Genet. *48*, 811–816.
65. O'Connell, J., Sharp, K., Shrine, N., Wain, L., Hall, I., Tobin, M., Zagury, J.F., Delaneau, O., and Marchini, J. (2016). Haplotype estimation for biobank-scale data sets. Nat. Genet. *48*, 817–820.
66. Pritchard, J.K., Pickrell, J.K., and Coop, G. (2010). The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr. Biol. *20*, R208–R215.
67. Pritchard, J.K., and Di Rienzo, A. (2010). Adaptation - not by sweeps alone. Nat. Rev. Genet. *11*, 665–667.
68. Turchin, M.C., Chiang, C.W., Palmer, C.D., Sankararaman, S., Reich, D., and Hirschhorn, J.N.; Genetic Investigation of ANthropometric Traits (GIANT) Consortium (2012). Evidence of widespread selection on standing variation in Europe at height-associated SNPs. Nat. Genet. *44*, 1015–1019.
69. Berg, J.J., and Coop, G. (2014). A population genetic signal of polygenic adaptation. PLoS Genet. *10*, e1004412.
70. Robinson, M.R., Hemani, G., Medina-Gomez, C., Mezzavilla, M., Esko, T., Shakhbazov, K., Powell, J.E., Vinkhuyzen, A., Berndt, S.I., Gustafsson, S., et al. (2015). Population genetic differentiation of height and body mass index across Europe. Nat. Genet. *47*, 1357–1362.
71. Yang, W.-Y., Novembre, J., Eskin, E., and Halperin, E. (2012). A model-based approach for analysis of spatial structure in genetic data. Nat. Genet. *44*, 725–731.
72. Baran, Y., Quintela, I., Carracedo, A., Pasaniuc, B., and Halperin, E. (2013). Enhanced localization of genetic samples through linkage-disequilibrium correction. Am. J. Hum. Genet. *92*, 882–894.
73. Baran, Y., and Halperin, E. (2015). A Note on the Relations Between Spatio-Genetic Models. J. Comput. Biol. *22*, 905–917.

The American Journal of Human Genetics, Volume *99*

Supplemental Data

Population Structure of UK Biobank and Ancient Eurasians Reveals Adaptation at Genes Influencing Blood Pressure

Kevin J. Galinsky, Po-Ru Loh, Swapan Mallick, Nick J. Patterson, and Alkes L. Price

Supplementary Figures

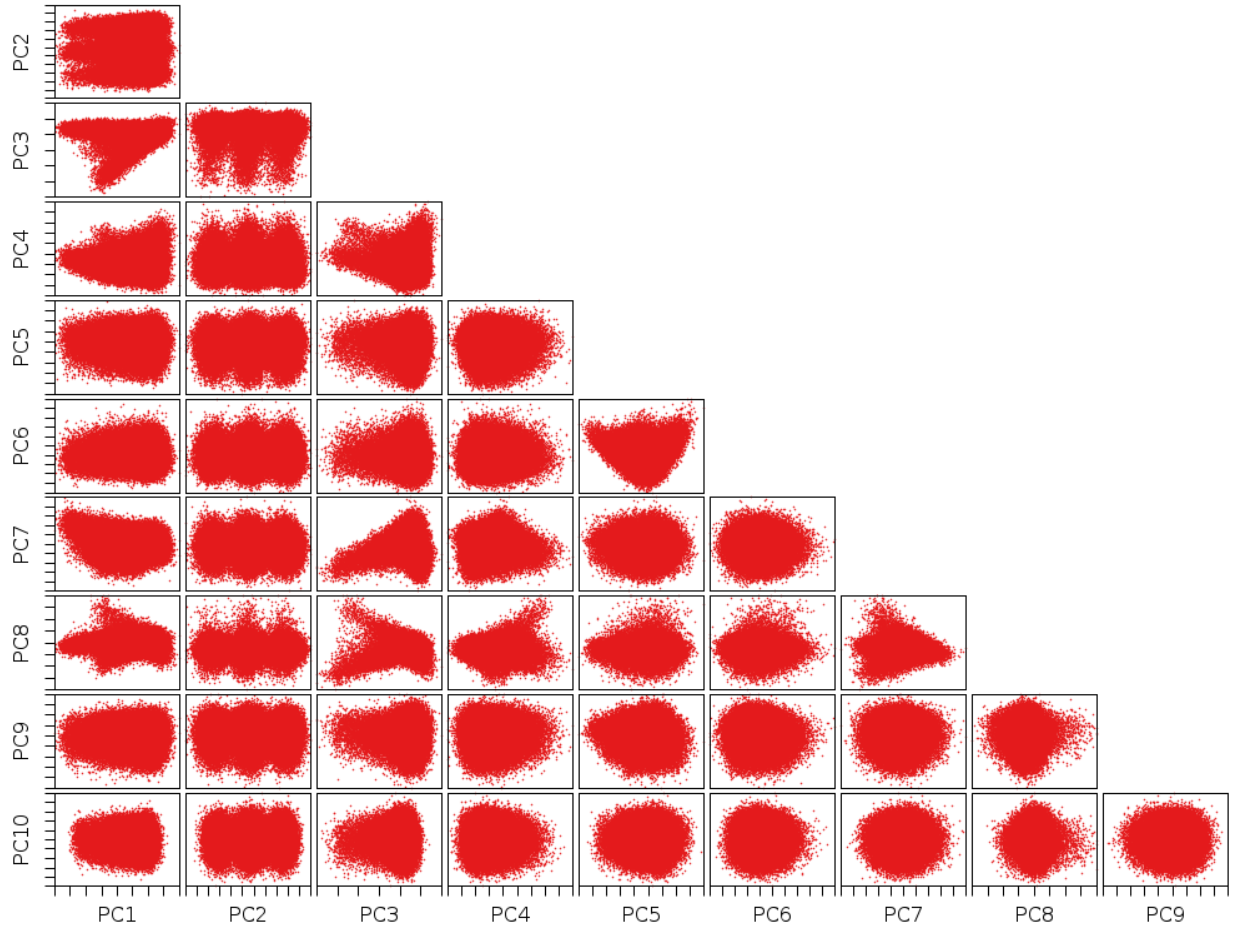

Figure S1 Results of initial PCA run

Shown are the PCA plots for PC1 to PC10 after the initial PCA run. Several of these PCs are dominated by regions of long-range LD. In particular, the three clusters along PC2 indicate 0, 1 or 2 copies of a chromosome 8 inversion variant.

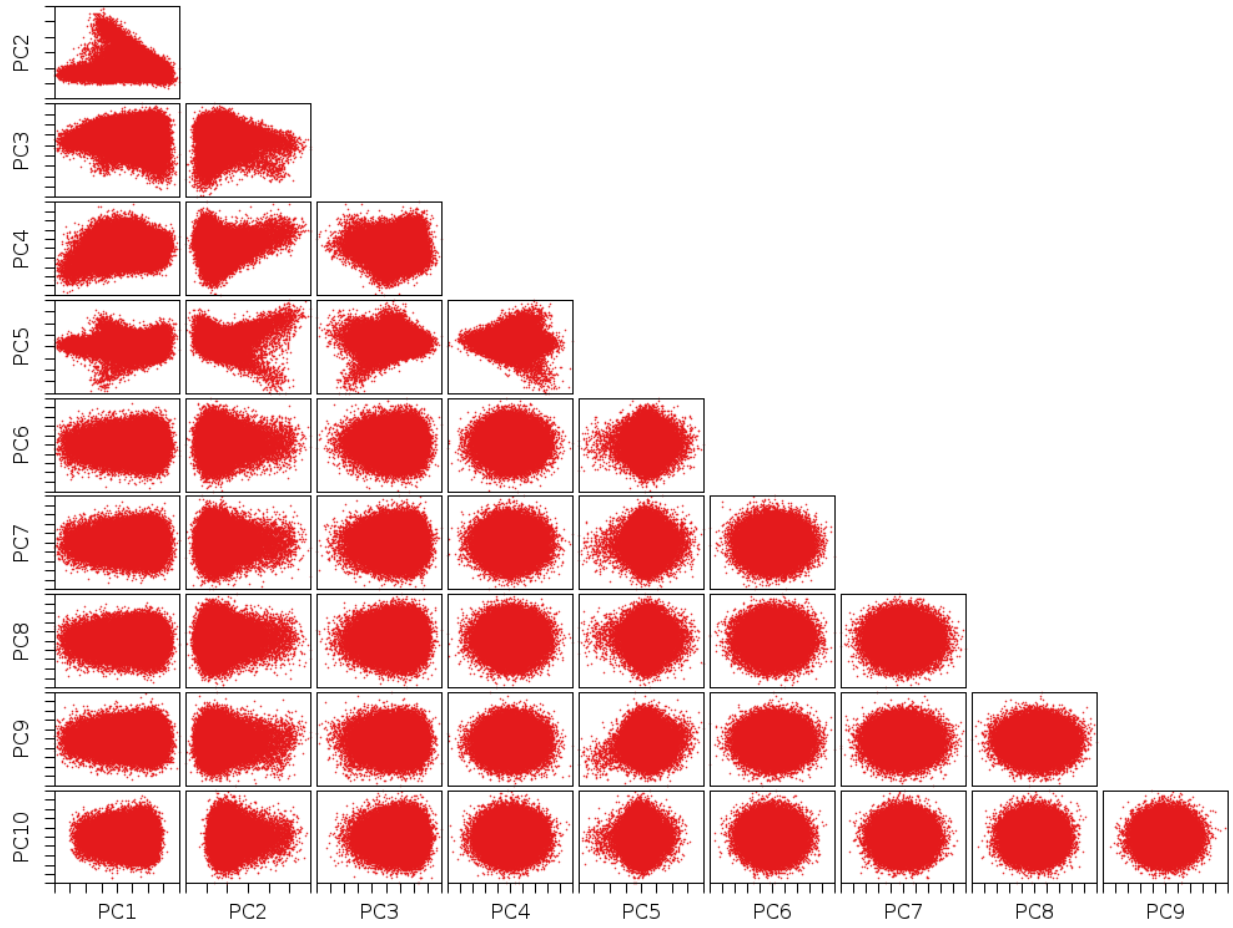

Figure S2 Results of PCA after removing long-range LD regions

Regions with high SNP weights from the first PCA run were removed and PCA was run on the remainder of the genome (see Methods). The resulting PCs are no longer influenced by long-range LD regions. A visual inspection suggests that PC1-PC5 have interesting population structure while PC6-PC10 do not.

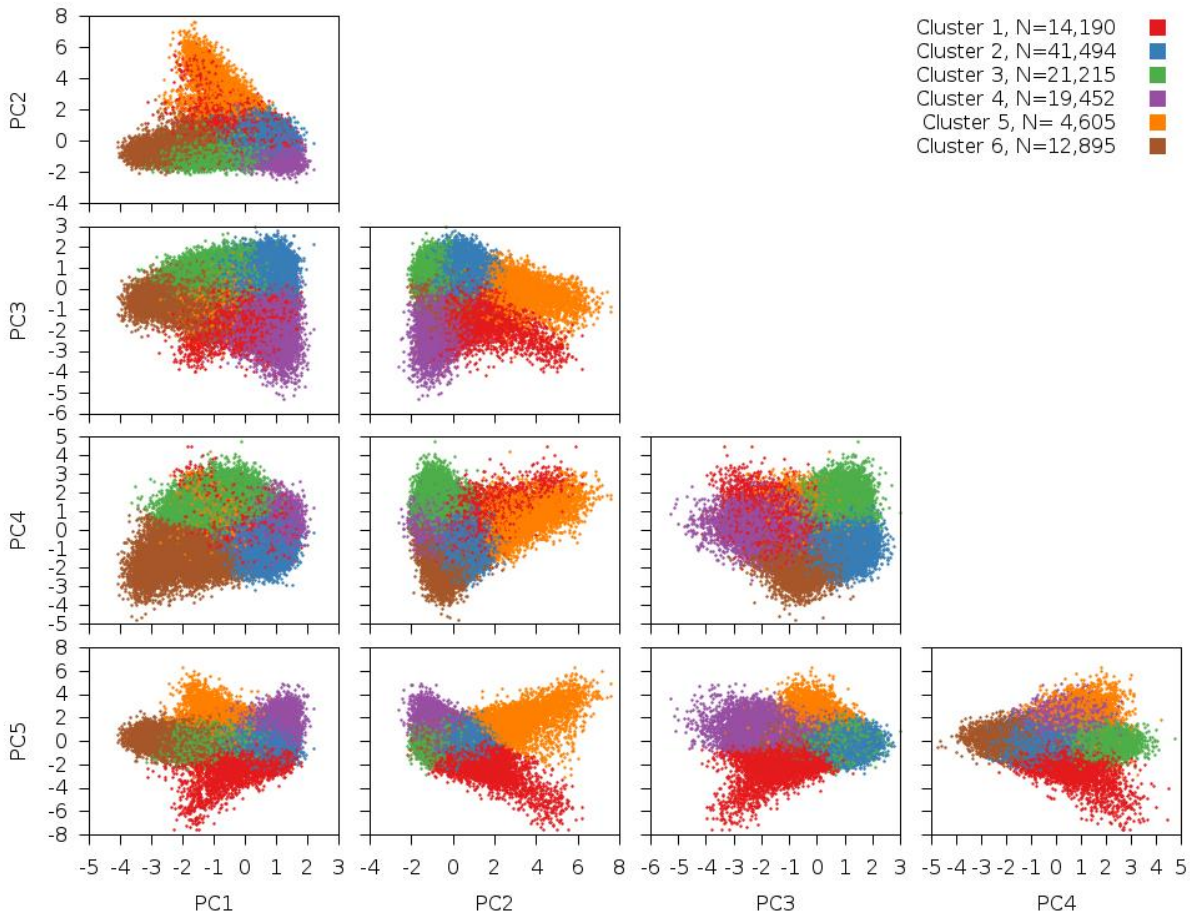

Figure S3 Results of PCA with *k*-means clustering for all PCs

This is an expanded set of plots similar to Figure 1, except that plots of all pairs of top PCs are displayed.

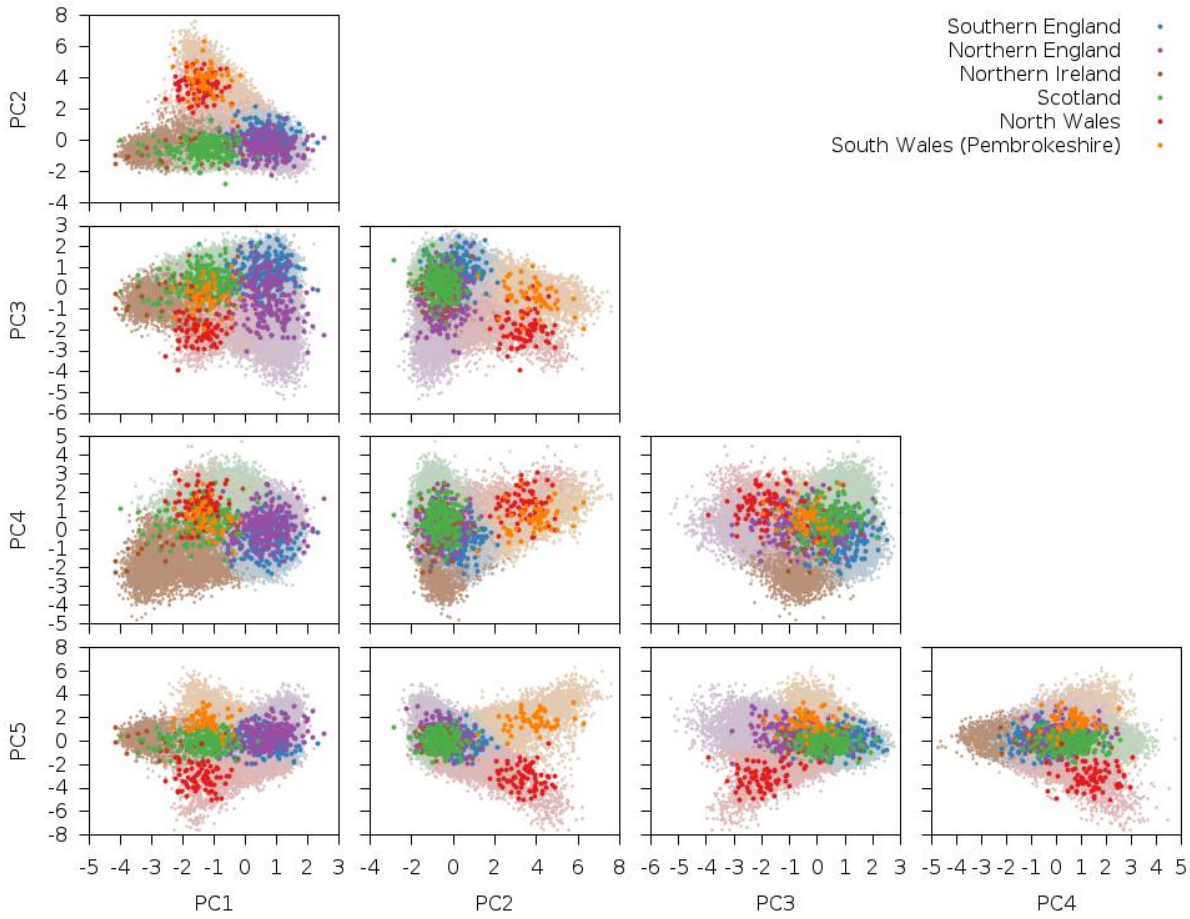

Figure S4 Results of PCA with projection of PoBI samples for all PCs

This is an expanded set of plots similar to Figure 2, except that plots of all pairs of top PCs are displayed.

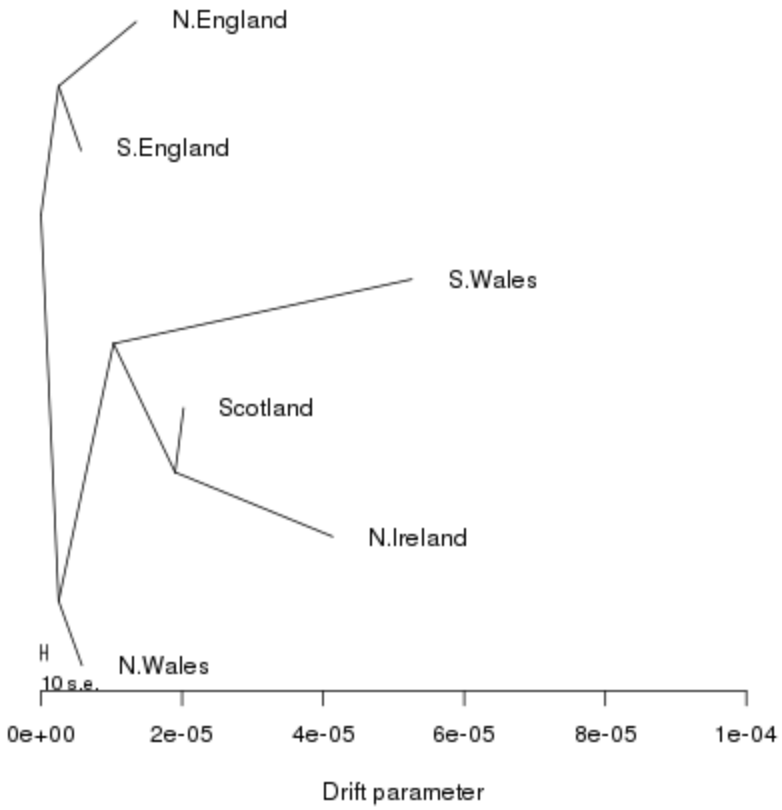

Figure S5 Tree-based clustering of UK Biobank subpopulation clusters

We clustered UK Biobank subpopulation clusters with TreeMix. We found that Northern and Southern England clusters were grouped together while the Celtic-related clusters formed a separate branch to the tree. The disparity between the north and south Welsh clusters is due to the north Wales cluster containing Saxon samples (see F_{ST} values in Table S6).

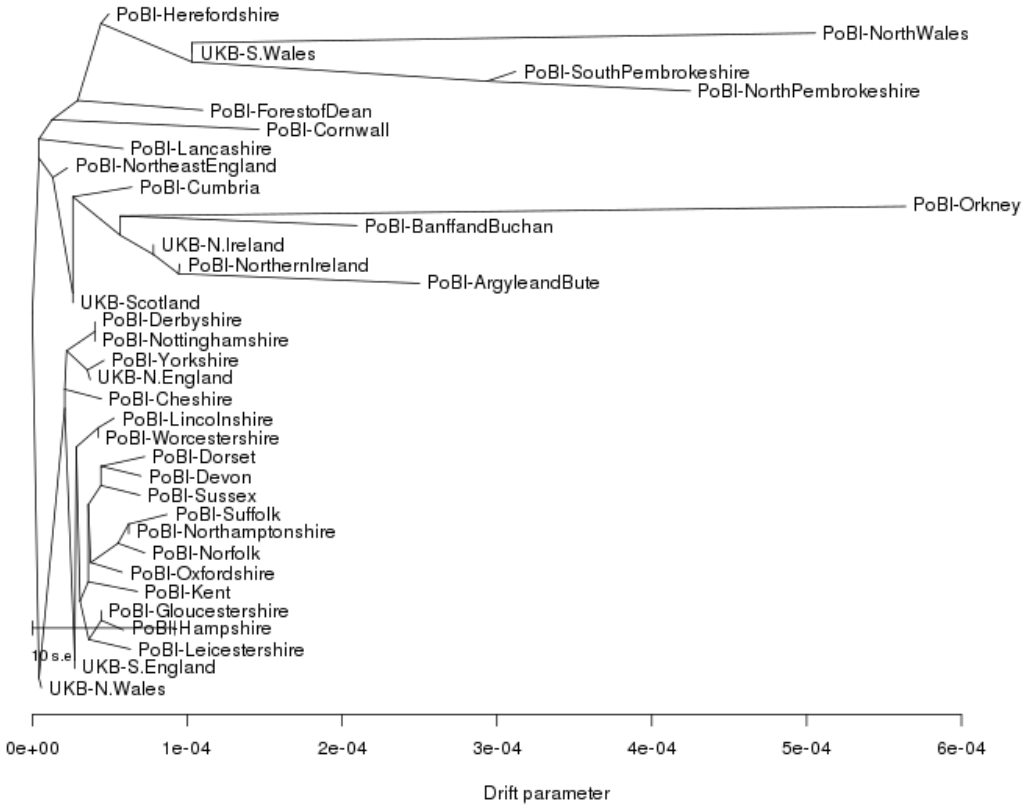

Figure S6 Tree-based clustering of UK Biobank clusters and PoBI populations

Here we see how the UK Biobank subpopulations cluster with the PoBI ones. As in Figure S5, the tree is split roughly into the "Celtic" (top) and "Saxon" (bottom) groups. The south Wales (UKB-S.Wales) cluster is grouped with the Welsh populations from PoBI while the north Wales (UKB-N.Wales) cluster is in the bottom group with the north and south England clusters. As in Figure S5, the disparity between the north and south Welsh clusters is due to the north Wales containing Saxon samples (see F_{ST} values in Table S6).

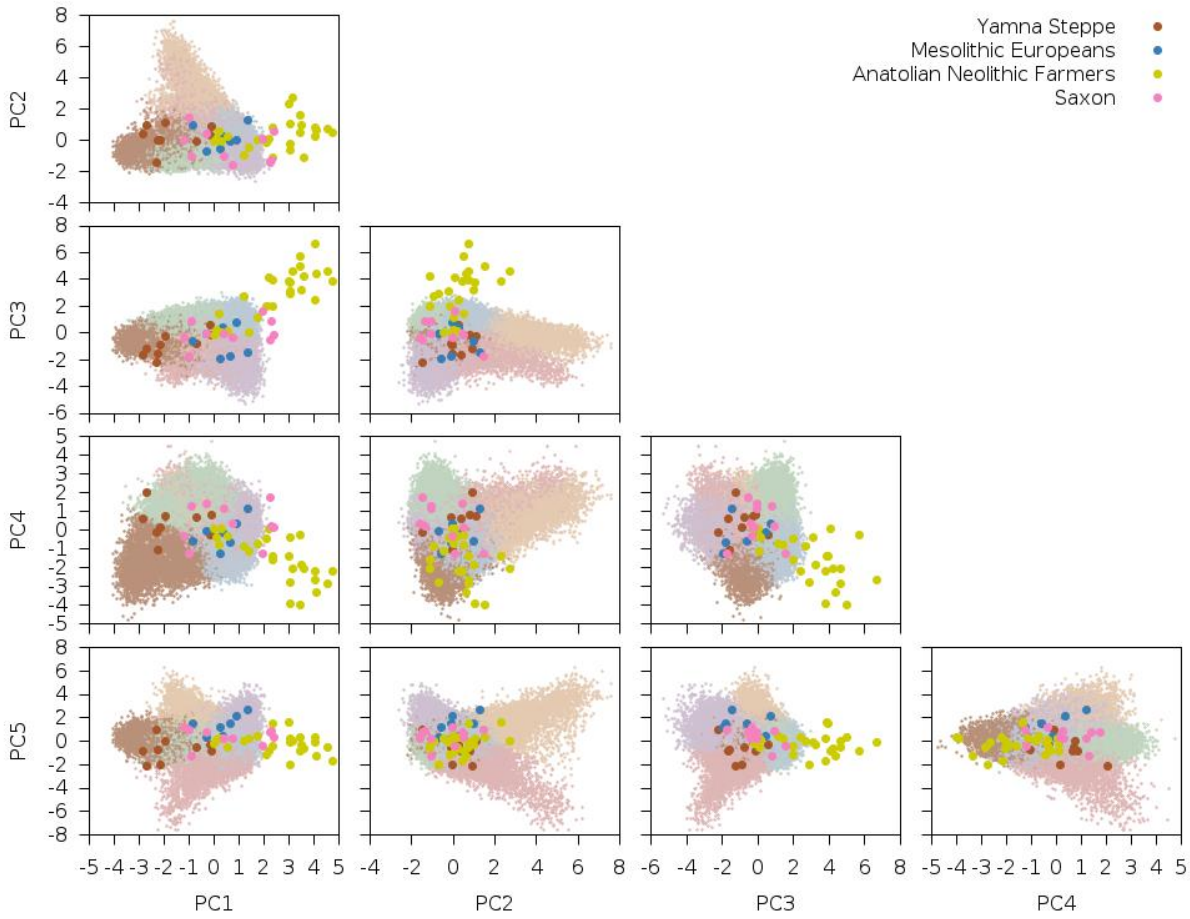

Figure S7 Results of PCA with projection of ancient samples for all PCs

This is an expanded set of plots similar Figure 3, except that plots of all pairs of top PCs are displayed.

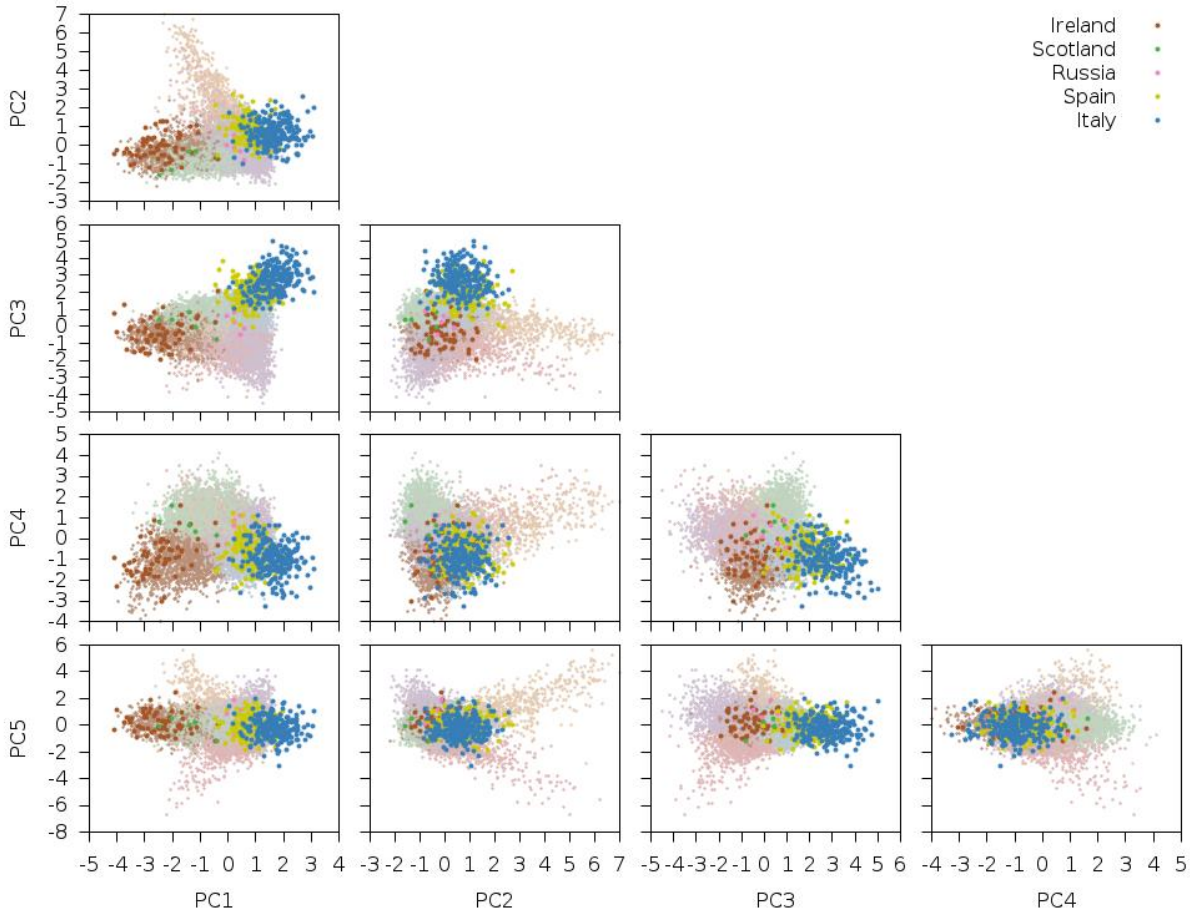

Figure S8 Results of PCA with projection of POPRES samples for all PCs

This set of plots is similar to Figure S3, except that POPRES samples are projected on top.

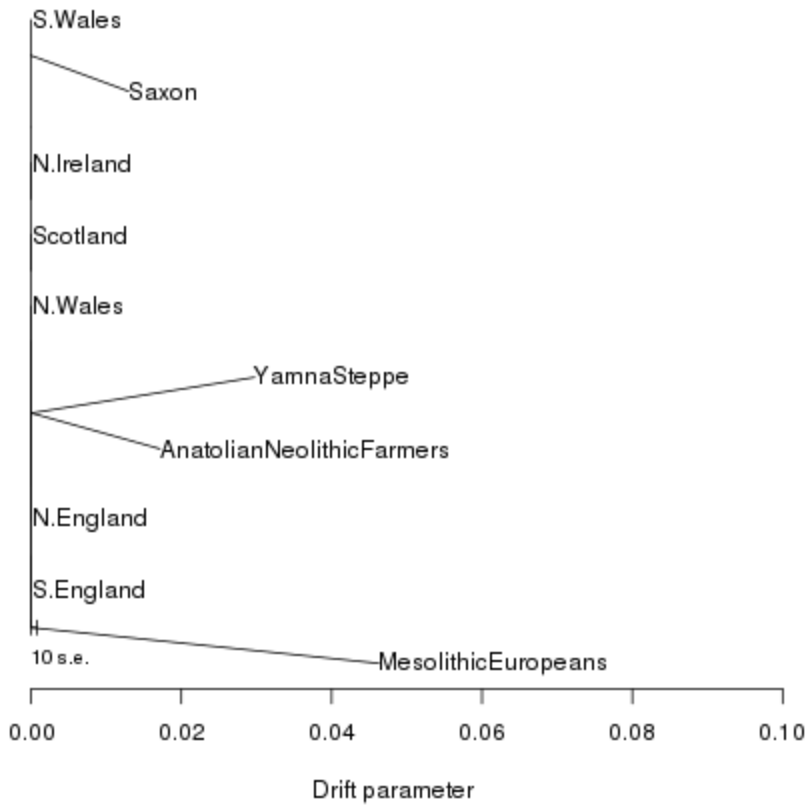

Figure S9 Tree-based clustering of UK Biobank clusters and ancient populations

We found that the ancient samples were too diverged from both the UK Biobank samples and from each other to form meaningful trees. Of note, the Mesolithich Europeans formed an outgroup and the Saxons were grouped with the south Wales cluster.

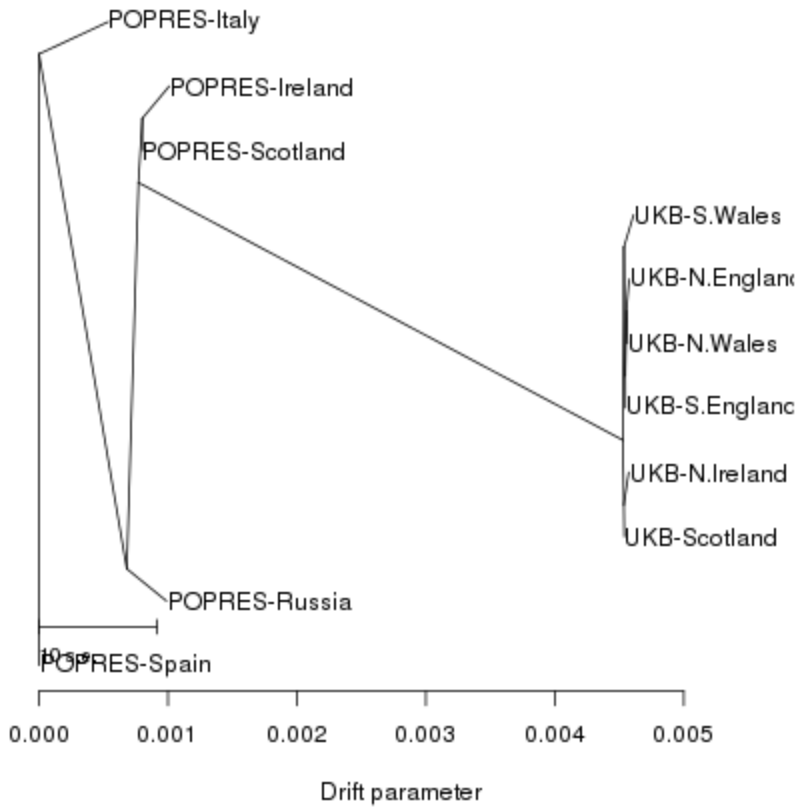

Figure S10 Tree-based clustering of UK Biobank clusters and POPRES populations

We see a bit of a batch effect differentiating the UK Biobank clusters from the POPRES populations. However, the UK Biobank samples do lie near the POPRES Irish and Italian samples. The Russian samples (which contain more Steppe ancestry) also cluster more closely with the UK Biobank samples than the Italian and Spanish samples.

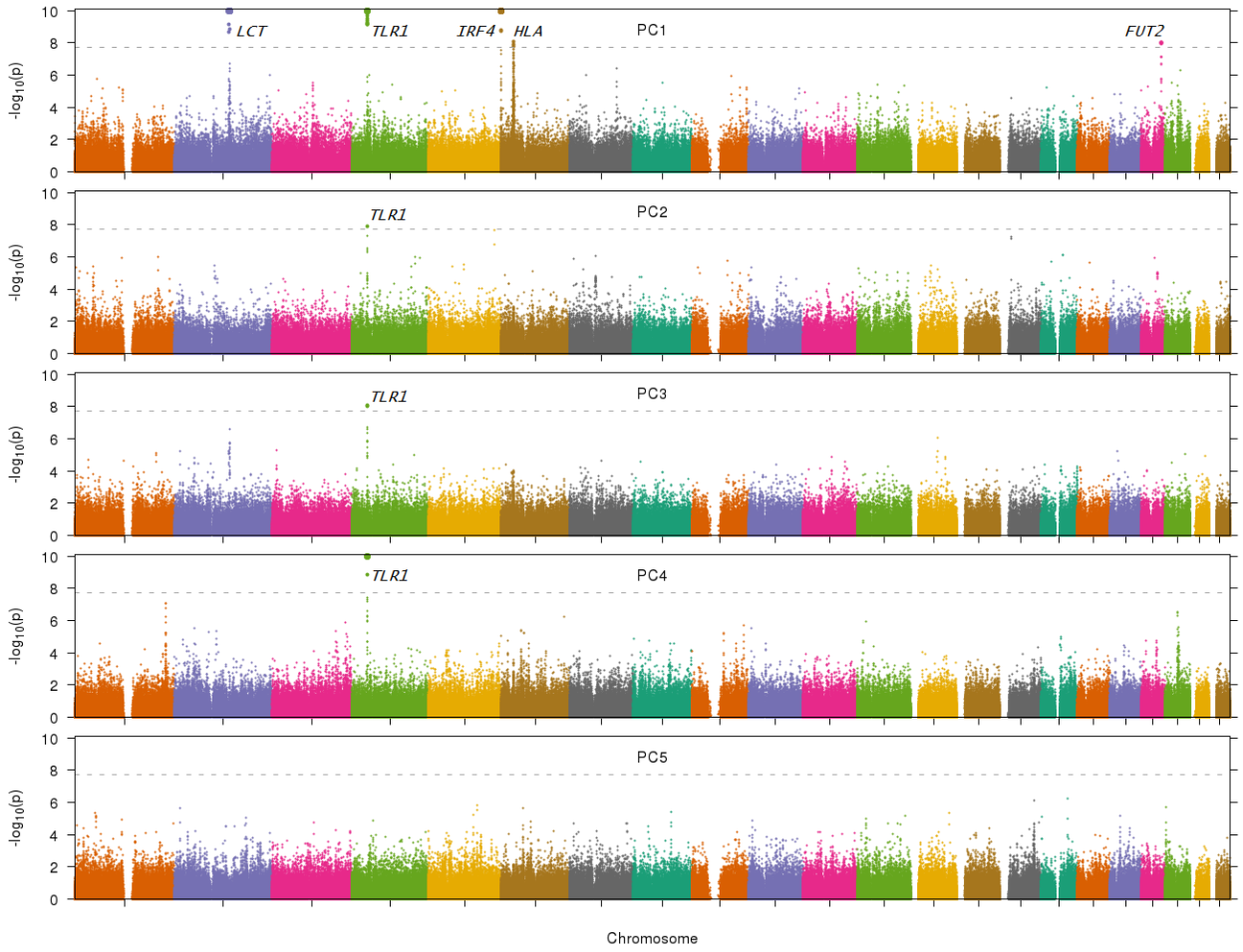

Figure S11 Selection statistic for UK Biobank along PC1-PC5

This is an expanded set of plots similar to Figure 4, except that plots for each of the top 5 PCs are displayed.

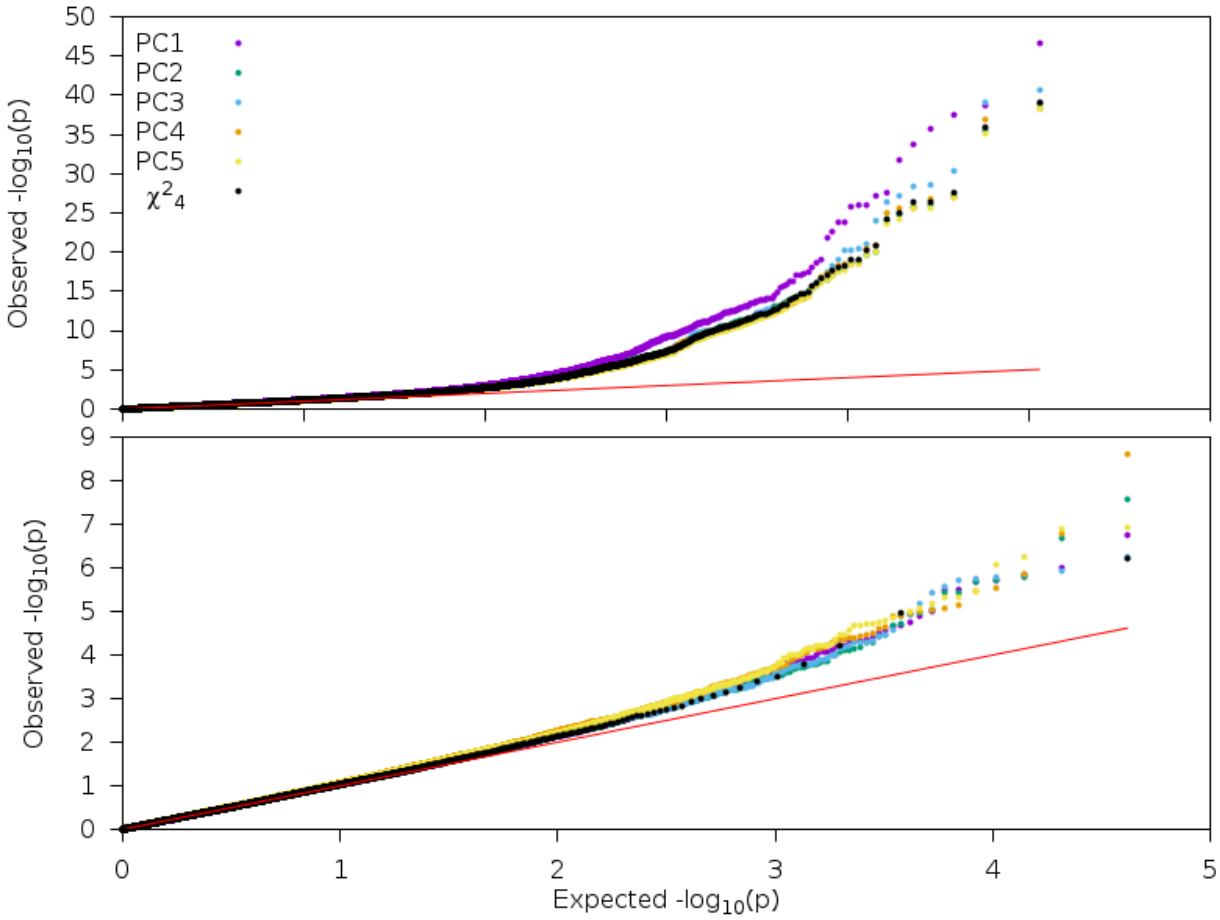

Figure S12 P-P plot of the combined selection statistic

We plot the observed $-\log_{10}(p)$ values compared to the expected distribution for the combined selection statistics as well as for the χ^2_4 statistic from Mathieson et al. The plot of all overlapping SNPs (top) suggests some inflation in the tails of the combined selection statistics, although λ_{GC} values were only slightly inflated (1.04-1.06; see Table S8) and this is largely due to the inflation of the χ^2_4 statistic from Mathieson et al. After LD-pruning by intersecting the SNPs with the PC dataset and removing two SNPs with $p < 5 \times 10^{-8}$ for the χ^2_4 statistic from Mathieson et al., we see much less inflation in the tails (λ_{GC} values were reduced to 1.01-1.03), indicating that the inflation in the tails was largely due to SNPs in LD with top hits from

Mathieson et al. In order to produce a maximally conservative statistic, we corrected our combined statistics by their λ_{GC} values (Table S8).

Supplementary Tables

Chrom	Locus (Mb)	PC	Best hit	*p*-value
1	2.2 - 2.2	3	rs79907870	4.68E-07
1	54.8 - 54.8	1	rs17390412	9.13E-07
1	56.0 - 56.1	8	rs1875068	1.86E-08
2	88.7 - 88.7	4	rs1713939	1.02E-07
2	133.3 - 144.0	1	rs7570971	7.21E-18
		4	rs1446585	3.02E-14
		7		3.09E-13
		10	rs72847650	<1e-50
2	159.8 - 160.0	7	rs1522699	3.89E-07
2	223.9 - 223.9	7	rs1900725	2.71E-07
3	46.3 - 46.4	7	rs9990343	2.80E-08
4	23.3 - 23.3	3	rs114557362	2.95E-07
4	38.7 - 38.9	1	rs4833095	9.29E-16
		3		8.54E-11
		4		1.21E-10
		7		2.18E-16
5	60.6 - 60.6	3	rs10471511	6.04E-07
5	101.5 - 101.6	8	rs411954	1.20E-08
5	114.8 - 114.8	8	rs895291	5.87E-07
5	164.8 - 164.9	3	rs77635680	6.70E-10
6	0.4 - 0.7	1	rs62389423	1.29E-47
		7		1.57E-09
6	23.9 - 36.7	1	rs151341075	3.76E-11
		3	rs2253908	5.43E-07
		4	rs151341075	3.35E-17
		5	rs3131618	<1e-50
		6	rs204999	<1e-50
		7	rs2596573	3.27E-54
		8	rs41268932	2.58E-23
		9	rs2596573	<1e-50
		10	rs9266258	4.14E-07
6	46.8 - 46.8	7	rs9395218	9.92E-07
6	86.0 - 87.0	10	rs2816583	2.23E-08
7	41.4 - 41.4	1	rs76920365	3.66E-07
7	64.4 - 66.4	3	rs79415723	1.17E-08
7	118.2 - 118.3	1	rs187417794	1.13E-07
8	7.2 - 12.7	2	rs11250099	<1e-50
9	14.0 - 14.0	3	rs12380860	4.45E-07
10	133.2 - 133.2	4	rs57105422	7.45E-07
15	28.4 - 28.4	3	rs12913832	9.17E-10
15	50.8 - 50.8	5	rs148783236	3.18E-194
		6		<1e-50
		8		2.18E-13
		9		<1e-50
16	9.5 - 9.5	4	rs12149526	6.26E-07
16	26.5 - 26.5	3	rs73528772	4.01E-07
16	53.7 - 53.7	3	rs61747071	1.36E-07
16	89.7 - 89.8	4	rs449882	5.37E-07
17	29.6 - 29.6	3	rs11655238	5.98E-07
19	33.8 - 33.8	3	rs41355649	3.65E-08
19	49.2 - 50.2	1	rs601338	1.05E-09
20	39.1 - 39.1	1	rs2143877	8.34E-08
22	32.9 - 32.9	5	rs115815765	7.40E-31
		6		<1e-50

Table S1 Significant or suggestive signals of selection in initial PCA run

We report significant or suggestive signals of selection in the initial PCA run. Neighboring SNPs <1Mb apart with genome-wide significant signals were grouped together into a single locus. The significant signals may represent either signals of selection or regions of long-range LD. All of these regions were removed from the main PCA run (see Methods).

			East-West		North-South	
	Eigenvalue	F_{ST}	**Correlation**	**p-value**	**Correlation**	**p-value**
PC1	20.99	1.76E-04	0.4154	<1e-50	-0.3981	<1e-50
PC2	9.35	7.33E-05	-0.0865	<1e-50	-0.4322	<1e-50
PC3	7.76	5.94E-05	0.1894	<1e-50	-0.1262	<1e-50
PC4	5.18	3.68E-05	-0.1418	<1e-50	0.3409	<1e-50
PC5	5.13	3.63E-05	0.0019	5.27E-01	-0.0124	4.14E-05
PC6	4.62	3.18E-05	-0.0163	6.84E-08	0.0150	6.93E-07
PC7	4.61	3.17E-05	-0.0025	4.01E-01	0.0049	1.04E-01
PC8	4.59	3.15E-05	0.0216	7.75E-13	0.0047	1.16E-01
PC9	4.59	3.15E-05	-0.0522	<1e-50	-0.0119	7.92E-05
PC10	4.57	3.14E-05	-0.0143	2.23E-06	0.0121	6.47E-05

Table S2 PC eigenvalues and geographical correlations

PC1-PC5 all had elevated eigenvalues, while PC6-PC10 had eigenvalues which were close to background levels. In the case where there are two equal-sized sample sets from distinct populations, the F_{ST} between the two populations can be estimated from the top eigenvalue (λ) via the following formula: $F_{ST} = (\lambda - 1)/N$, where N is the total number of samples. The top eigenvalue reflects an F_{ST} of 1.76×10^{-4}, indicating very subtle population structure within the UK. PC1 was most strongly correlated with east-west birth coordinate and PC2 was most strongly correlated with north-south birth coordinate.

		Pop2				
Grouping	***Pop1***	Norfolk	Suffolk	Hampshire	Kent	Devon
Saxon	Saxon	5.118	5.268	2.543	3.953	3.32
Scotland	Argyll and Bute	9.560	9.370	3.323	6.411	6.223
	Banff and Buchan	7.609	77.545	1.234	4.440	4.379
	Orkney	11.229	10.583	3.620	7.310	7.259
N. Wales	North Wales	8.490	8.393	1.918	5.163	5.239
S. Wales	North Pembrokeshire	7.124	7.287	1.759	4.542	4.430
	South Pembrokeshire	6.301	6.189	2.315	4.336	4.171

Table S3 Expanded results of f_4 statistics in ancient and modern British samples

We report f_4 statistics of the form $f_4(Steppe, Neolithic\ Farmer; Pop1, Pop2)$, representing a z-score with positive values indicating more Steppe ancestry in *Pop1* than *Pop2*. Samples for *Pop1* were either modern Celtic (Scotland and Wales) or ancient Saxon. Samples for *Pop2* were modern Anglo-Saxon (southern and eastern England).

Annotation	Chromosome	Locus (Mb)	PC	Best hit	p-value
-	1	208.8 - 208.8	2	rs75602597	9.71e-07
ABCD3	1	226.4 - 226.8	4	rs72759068	8.62e-08
-	4	45.2 - 45.2	1	rs77147311	9.78e-07
-	5	164.8 - 164.9	2	rs77635680	2.13e-08
ZDHHC14	6	158.1 - 158.1	4	rs73584091	5.46e-07
-	7	64.9 - 64.9	2	rs79415723	8.81e-07
-	7	118.2 - 118.2	1	rs187417794	3.66e-07
-	13	66.2 - 66.2	3	rs1417218	8.38e-07
OCA2[1,2]	15	28.4 - 28.4	2	rs12913832	5.55e-08
RPGRIP1L	16	53.7 - 53.7	2	rs61747071	7.81e-07
BPIFB9P	20	31.8 - 32.0	4	rs293709	3.00e-07
-	20	39.1 - 39.1	1	rs2143877	5.03e-07

Table S4 Suggestive signals of selection in UK Biobank

We report the top signal of natural selection for each locus not reaching genome-wide significance ($p > 1.96 \times 10^{-8}$) but yielding a suggestive signal ($p < 1.00 \times 10^{-6}$) along any of the top five PCs. Neighboring SNPs <1Mb apart with suggestive significant signals were grouped together into a single locus.

Dataset	Cluster	rs601338 (G/A)	rs492602 (G/A)	rs676388 (T/C)
UK Biobank	Northern Ireland	0.4406	0.4413	0.4207
	Northern England	0.4633	0.4638	0.4407
	Pembrokeshire	0.4864	0.4871	0.4580
	North Wales	0.5006	0.501	0.4781
	Yorkshire	0.5025	0.503	0.4763
	Southern England	0.5109	0.5111	0.4847
GERA	Irish	-	0.4754	0.4522
	Northern European	-	0.5215	0.4962
	Southern European	-	0.5248	0.5021
	Ashkenazi Jewish	-	0.5530	0.5200
	Eastern European	-	0.5840	0.5586
PoBI	Argyll and Bute	-	-	0.2949
	North Pembrokeshire	-	-	0.3171
	Banff and Buchan	-	-	0.3365
	Northern Ireland	-	-	0.3667
	Cumbria	-	-	0.4039
	Derbyshire	-	-	0.4091
	Dorset	-	-	0.4125
	Herefordshire	-	-	0.4259
	Worcestershire	-	-	0.4265
	Lancashire	-	-	0.4306
	Devon	-	-	0.4416
	Yorkshire	-	-	0.4517
	Orkney	-	-	0.4531
	Lincolnshire	-	-	0.4619
	Kent	-	-	0.4661
	Suffolk	-	-	0.4699
	Cornwall	-	-	0.4716
	Leicestershire	-	-	0.4726
	North Wales	-	-	0.4737
	Northeast England	-	-	0.4844
	Cheshire	-	-	0.4875
	Forest of Dean	-	-	0.4881
	Norfolk	-	-	0.4951
	Nottinghamshire	-	-	0.5000
	Northamptonshire	-	-	0.5000
	Oxfordshire	-	-	0.5054
	Sussex	-	-	0.5167
	Gloucestershire	-	-	0.5417
	South Pembrokeshire	-	-	0.5417
	Hampshire	-	-	0.5833

Table S5 Allele frequency of FUT2 alleles

We report the allele frequency of the most significant hit, rs601338, along with two other linked SNPs in GERA and the PoBI datasets.

Population 1	Population 2	F_{ST}	rs601338	rs492602	rs676388
N. England	S. England	7.36E-05	2.22E-01	2.29E-01	2.13E-01
	N. Ireland	2.96E-04	1.28E-06	1.42E-06	1.31E-05
	Scotland	1.48E-04	2.38E-05	2.50E-05	1.25E-04
	N. Wales	8.53E-05	7.94E-01	7.99E-01	8.15E-01
	S. Wales	3.75E-04	2.77E-01	2.85E-01	2.17E-01
S. England	N. Ireland	2.67E-04	**6.04E-09**	7.35E-09	1.07E-07
	Scotland	1.10E-04	**2.61E-09**	**3.11E-09**	3.38E-08
	N. Wales	7.46E-05	1.42E-01	6.86E-02	3.41E-01
	S. Wales	2.90E-04	6.45E-02	6.86E-02	4.27E-02
N. Ireland	Scotland	1.22E-04	9.31E-03	9.86E-03	2.12E-02
	N. Wales	2.23E-04	1.42E-07	1.58E-07	4.45E-07
	S. Wales	3.43E-04	1.47E-03	1.48E-03	9.38E-03
Scotland	N. Wales	1.23E-04	1.95E-05	2.00E-05	1.81E-05
	S. Wales	3.10E-04	9.12E-02	8.93E-02	2.06E-01
N. Wales	S. Wales	2.60E-04	2.72E-01	2.78E-01	1.18E-01

Table S6 Discrete test for natural selection at *FUT2* in UK Biobank

We report results of tests for selection using discrete subpopulations for *FUT2* in the UK Biobank data set, using the UK Biobank subpopulations derived from *k*-means clustering. With 15 comparisons per SNP and 510,665 SNPs (*p*-value threshold of 6.53×10^{-9}), we are still able to find genome-wide-significant results when comparing southern England with Scotland as well as Northern Ireland (emphasized in bold).

Population 1	Population 2	F_{ST}	rs492602	rs676388
Ashkenazi Jewish	Eastern European	6.84E-03	5.96E-01	5.13E-01
	Irish	6.71E-03	1.84E-01	2.45E-01
	Northern European	6.54E-03	5.83E-01	6.79E-01
	Southeast European	3.45E-03	5.05E-01	6.73E-01
Eastern European	Irish	9.44E-04	***1.48E-06***	***2.49E-06***
	Northern European	7.23E-04	1.58E-03	1.69E-03
	Southeast European	2.39E-03	9.25E-02	1.10E-01
Irish	Northern European	1.26E-04	***1.34E-07***	***4.53E-07***
	Southeast European	1.91E-03	1.16E-01	1.12E-01
Northern European	Southeast European	1.80E-03	9.12E-01	8.46E-01

Table S7 Discrete test for natural selection at *FUT2* in GERA

We report results of tests for selection using discrete subpopulations for *FUT2* in the GERA data set. *FUT2* does not reach genome-wide significance in the GERA dataset, however there are several suggestive signals when comparing the "Irish" subgroup with the Northern European and Eastern European subpopulation (emphasized in bold italic).

	Inflation			
	Genome-wide	**Overlap**	**Combined**	**Correlation**
Ancient	1.00	1.07	-	-
UKB PC1	1.02	1.08	1.06	18.8%
UKB PC2	0.95	1.00	1.05	2.8%
UKB PC3	0.95	1.00	1.05	6.5%
UKB PC4	0.88	0.94	1.04	2.5%
UKB PC5	0.86	0.91	1.04	0.0%

Table S8 Independence of UK Biobank and ancient Eurasian scans for selection

The UK Biobank and ancient Eurasian selection statistics were not substantially inflated genome-wide, nor at the overlapping SNPs in both datasets. Similarly, the combined selection statistics were not substantially inflated. The correlation between the two statistics is also small, with the UK Biobank PC1 and ancient Eurasian statistics being most correlated with $r = 0.188$.

Locus	Chr	Locus (Mb)	Phenotype	Top SNP	*p*-value
LCT	2	134.9 - 137.0	lung_FVCzSMOKE	rs6716536	4.90E-10
TLR1	4	38.8 - 38.9	disease_ALLERGY_ECZEMA_DIAGNOSED	rs5743614	5.80E-26
SLC22A4	5	131.4 - 131.8	body_HEIGHTz	rs1050152	3.70E-18
			disease_ASTHMA_DIAGNOSED	rs2188962	6.00E-09
F12	5	176.8 - 176.8	body_HEIGHTz	rs2545801	4.80E-11
HLA	6	28.3 - 33.0	body_HEIGHTz	rs2256183	4.10E-23
			body_WHRadjBMIz	rs521977	1.80E-10
			bp_DIASTOLICadjMEDz		2.00E-10
			disease_ALLERGY_ECZEMA_DIAGNOSED	rs3135377	1.60E-11
			disease_ASTHMA_DIAGNOSED	rs204993	5.10E-09
			lung_FEV1FVCzSMOKE	rs3891175	3.50E-21
			lung_FVCzSMOKE	rs6456834	3.70E-09
FADS1	11	61.5 - 61.6	bmd_HEEL_TSCOREz	rs174548	1.20E-08
ATXN2/ SH2B3	12	111.9 - 112.9	bp_DIASTOLICadjMEDz	rs3184504	8.00E-33
			bp_SYSTOLICadjMEDz		1.90E-13
			disease_HYPERTENSION_DIAGNOSED		1.30E-09
CYP1A2/ CSK	15	75.0 - 75.1	bp_DIASTOLICadjMEDz	rs2472304	1.10E-19
			bp_SYSTOLICadjMEDz		4.20E-10
			disease_HYPERTENSION_DIAGNOSED		2.60E-09

Table S9 Phenotype associations at SNPs with signals of selection

We tested SNPs with genome-wide significant signals of selection in the constituent UK Biobank or ancient Eurasian scans or the combined scan for association with 15 phenotypes in the UK Biobank data set, using the top 5 PCs as covariates.

Phenotype	PC1	PC2	PC3	PC4	PC5
bmd_HEEL_TSCOREz	2.33E-01	4.75E-01	2.39E-23	3.54E-07	2.97E-01
body_BMIz	2.36E-27	7.98E-01	2.59E-11	2.21E-03	1.14E-01
body_HEIGHTz	<1e-50	4.66E-01	<1e-50	8.38E-03	2.41E-03
body_WHRadjBMIz	2.75E-06	3.35E-01	3.71E-03	2.94E-24	1.06E-01
bp_DIASTOLICadjMEDz	7.34E-01	5.42E-01	2.70E-08	6.16E-13	2.37E-01
bp_SYSTOLICadjMEDz	1.75E-03	7.67E-02	6.31E-01	6.15E-07	3.09E-02
cov_EDU_COLLEGE	6.73E-01	9.75E-25	2.01E-29	8.59E-36	7.96E-07
cov_SMOKING_STATUS	7.13E-01	5.67E-01	3.61E-03	9.63E-07	8.08E-01
disease_ALLERGY_ECZEMA_DIAGNOSED	1.76E-16	1.50E-03	1.07E-08	7.15E-03	7.87E-01
disease_ASTHMA_DIAGNOSED	7.04E-01	2.50E-06	6.12E-01	2.89E-05	3.84E-01
disease_HYPERTENSION_DIAGNOSED	7.84E-01	2.45E-01	1.09E-03	5.92E-01	3.32E-01
lung_FEV1FVCzSMOKE	9.85E-01	3.31E-07	1.89E-13	5.58E-10	3.43E-07
lung_FVCzSMOKE	8.69E-02	1.93E-45	2.21E-03	3.48E-40	3.99E-04
repro_MENARCHE_AGE	1.14E-04	1.11E-05	1.93E-03	3.11E-01	3.75E-02
repro_MENOPAUSE_AGE	1.46E-15	7.51E-04	9.54E-07	1.27E-03	1.93E-01

Table S10 PC-phenotype associations in UK Biobank

We report the results of tests of associations (*p*-value) between top PCs and 15 phenotypes in UK Biobank. This analysis does not distinguish between environmental and genetic effects.

References

1. Voight, B.F., Kudaravalli, S., Wen, X., and Pritchard, J.K. (2006). A Map of Recent Positive Selection in the Human Genome. PLoS Biol *4*, e72.

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