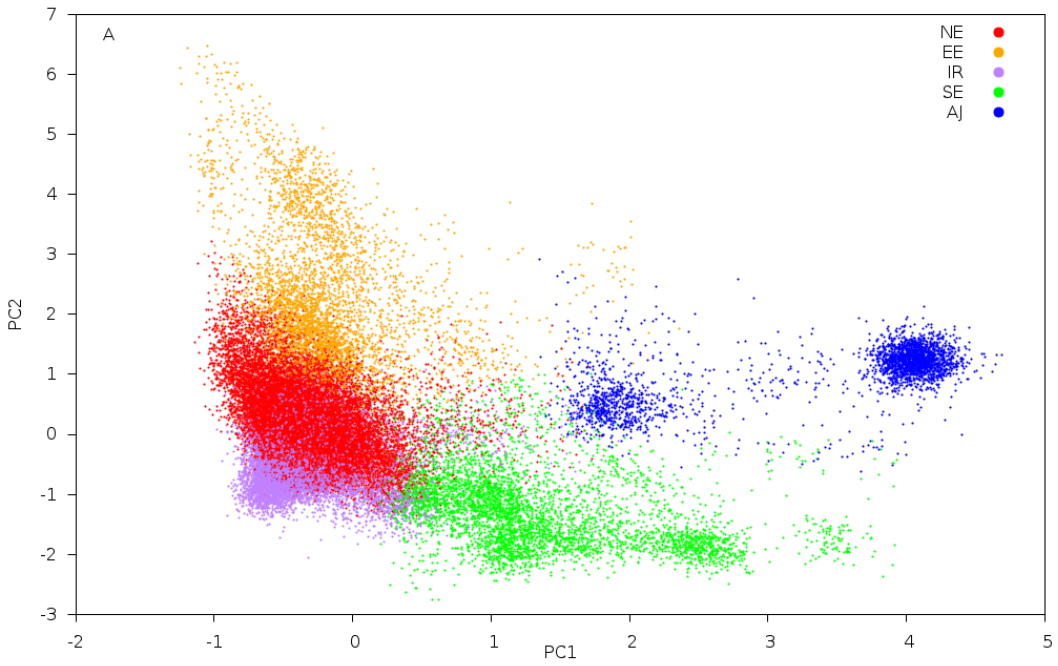
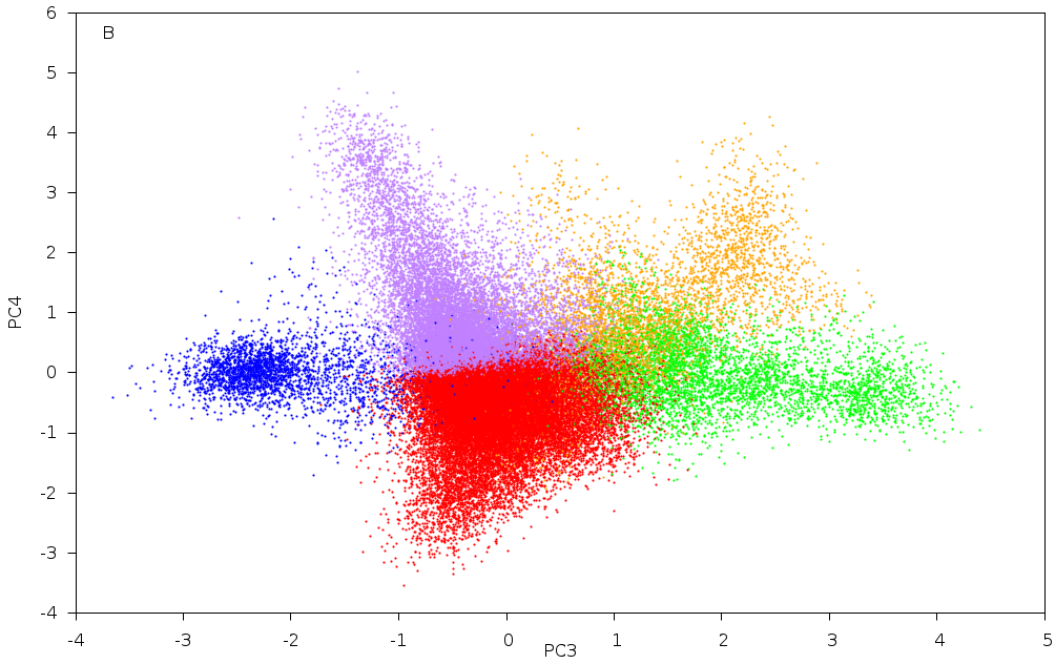
Supplementary Figures

Supplementary Figure 1. k-Means clustering confirms visually-observed subpopulations.

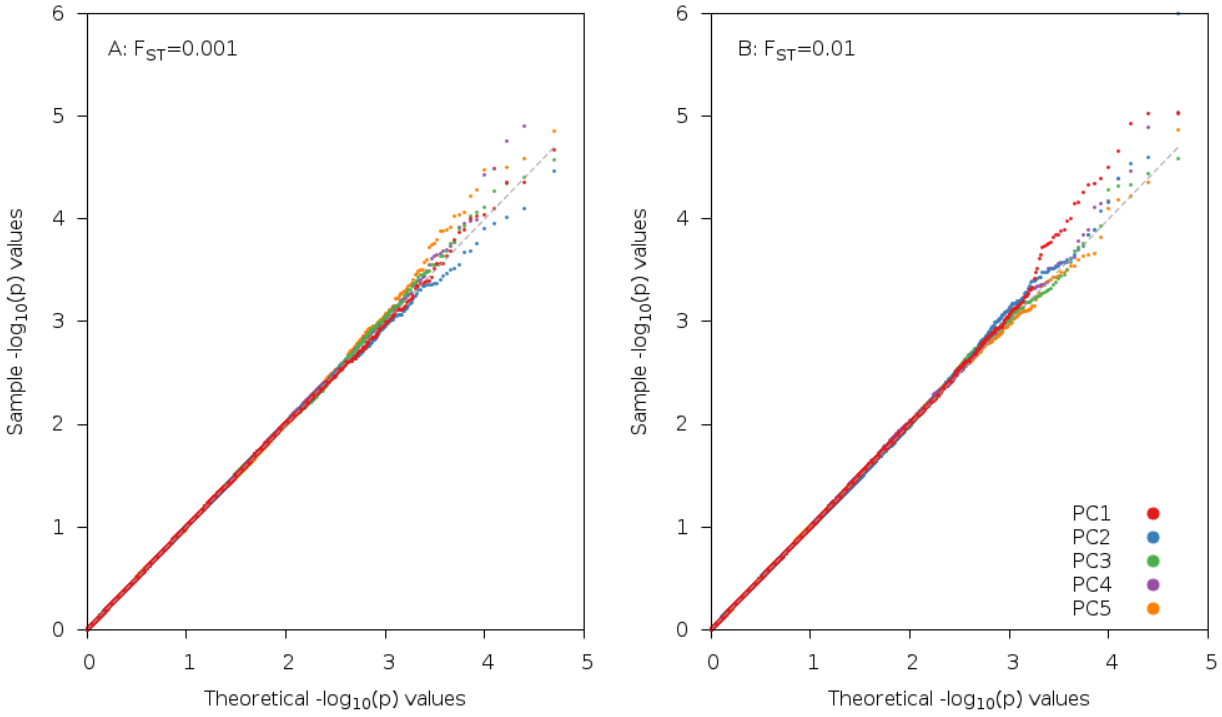
Individuals were clustered using k-means clustering with $k = 5$ on the top 4 PCs. 5 clusters were the minimum number of clusters that produced results consistent between runs. Clusters were labeled and assigned colors based upon where they fell relative to predicted fractional ancestry and where projected populations lay.



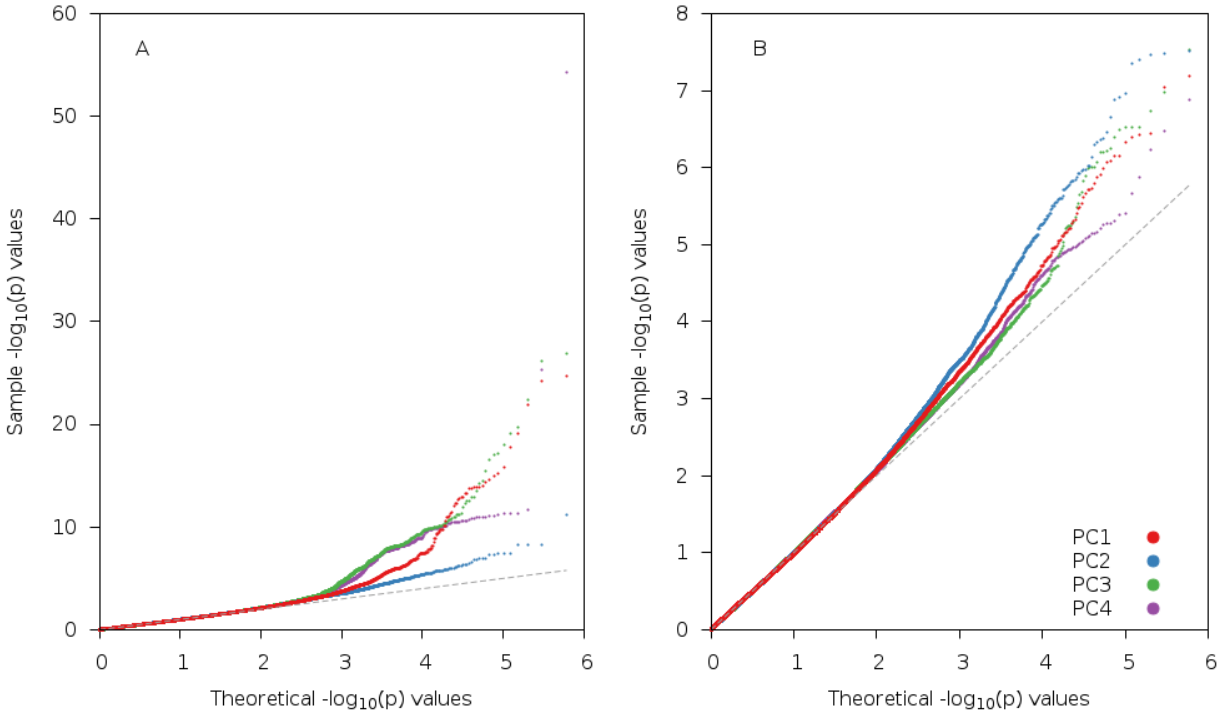


Supplementary Figure 2. QQ-plot of the selection statistic in null simulations.

The selection statistic was generated for null simulations containing 6 populations and differing by $F_{ST} = 0.001$ and $F_{ST} = 0.01$. The p-values of the selection statistic for the first PC did not significantly deviate from the distribution expected under null.

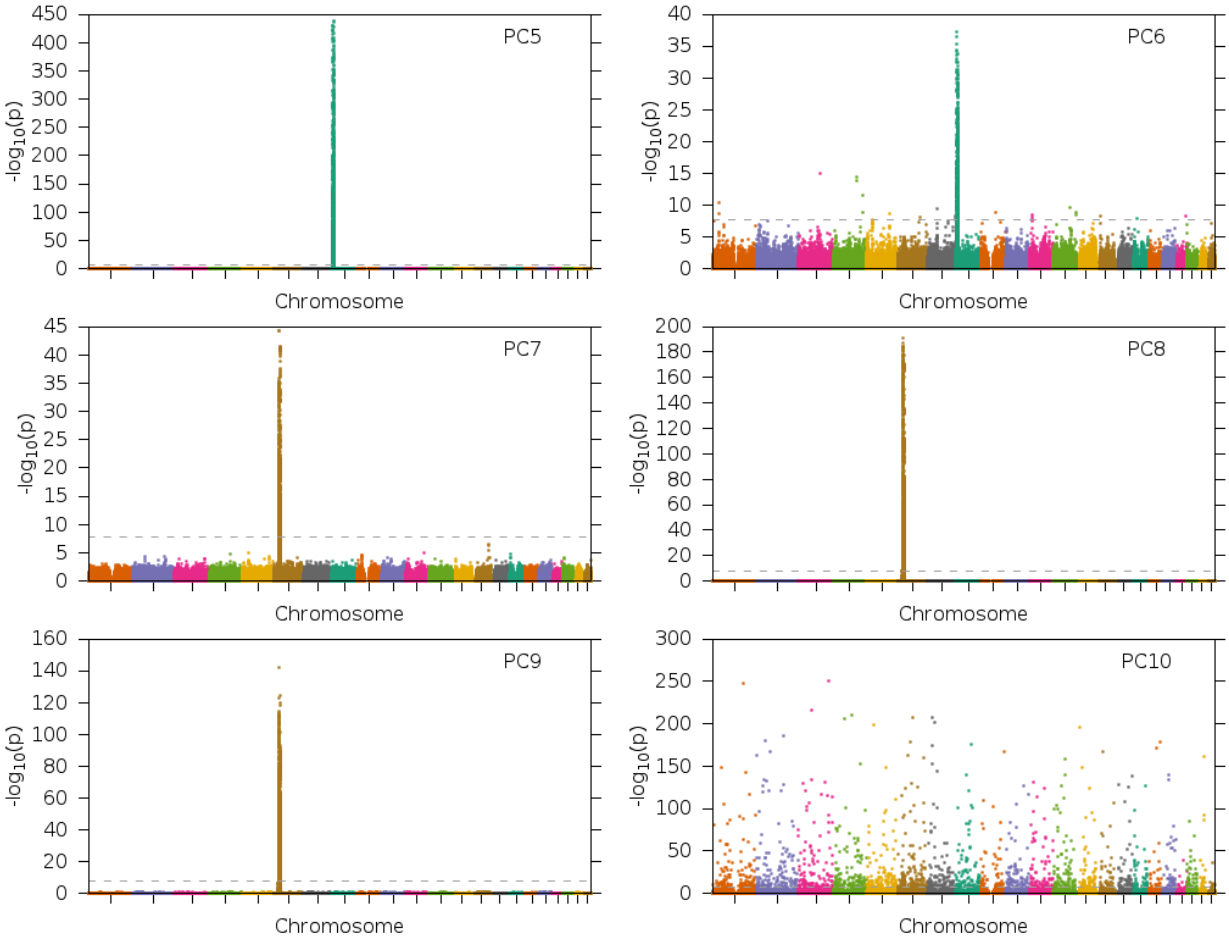


Supplementary Figure 3. QQ-plot of the selection statistic for PCs 1-4 in GERA data.
QQ-plots of actual vs. theoretical p-values are provided for (A) selection statistics for 608,981 SNPs in the GERA sample that passed the first stage of QC, and (B) selection statistics for 599,992 SNPs excluding the genome-wide significant loci listed in **Error! Reference source not found.**. Despite clear evidence of signal at the extreme tails, the overall distribution of test statistic was not inflated in the original set of SNPs ($.928 \leq \lambda_{GC} \leq .990$) nor in the filtered set ($.966 \leq \lambda_{GC} \leq 1.01$).



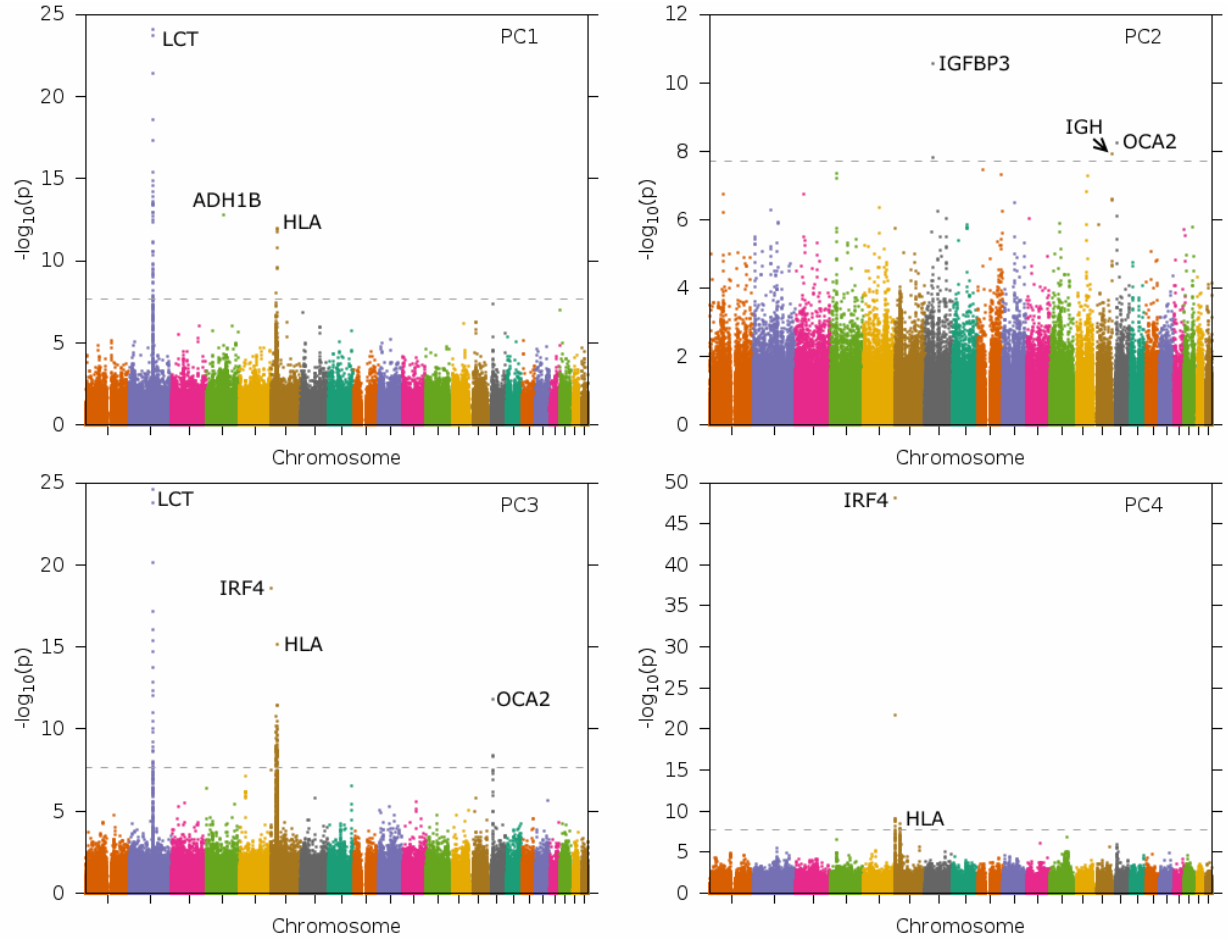
Supplementary Figure 4. Selection statistics for PCs 5-10 in GERA data.

The selection statistics for PCs 5-10 were dominated by exceedingly large signals at one locus (PCs 5-9) or substantial correlation with missing data rate per individual (PC 10; $\rho = 0.07, p < 2.2 \times 10^{-16}$), suggesting that these PCs are caused by PC artifacts and do not represent true population structure. PCs 1-4 were not significantly correlated with missing data.



Supplementary Figure 5. Selection statistics for PCs 1-4 in GERA data after removing significant regions.

We removed the genome-wide significant regions listed in **Error! Reference source not found.**, reran FastPCA and calculated the selection statistic across the genome. The significant hits in PCs 1-4 remain largely unchanged (Supplementary Table 4). The notable exception is the inversion on chromosome 8 spanning from 8-12 Mb. This indicates that the signal in that region was artifactual.



Supplementary Tables

Supplementary Table 1. CPU time and memory requirements of FastPCA and other methods.

We report the running time (in CPU seconds) and memory usage (GB) of PCA implementations, with standard deviation in parentheses. Runs in which smartpca, PLINK2-pca and flashpca exceeded the time constraint (100 hours) or memory constraint (40GB) are denoted as blank entries. When there are few individuals, PLINK2-pca ran faster and consumed less memory than FastPCA. However, FastPCA was able to run on 100k individuals and 100k SNPs in 56 minutes using 3.2GB of memory.

SAMPLES	**FASTPCA**		**FLASHPCA**		**PLINK2-PCA**		**SMARTPCA**	
(×1000)	**CPU**	**MEMORY**	**CPU**	**MEMORY**	**CPU**	**MEMORY**	**CPU**	**MEMORY**
1	0:01:42	0.54	0:00:55	1.25	0:00:19	0.02	0:02:10	0.17
	(0:06)	(0.00)	(0:01)	(0.00)	(0:01)	(0.00)	(0:10)	(0.00)
1.5	0:02:00	0.55	0:01:41	1.64	0:00:42	0.03	0:05:39	0.25
	(0:04)	(0.00)	(0:01)	(0.00)	(0:01)	(0.00)	(0:33)	(0.00)
2	0:02:18	0.57	0:02:44	2.03	0:01:15	0.05	0:10:11	0.35
	(0:06)	(0.00)	(0:01)	(0.00)	(0:01)	(0.00)	(0:48)	(0.00)
3	0:02:53	0.59	0:05:38	2.82	0:02:53	0.09	0:23:38	0.58
	(0:07)	(0.00)	(0:02)	(0.00)	(0:04)	(0.00)	(1:18)	(0.00)
5	0:03:58	0.64	0:14:31	4.44	0:08:20	0.25	1:11:21	1.19
	(0:08)	(0.00)	(0:06)	(0.00)	(0:17)	(0.00)	(7:09)	(0.00)
7	0:05:08	0.69	0:27:24	6.13	0:17:13	0.47	2:21:24	2.02
	(0:07)	(0.00)	(0:04)	(0.00)	(0:19)	(0.00)	(8:13)	(0.00)
10	0:06:56	0.77	0:54:37	9.11	0:39:15	0.94	5:15:58	3.64
	(0:05)	(0.00)	(0:16)	(0.00)	(1:08)	(0.00)	(16:59)	(0.00)
15	0:09:50	0.89	2:01:16	14.71	1:45:43	2.10	14:13:13	7.39
	(0:08)	(0.00)	(0:42)	(0.00)	(3:51)	(0.00)	(38:46)	(0.00)
20	0:13:05	0.98	3:32:55	21.04	3:41:55	3.70	29:34:22	12.44
	(0:09)	(0.00)	(0:55)	(0.00)	(10:06)	(0.00)	(41:27)	(0.00)
30	0:19:36	1.22	7:53:56	35.96	11:41:39	8.27	73:30:37	26.46
	(0:10)	(0.00)	(2:00)	(0.00)	(12:20)	(0.00)	(23:53)	(0.00)
50	0:29:57	1.69			47:16:16	22.87		
	(0:36)	(0.00)			(50:39)	(0.00)		
70	0:41:18	2.30						
	(1:16)	(0.00)						
100	0:56:00	3.20						
	(1:25)	(0.00)						

Supplementary Table 2. Proportion of significant SNPs at different thresholds in null simulations. We ran 10 simulations containing 50k SNPs and 10k simulated individuals and calculated the selection statistic under the null. We report the proportion of SNPs that meet significance at different thresholds (s.e. in parenthesis). In order for a SNP to be genome-wide significant in a simulation, its p-value must be less than 10^{-6} at $\alpha = 0.05$. This table shows that the selection statistic is well behaved under the null.

Threshold	$\boldsymbol{F_{ST} = 0.001}$	$\boldsymbol{F_{ST} = 0.01}$
$\mathbf{10^{-6}}$	2×10^{-6} (6.32×10^{-6})	0 (n/a)
$\mathbf{10^{-5}}$	8×10^{-6} (1×10^{-5})	1.2×10^{-5} (1.69×10^{-5})
$\mathbf{10^{-4}}$	9.8×10^{-5} (3.82×10^{-5})	1.2×10^{-4} (6.32×10^{-5})
$\mathbf{10^{-3}}$	9.98×10^{-4} (1.38×10^{-4})	1.04×10^{-3} (9.18×10^{-5})
$\mathbf{10^{-2}}$	9.98×10^{-3} (3.82×10^{-4})	9.85×10^{-3} (3.95×10^{-4})
$\mathbf{10^{-1}}$	9.99×10^{-2} (6.22×10^{-4})	9.96×10^{-2} (7.70×10^{-4})

Supplementary Table 3. Suggestive signals of selection in GERA data.
We report the regions with suggestive ($10^{-6} < p < 2.05 \times 10^{-8}$) evidence of selection (analogous to **Error! Reference source not found.**).

Locus	Chromosome	Region (Mb)	PC	Best Hit	p-value
	1	79.3 - 79.4	2	rs17590370	1.47e-7
INPP4A	2	98.5 - 98.5	2	rs78108890	5.00e-7
ANO10	3	43.7 - 43.7	2	rs116086673	1.57e-7
	4	4.8 - 4.8	3	rs12186237	3.90e-7
ARAP2	4	35.9 - 35.9	2	rs116105213	3.78e-8
TLR1[36]	4	38.5 - 38.5	2	rs5743611	5.42e-8
			4	rs4833095	6.52e-7
SLC45A2[38]	5	34.0 - 34.0	3	rs16891982	6.89e-8
	5	89.5 - 89.5	2	rs72779178	4.22e-7
	6	93.7 - 93.7	1	rs1538270	5.80e-7
DGKB	7	14.2 - 14.2	1	rs59706690	1.43e-7
CCDC146	7	76.8 - 76.8	2	rs17151162	5.96e-7
CADPS2	7	121.8 - 121.8	2	rs6947805	8.58e-7
PVT1	8	129.1 - 129.1	3	rs12676558	2.26e-7
EQTN	9	27.3 - 27.3	2	rs41305329	4.25e-8
RALGPS1	9	128.8 - 128.8	2	rs76798990	4.88e-8
	9	135.4 - 135.4	2	rs79784812	5.65e-7
TET1	10	70.1 - 70.1	2	rs7896856	2.71e-7
	12	94.5 - 94.5	4	rs79822723	2.64e-7
	13	77.2 - 77.2	2	rs75892602	1.30e-7
	13	80.4 - 80.4	2	rs117888143	4.13e-8
	13	83.0 - 83.0	1	rs73234476	7.14e-7
	14	40.2 - 40.2	1	rs8021234	5.55e-7
	20	1.8 - 1.8	1	rs6045087	1.05e-7

Supplementary Table 4. Top signals of selection in GERA data using PCs computed from SNPs in other regions.

After removing **Error! Reference source not found.** regions from the set of SNPs used to compute PCs, the selected loci remained the same except for the inversion on chromosome 8.

Locus	**Chromosome**	**Region (Mb)**	**PC**	**Best Hit**	p**-value**
LCT	2	134.8 – 137.6	1	rs6754311	8.27×10^{-25}
			3	rs4988235	2.50×10^{-25}
ADH1B	**4**	**100.5**	**1**	**rs1229984**	$\mathbf{1.57 \times 10^{-13}}$
IRF4	6	0.3 – 0.5	3	rs12203592	2.72×10^{-19}
			4	rs12203592	6.99×10^{-49}
HLA	6	30.8 – 32.9	1	rs382259	1.12×10^{-12}
			3	rs9268628	6.98×10^{-16}
			4	rs1265103	3.14×10^{-9}
IGFBP3	**7**	**45.3-45.9**	**2**	**rs150353309**	$\mathbf{2.69 \times 10^{-11}}$
IGH	**14**	**106.0-106.1**	**2**	**rs34614900**	$\mathbf{1.19 \times 10^{-8}}$
OCA2	15	25.9 – 26.2	2	rs12916300	5.81×10^{-9}
			3	rs12916300	1.55×10^{-12}

Supplementary Table 5. Performance of natural selection statistic in subsampled data.

The selection statistic was computed in random subsets of individuals of specified size for each SNP in Table 1 (except for the chromosome 8 inversion region) and the known selection regions TLR and SLC45A2 in Supplementary Table 3. We report the median selection statistic P-value across 100 random subsets.

		Sample size						
Locus	**SNP**	**Full data set**	**1k**	**2k**	**5k**	**10k**	**20k**	**50k**
LCT	rs6754311	2.15e-25	4.91e-17	2.97e-20	1.53e-23	1.17e-24	2.63e-25	1.02e-26
	rs4988235	1.15e-27	7.44e-17	9.80e-20	4.64e-23	3.11e-24	2.69e-25	1.62e-27
	rs17346504	8.41e-7	2.86e-2	1.25e-2	9.49e-4	6.03e-5	8.12e-6	9.80e-7
ADH1B	rs1229984	1.26e-13	3.91e-9	3.51e-11	1.97e-12	5.54e-13	1.50e-13	1.31e-13
IRF4	rs12203592	5.52e-55	3.15e-6	9.18e-12	7.47e-25	7.21e-36	7.02e-45	2.19e-54
HLA	rs382259	5.38e-13	8.68e-9	1.23e-10	7.07e-12	1.85e-12	7.51e-13	5.77e-13
	rs9268628	8.66e-18	3.62e-5	3.41e-7	5.97e-12	2.10e-14	2.68e-16	1.00e-17
	rs4394275	9.36e-12	8.40e-2	1.94e-3	1.44e-5	4.00e-8	7.86e-10	1.24e-11
IGFBP3	rs150353309	5.82e-12	5.90e-4	1.49e-5	2.72e-8	3.61e-10	3.34e-11	6.61e-12
IGH	rs34614900	5.23e-9	6.33e-3	2.24e-4	2.26e-6	2.01e-7	3.32e-8	5.32e-9
OCA2	rs12916300	2.80e-13	6.29e-6	1.07e-7	3.67e-9	1.94e-11	5.29e-12	3.11e-13
	rs2703951	5.11e-7	1.12e-1	2.45e-2	7.96e-4	7.17e-5	4.52e-6	5.74e-7
TLR1	rs5743611	5.42e-8	8.05e-3	4.27e-4	9.41e-6	1.19e-6	2.17e-7	5.60e-8
	rs4833095	6.52e-7	6.07e-4	3.37e-4	7.35e-5	3.64e-5	6.03e-6	7.10e-7
SLC45A2	rs16891982	6.89e-8	8.25e-4	2.17e-4	1.93e-5	4.55e-6	2.46e-7	7.31e-8

Supplementary Table 6. Allele frequencies for novel loci in GERA subpopulations.

The GERA sample was clustered into 5 discrete subpopulations using k-means clustering run on the top 4 PCs. Individual clusters were labelled to coincide with SNPweights and projected POPRES individuals. These were Ashkenazi Jewish (AJ), Eastern European (EE), Irish (IR), Northern European (NE) and South-east European (SE). Results are reported only for genome-wide significant SNPs at novel loci. We also report F_{ST} between each pair of subpopulations.

		AJ	EE	IR	NE	SE
Count		2,750	4,196	14,771	28,439	4,578
ADH1B	rs1229984	21.37%	4.99%	2.66%	2.96%	9.58%
IGFBP3	rs150353309	1.66%	4.38%	0.76%	1.10%	0.79%
	rs35751739	2.47%	7.71%	2.68%	3.06%	2.19%
IGH	rs34614900	13.63%	26.78%	17.29%	18.92%	12.73%

	AJ	EE	IR	NE
EE	0.00684			
IR	0.00671	0.00095		
NE	0.00655	0.00073	0.00013	
SE	0.00345	0.00239	0.00193	0.00182

Supplementary Table 7. Natural selection at ADH1B between discrete subpopulations.

The discrete-population selection statistic[21] (see Online Methods) for each pair of populations was calculated (below the diagonal) as well as the statistic comparing the frequency of rs1229984 in that population with the set of remaining individuals (diagonal). Genome-wide significant comparisons are those with $p < 5.47 \times 10^{-9}$ (608,981 SNPs × 15 subpopulation comparisons = 9,134,715 tests with $\alpha = 0.05$).

rs1229984	**AJ**	**EE**	**IR**	**NE**	**SE**
AJ	1.47e-06				
EE	4.15e-05	0.556			
IR	8.31e-07	0.00731	1.83e-08		
NE	1.04e-06	0.00932	0.293	2.61e-10	
SE	0.000121	0.0126	4.98e-06	8.84e-06	0.00012

Supplementary Table 8. ADH1B haplotypes in 1000 genomes.

Densities of known haplotypes in 1000 genomes Asian and European populations were calculated. 9 SNPs were used to determine haplotype and novel haplotypes were excluded from the analysis. 98% of the European haplotypes did not contain the derived allele (T) at rs122998 (above double bar line) compared to 20.8% of Asian haplotypes. The derived allele (A) of the regulatory SNP rs3811801 was not found at all in European populations, while haplotype H7 which contains this allele is the most common haplotype in Asian populations.

Haplotype	rs169343	rs381180	rs115991	rs122998	rs414753	rs207563	rs206670	rs17033	rs104202	**Asian** (CHB, CHS, JPT)	**European** (CEU, FIN, GBR, IBS, TSI)	**African** (ASW, LWK, YRI)
H1b	G	G	C	C	C	T	G	T	T	1.96%	40.11%	14.97%
H1c	G	G	C	C	A	T	G	T	T	0%	0.14%	5.21%
H2	G	G	A	C	C	T	G	T	T	0%	0.84%	18.66%
H2b	G	G	A	C	C	T	G	C	T	9.46%	10.10%	9.33%
H3	G	G	C	C	C	C	A	T	C	8.04%	27.21%	4.34%
H3c	G	G	C	C	C	C	G	T	T	0%	0%	0.43%
H4	G	G	A	C	A	T	G	T	T	6.96%	17.67%	46.42%
H4b	A	G	A	C	A	T	G	T	T	0%	1.96%	0.65%
H5	G	G	C	T	C	T	G	T	T	0.36%	1.12%	0%
H5b	A	G	A	T	A	T	G	T	T	0.18%	0.56%	0%
H6	G	G	C	T	C	C	A	T	C	12.14%	0.28%	0%
H7	G	A	C	T	C	C	A	T	C	60.89%	0%	0%

Supplementary Table 9. Natural selection at IGFBP3 between discrete subpopulations.

As in Supplementary Table 7, but for SNPs rs150353309 and rs150353309 in IGFBP3 which were under selection. Genome-wide significant comparisons are those with $p < 5.47 \times 10^{-9}$ (608,981 SNPs × 15 subpopulation comparisons = 9,134,715 tests with $\alpha = 0.05$).

rs150353309	**AJ**	**EE**	**IR**	**NE**	**SE**
AJ	0.755				
EE	0.178	4.07e-07			
IR	0.48	4.38e-07	0.00441		
NE	0.678	4.62e-07	0.0429	0.217	
SE	0.351	0.0014	0.955	0.6	0.374

rs35751739	**AJ**	**EE**	**IR**	**NE**	**SE**
AJ	0.675				
EE	0.0438	1.24e-07			
IR	0.909	5.99e-07	0.0703		
NE	0.757	2.33e-07	0.207	0.451	
SE	0.827	0.000332	0.614	0.379	0.233

Supplementary Table 10. Natural selection at IGH between discrete subpopulations.

As in Supplementary Table 7, but for SNP rs34614900 in IGH which was under selection and SNPs rs35237072 and rs34479337 were suggestive with p-value $< 10^{-6}$. Genome-wide significant comparisons are those with $p < 5.47 \times 10^{-9}$ (608,981 SNPs × 15 subpopulation comparisons = 9,134,715 tests with $\alpha = 0.05$).

rs34614900	**AJ**	**EE**	**IR**	**NE**	**SE**
AJ	0.23				
EE	0.00557	8.17e-08			
IR	0.386	4.43e-07	0.12		
NE	0.214	2.65e-06	0.0165	0.173	
SE	0.754	6.35e-07	0.0437	0.00577	0.00347

rs35237072	**AJ**	**EE**	**IR**	**NE**	**SE**
AJ	0.378				
EE	0.0151	2.76e-07			
IR	0.554	1.37e-06	0.151		
NE	0.373	3.21e-06	0.0569	0.432	
SE	0.771	1.13e-05	0.139	0.0384	0.0245

rs34479337	**AJ**	**EE**	**IR**	**NE**	**SE**
AJ	0.616				
EE	0.0472	1.52e-06			
IR	0.745	1.39e-05	0.371		
NE	0.613	9.15e-06	0.247	0.655	
SE	0.305	6.72e-06	0.0489	0.0183	0.0079