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

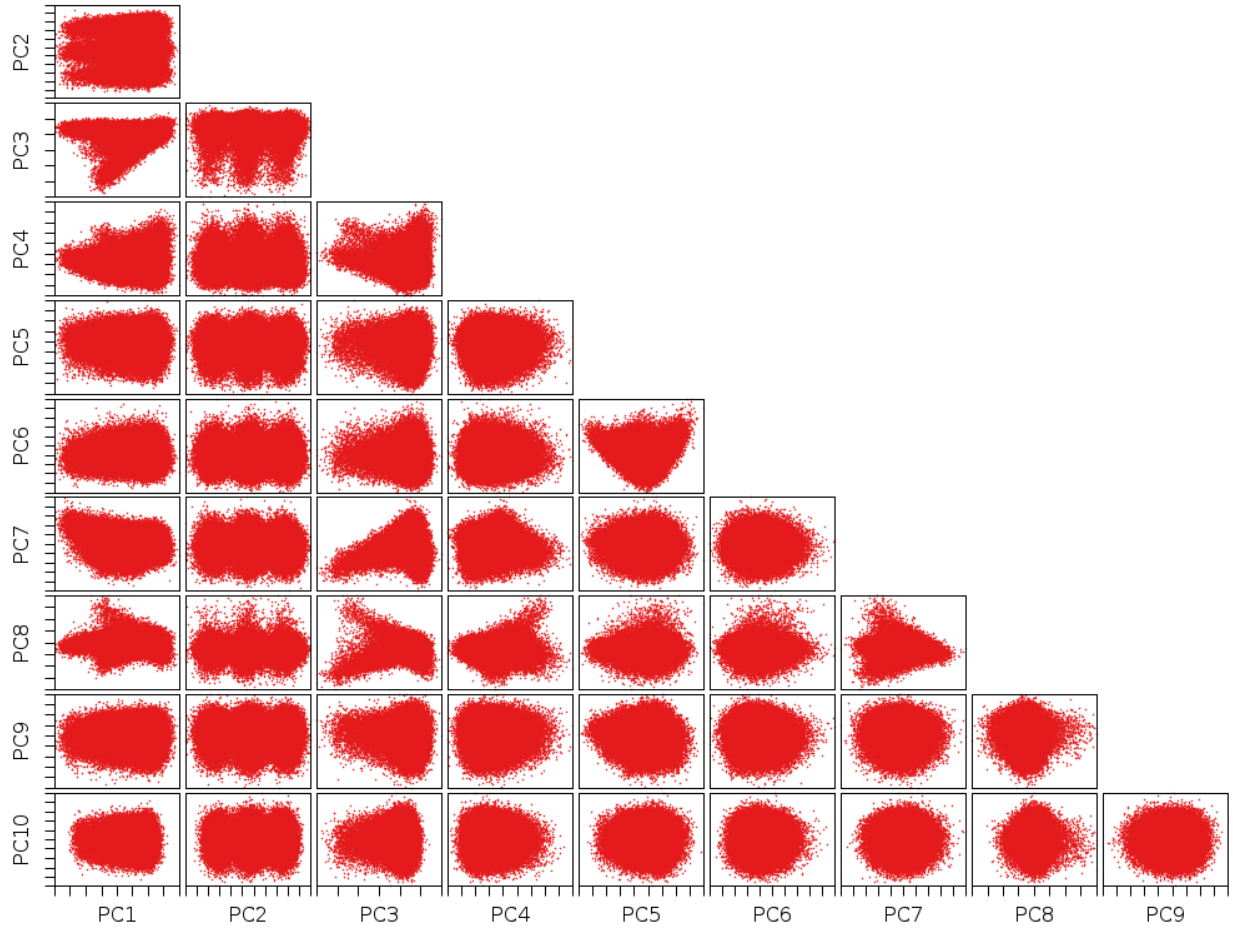

Figure S1 Results of initial PCA run

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.

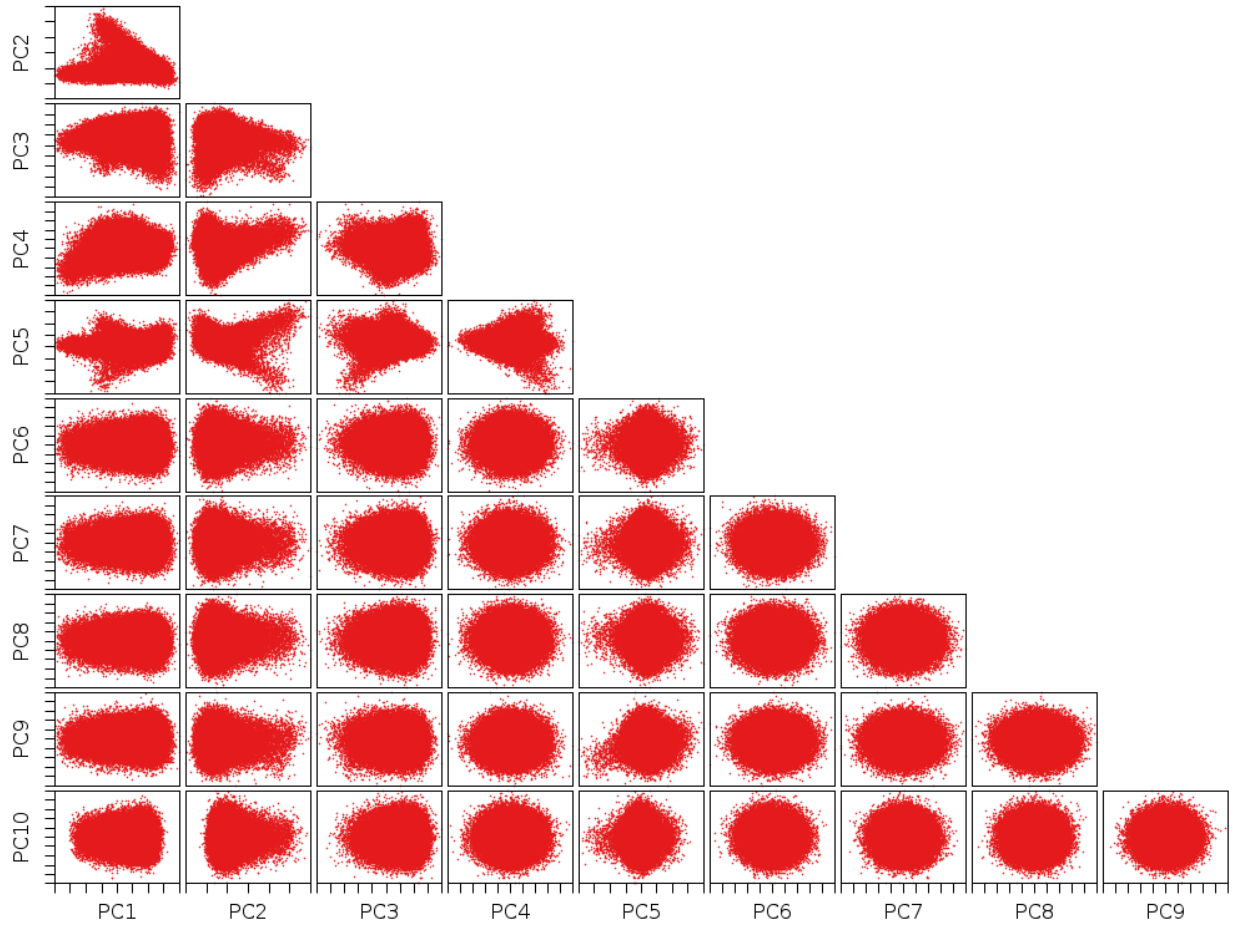

Figure S2 Results of PCA after removing long-range LD regions

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.

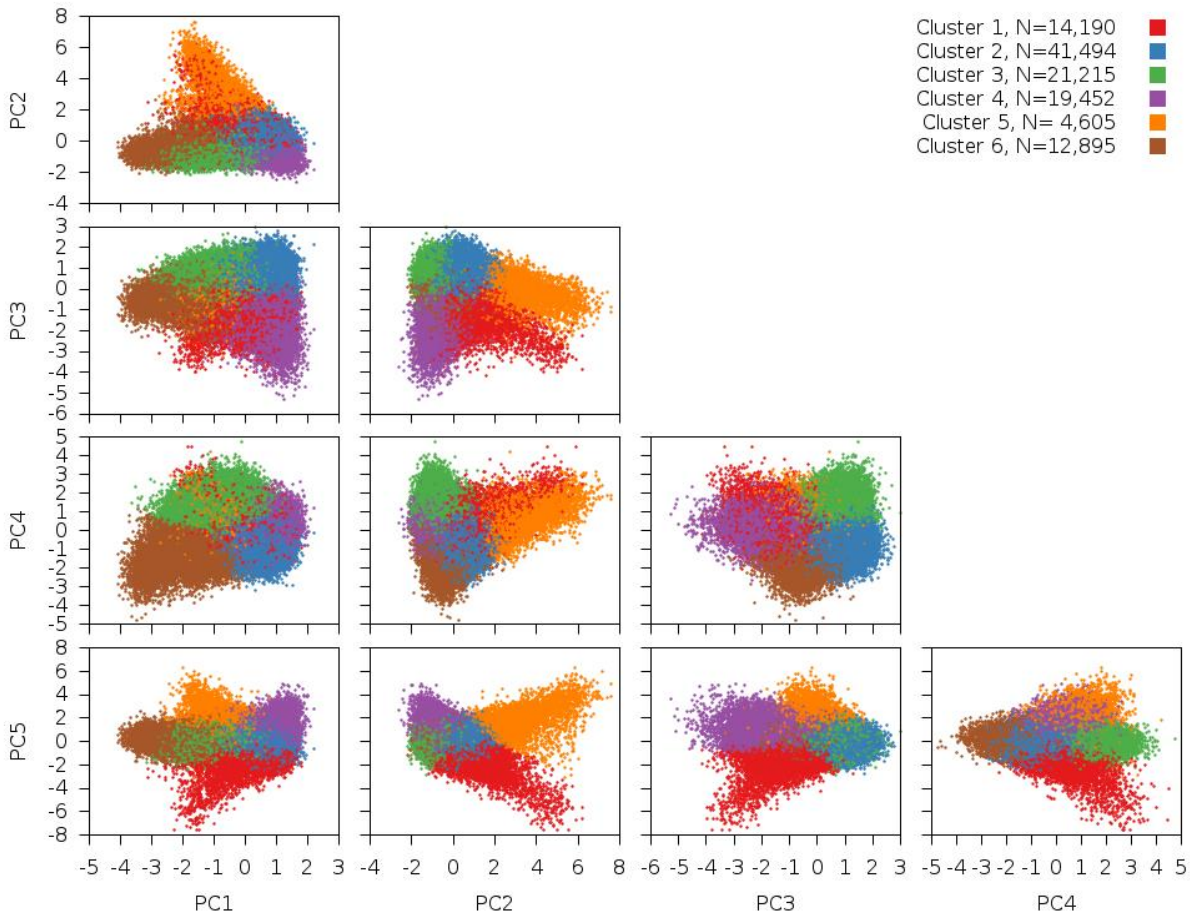

Figure S3 Results of PCA with *k*-means clustering for all PCs

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

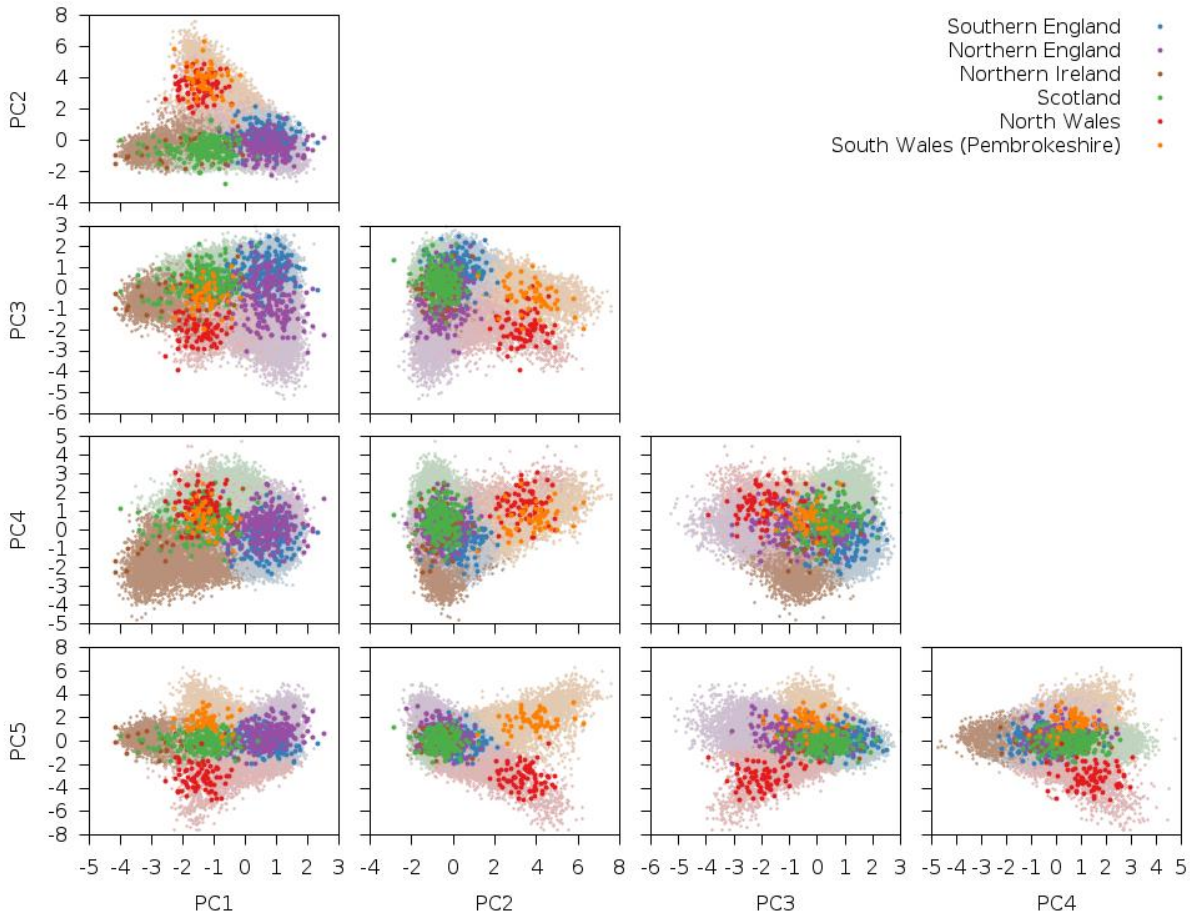

Figure S4 Results of PCA with projection of PoBI samples for all PCs

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

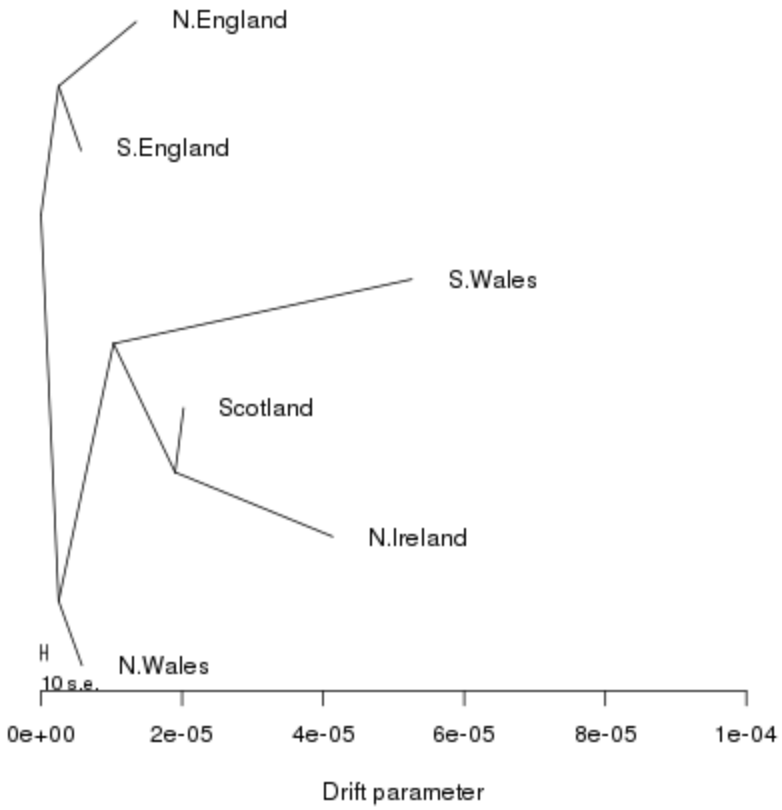

Figure S5 Tree-based clustering of UK Biobank subpopulation clusters

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).

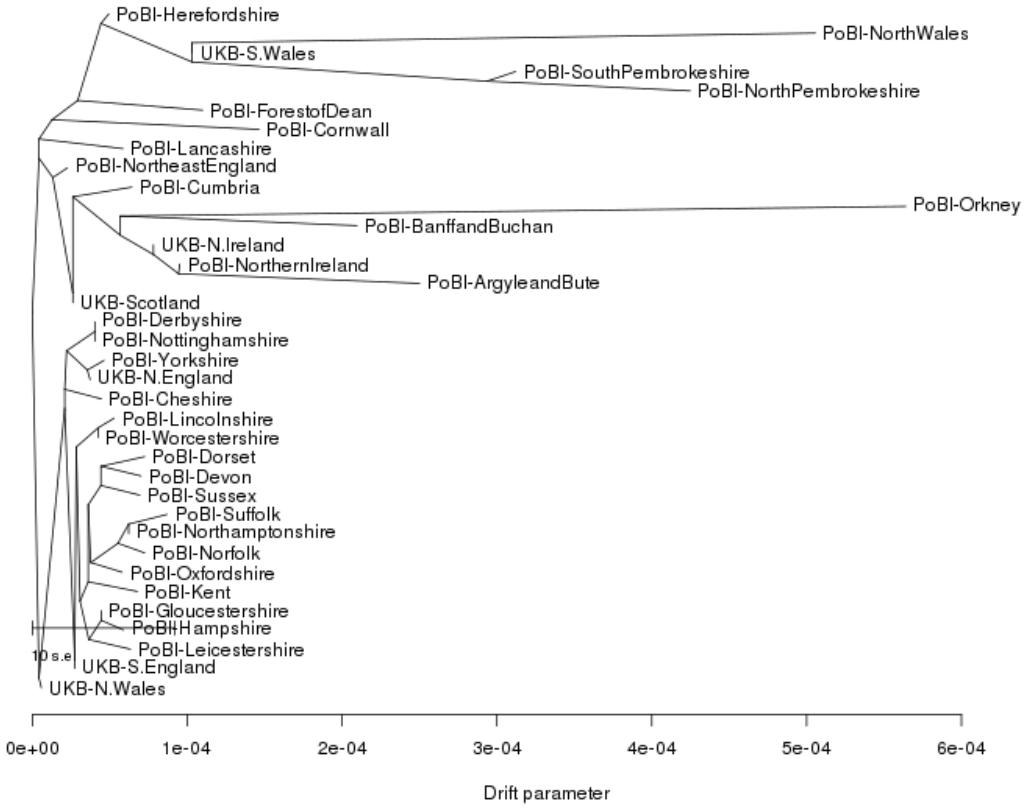

Figure S6 Tree-based clustering of UK Biobank clusters and PoBI populations

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).

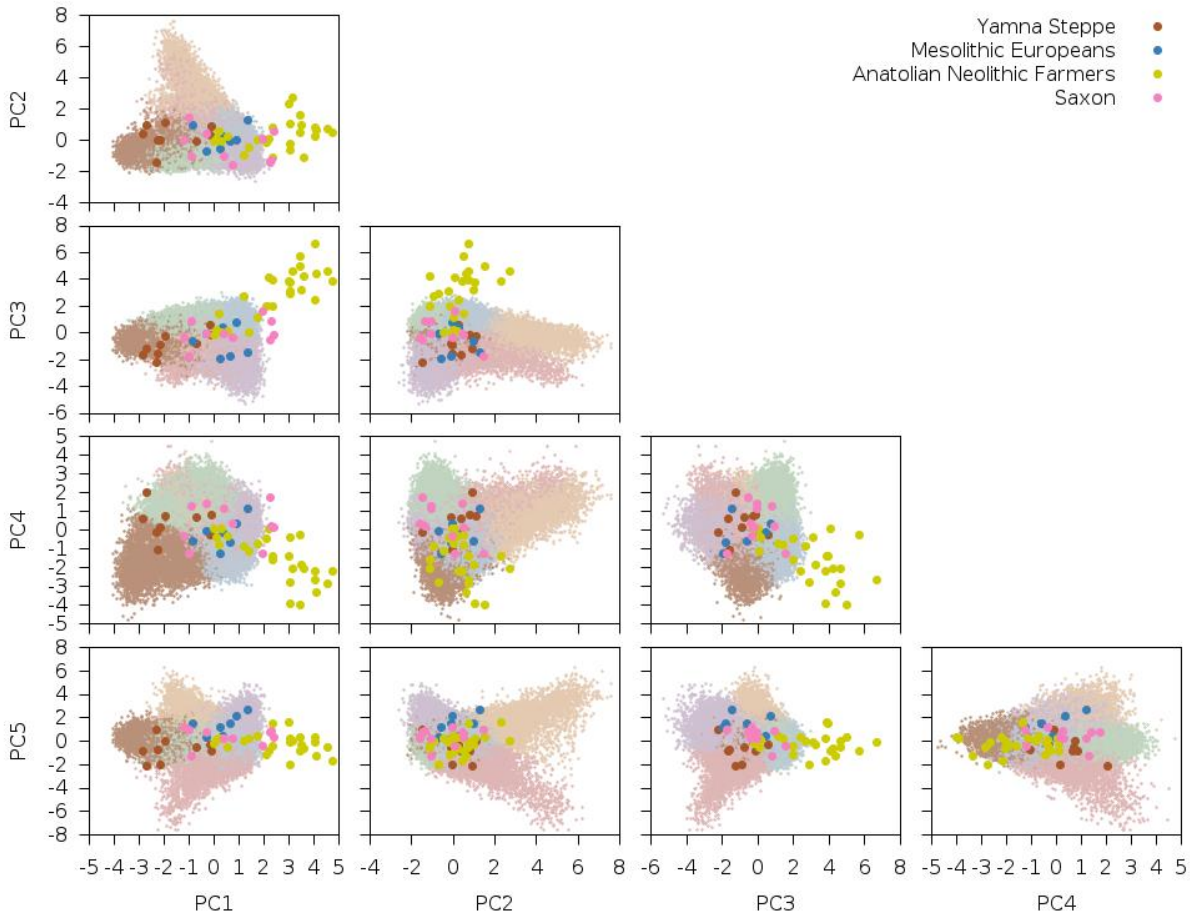

Figure S7 Results of PCA with projection of ancient samples for all PCs

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

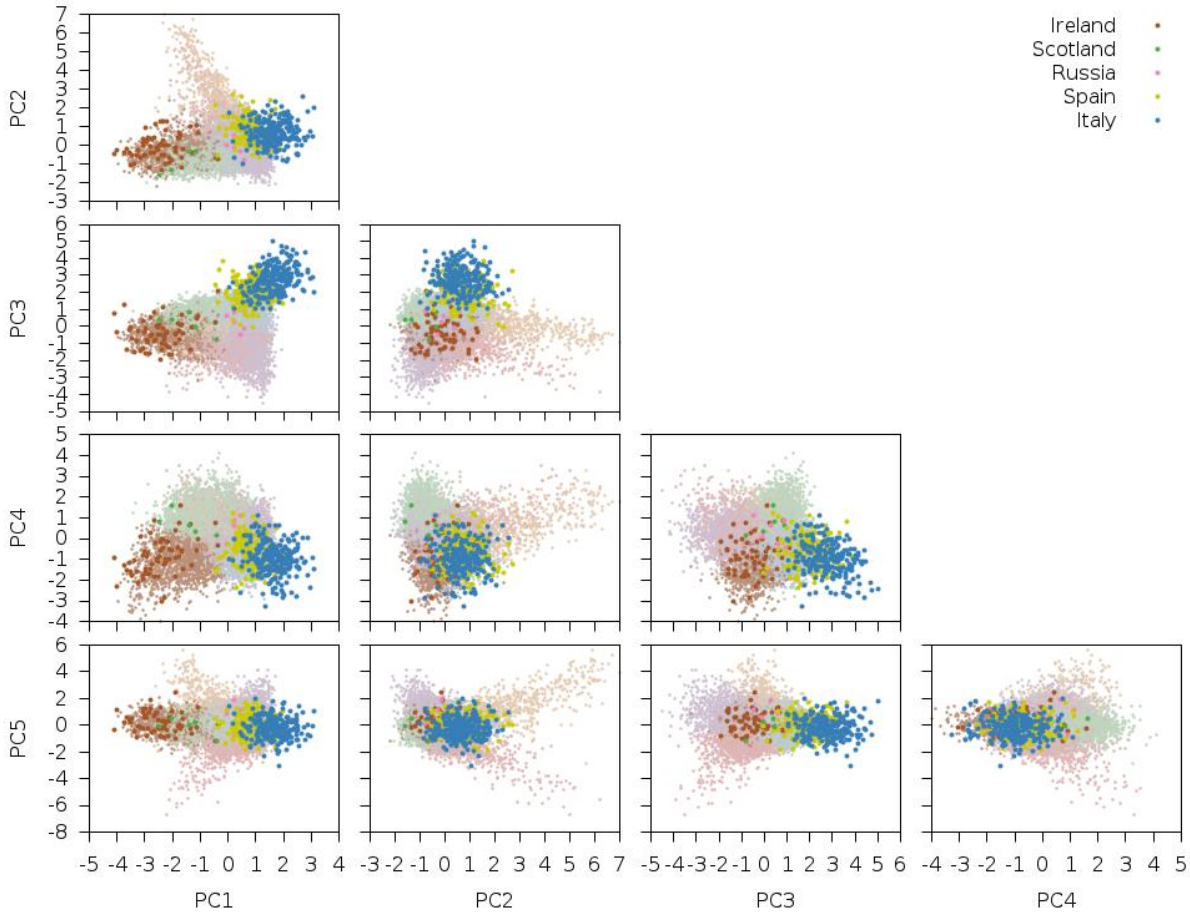

Figure S8 Results of PCA with projection of POPRES samples for all PCs

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

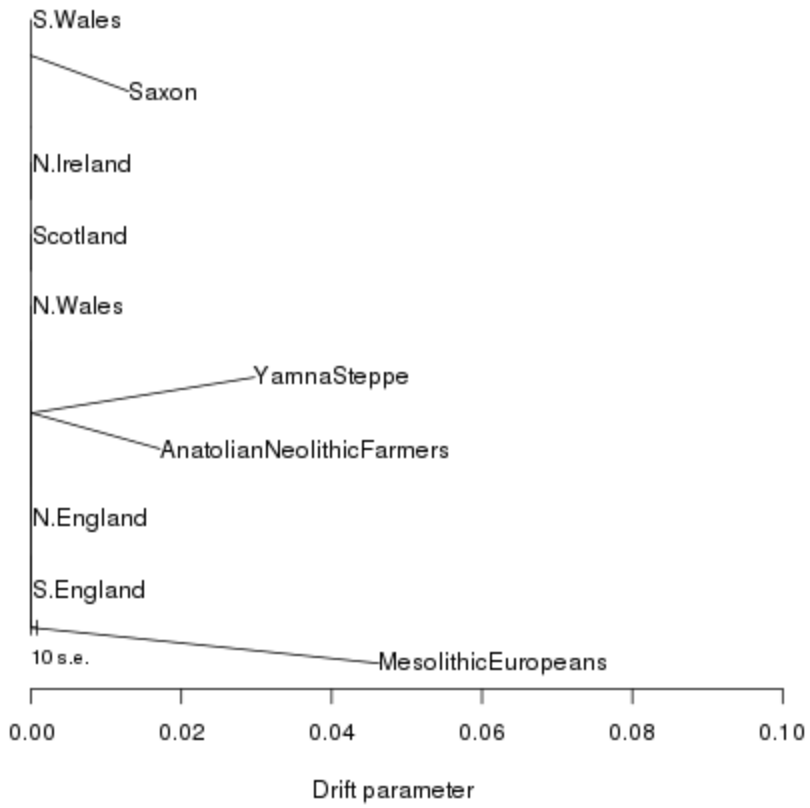

Figure S9 Tree-based clustering of UK Biobank clusters and ancient populations

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.

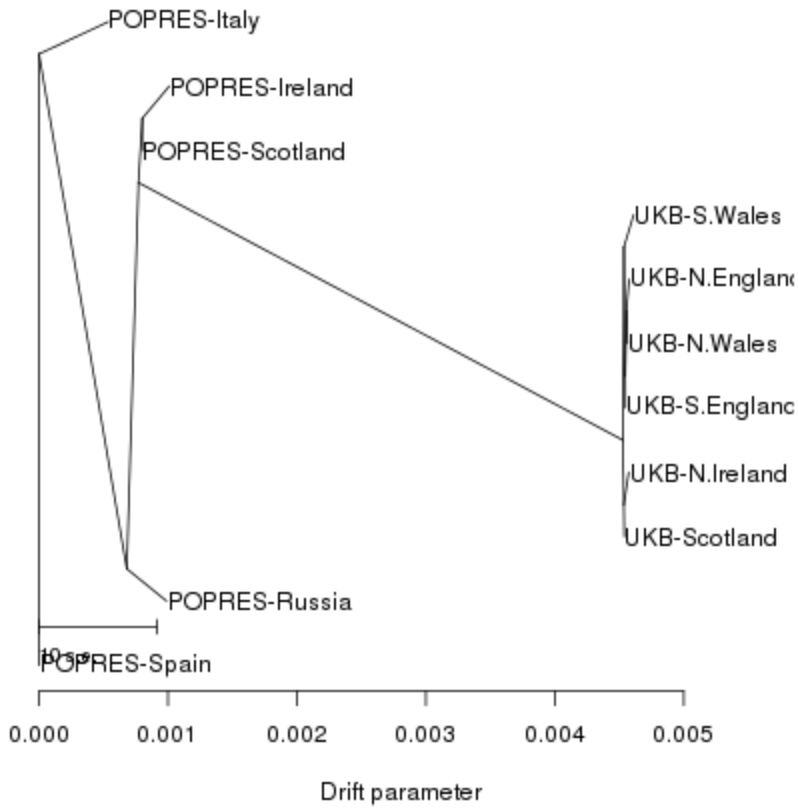

Figure S10 Tree-based clustering of UK Biobank clusters and POPRES populations

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.

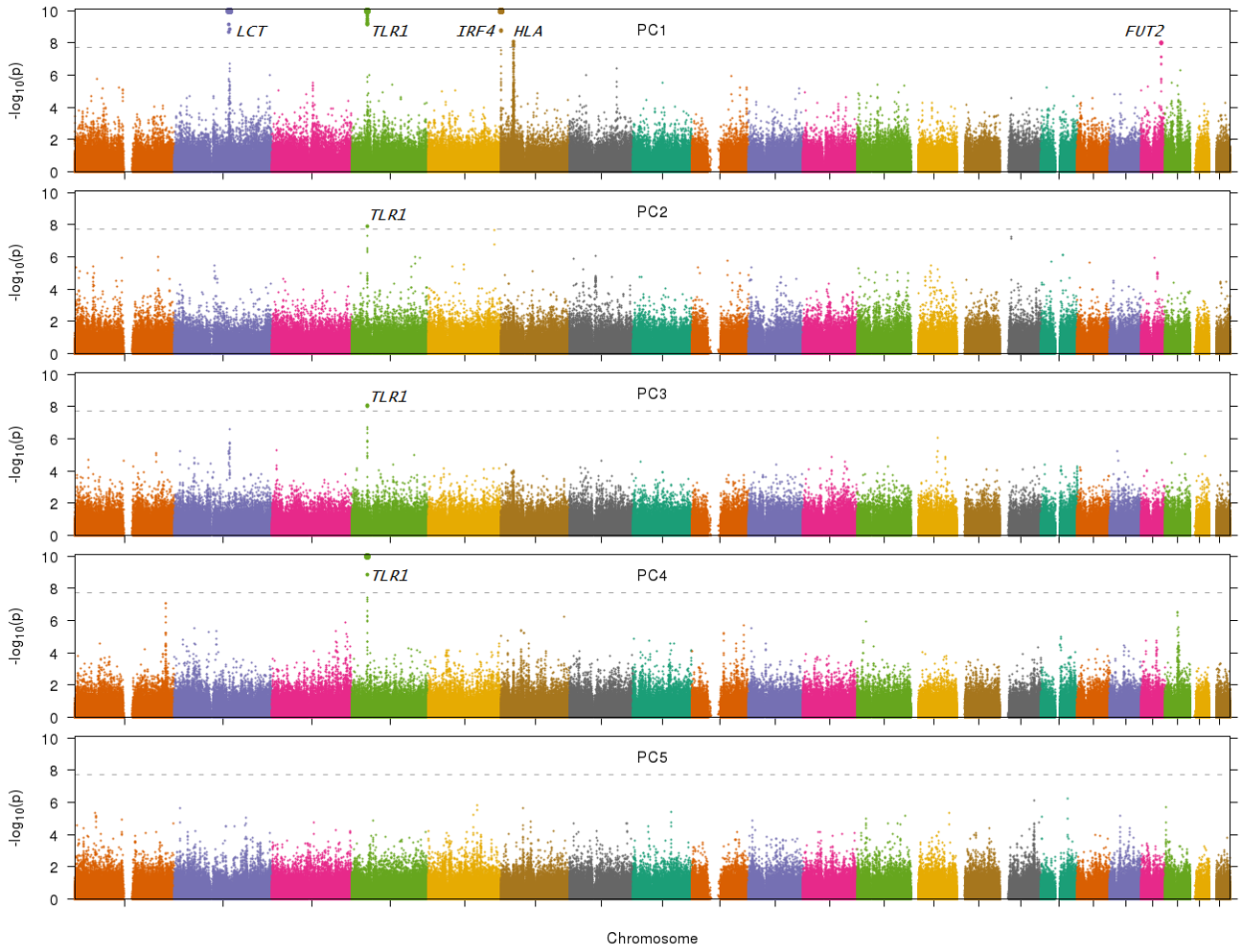

Figure S11 Selection statistic for UK Biobank along PC1-PC5

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

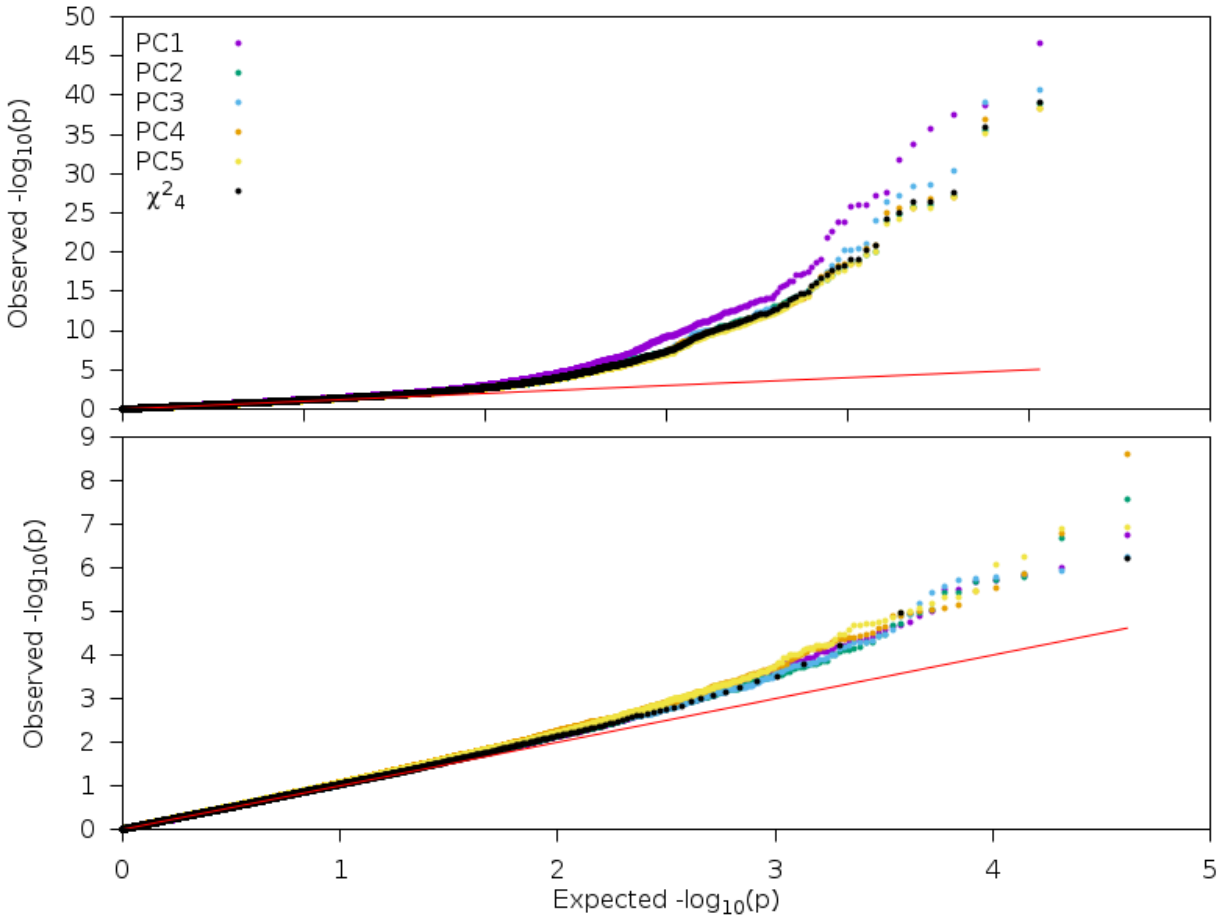

Figure S12 P-P plot of the combined selection statistic

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.