The American Journal of Human Genetics, Volume *98*

Supplemental Information

Fast Principal-Component Analysis Reveals Convergent Evolution of *ADH1B* in Europe and East Asia

Kevin J. Galinsky, Gaurav Bhatia, Po-Ru Loh, Stoyan Georgiev, Sayan Mukherjee, Nick J. Patterson, and Alkes L. Price

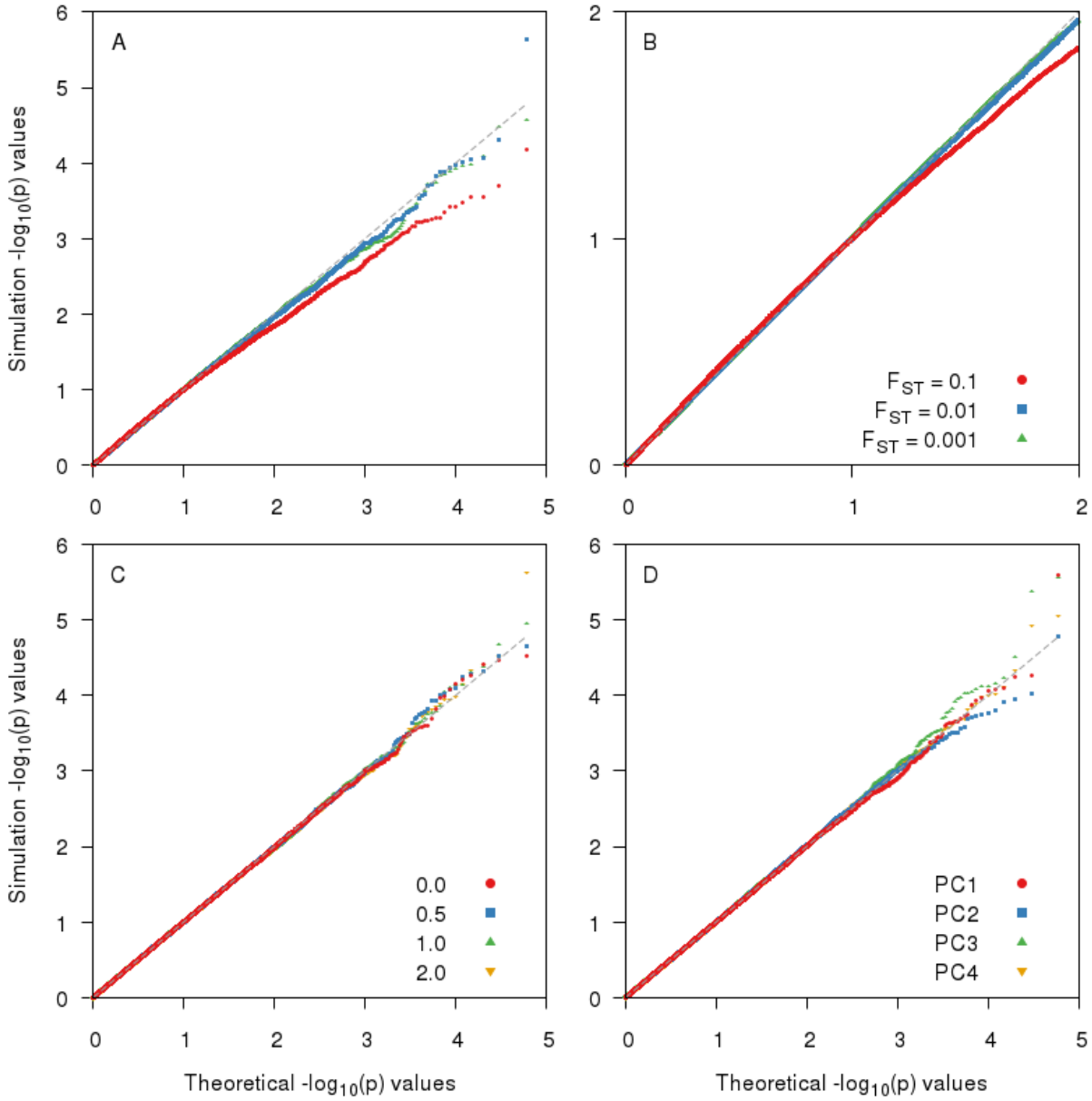

Figure S1. QQ-plot of the selection statistic in null simulations.

The selection statistic was generated for the first PC in null simulations containing 2 populations and differing by $F_{ST} = 0.001$, 0.01 and 0.1. (a) Examining all the p-values, the selection statistic was well calibrated for $F_{ST} = 0.001$ and 0.01, with deflation in the tails for $F_{ST} = 0.1$. (b) Looking only at p-values greater than 0.01, the selection statistic was well calibrated for $F_{ST} = 0.001$ and 0.01, but slightly inflated for p-values greater than 0.1 for $F_{ST} = 0.1$. This explains the results in Table S2. (c) In the case

with 2 populations differing by $F_{ST} = 0.001$, admixed individuals were generated with admixture proportion drawn from a $Beta(a, a)$ distribution, where increasing a means more admixture. (d) Five subpopulations were generated from a phylogenetic structure (see Methods), where the F_{ST} between populations 3, 4 and 5 was 0.001 and the F_{ST} between any other pair of populations was 0.01. In this case with five subpopulations, four principal components are sufficient to describe the population structure. For both examples with more complicated population structure, (c) and (d), the selection statistic remains well calibrated.

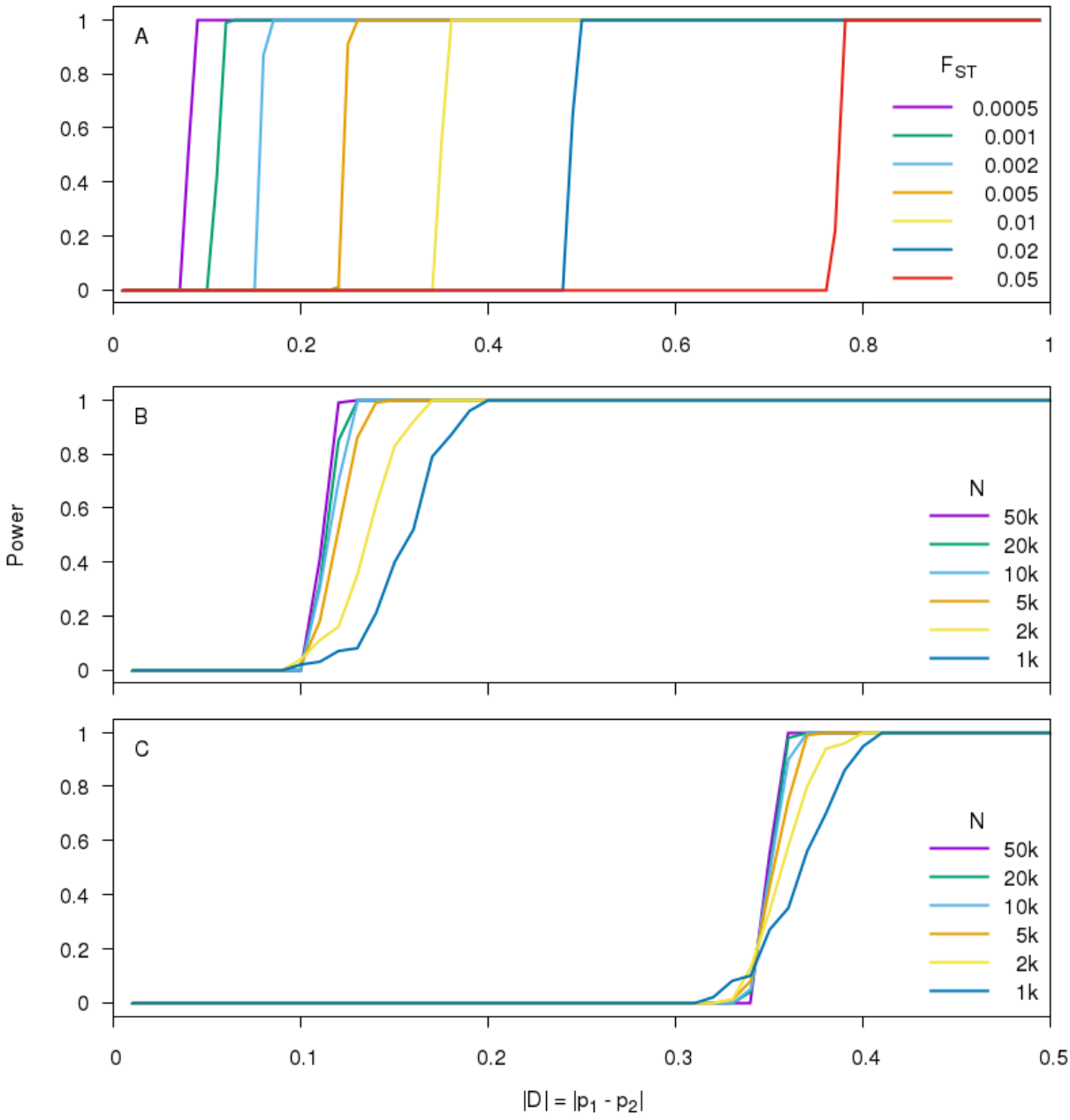

Figure S2. Power of the discrete-population selection statistic.

We ran the discrete-population selection statistic on the same simulations as in Figure 3 and found that the discrete-population and the PC-based selection statistics performed nearly identically in these regimes.

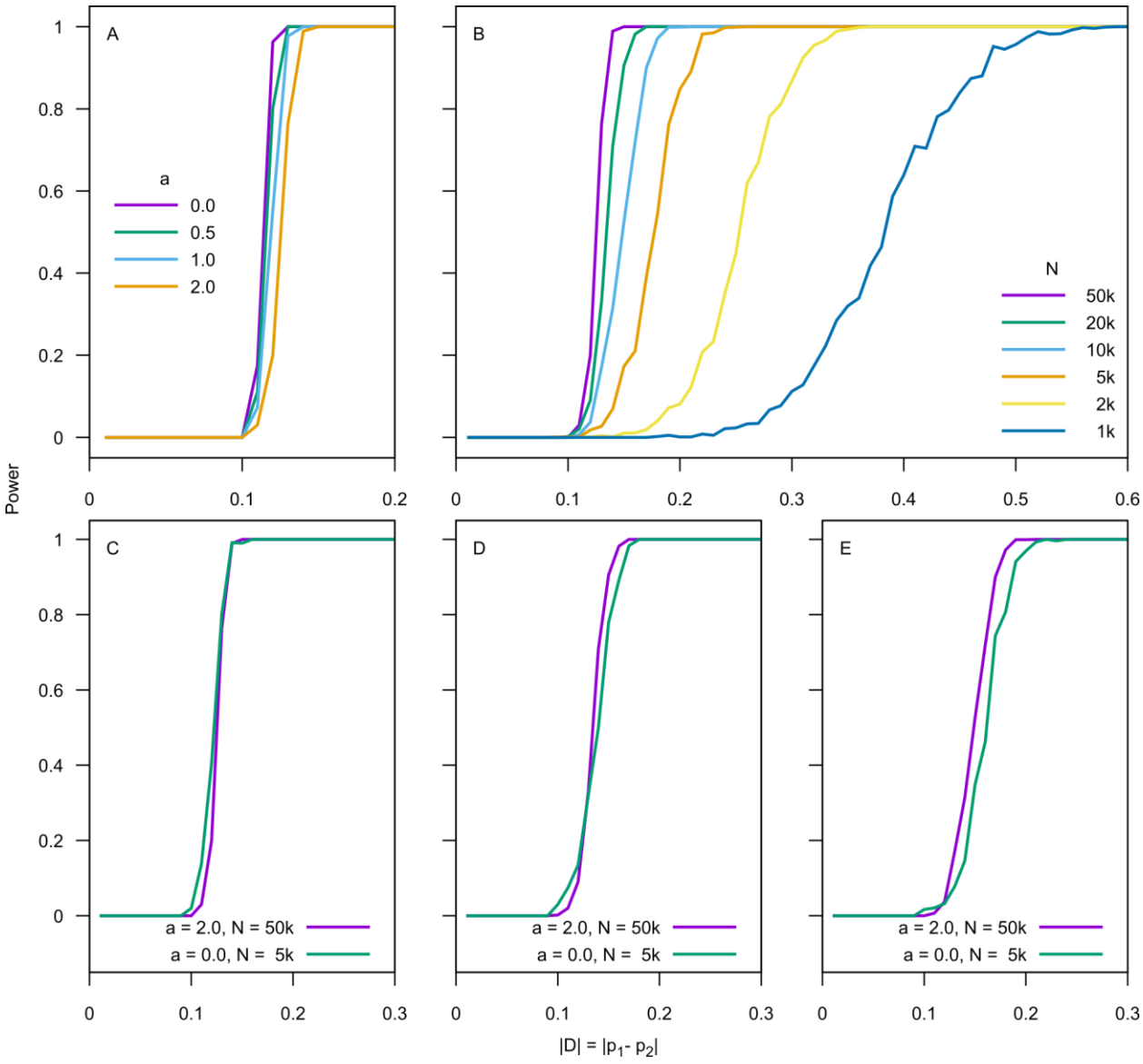

Figure S3. Power of the PC-based selection statistic in the presence of admixture.

Admixture or clinal variation in allele frequencies was simulated by samplings ancestry fraction between two ancestral populations from a $Beta(a, a)$ distribution. The two populations were differentiated by $F_{ST} = 0.001$. (a) Increasing a has a similar effect to reducing sample size (Figure 3). (b) Varying the number of samples when $a = 2.0$ had a dramatic effect, indicating that sample size is quite important in real data which will have small F_{ST} and non-discrete populations. (c-e) Setting $a = 2$ is roughly the same as having 10% of the data in a dataset with discrete populations.

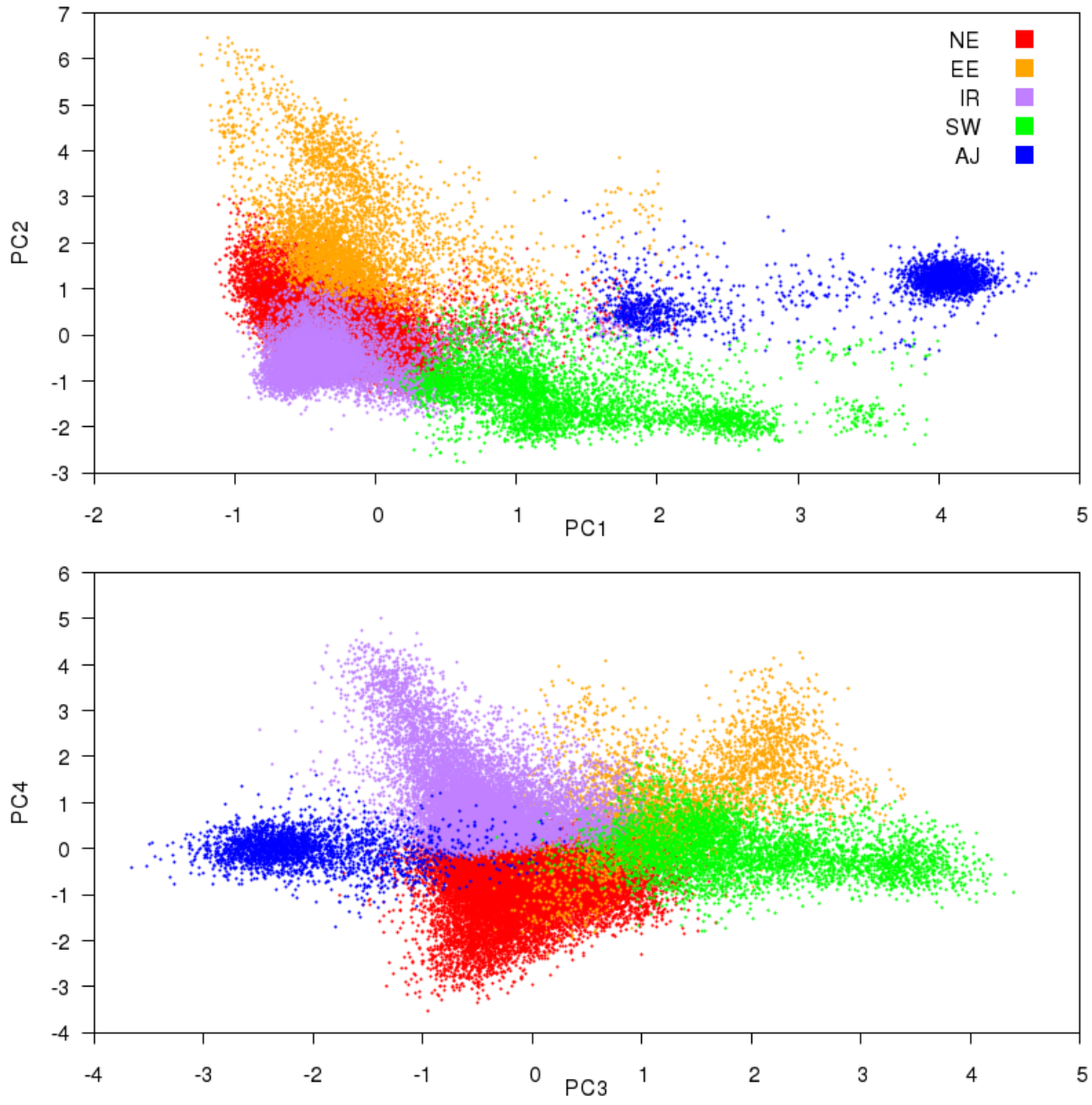

Figure S4. k-Means clustering confirms visually-observed subpopulations.

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.

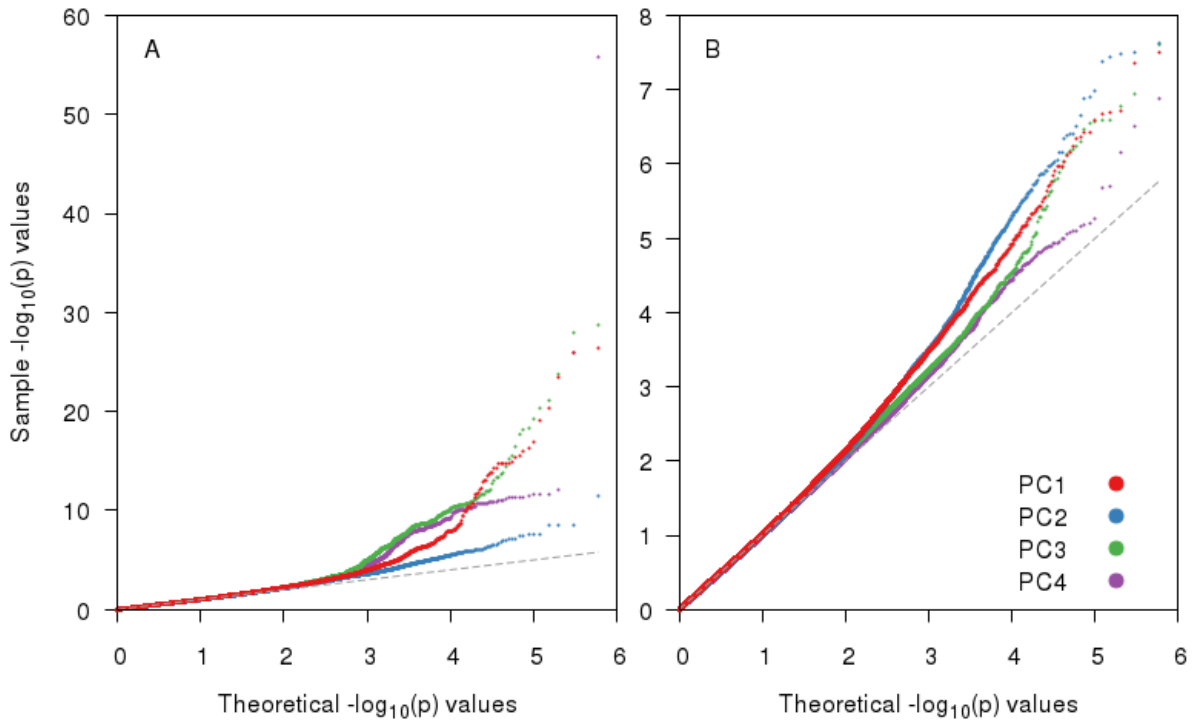

Figure S5. QQ-plot of the selection statistic for PCs 1-4 in GERA data.

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 Table 1. Despite clear evidence of signal at the extreme tails, the overall distribution of test statistic was not inflated in the original set of SNPs ($0.96 \leq \lambda_{GC} \leq 1.06$) nor in the filtered set ($0.94 \leq \lambda_{GC} \leq 1.05$).

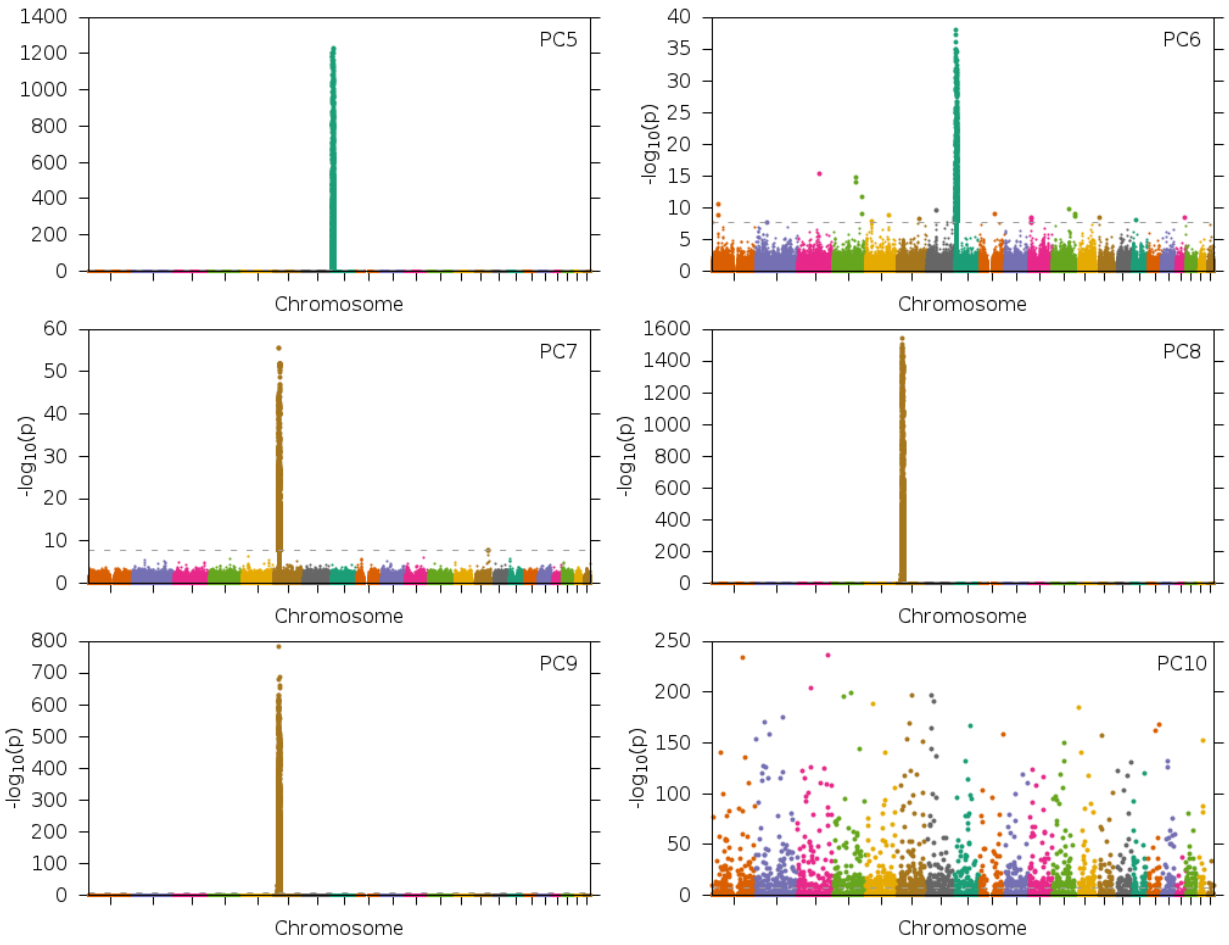

Figure S6. Selection statistics for PCs 5-10 in GERA data.

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.

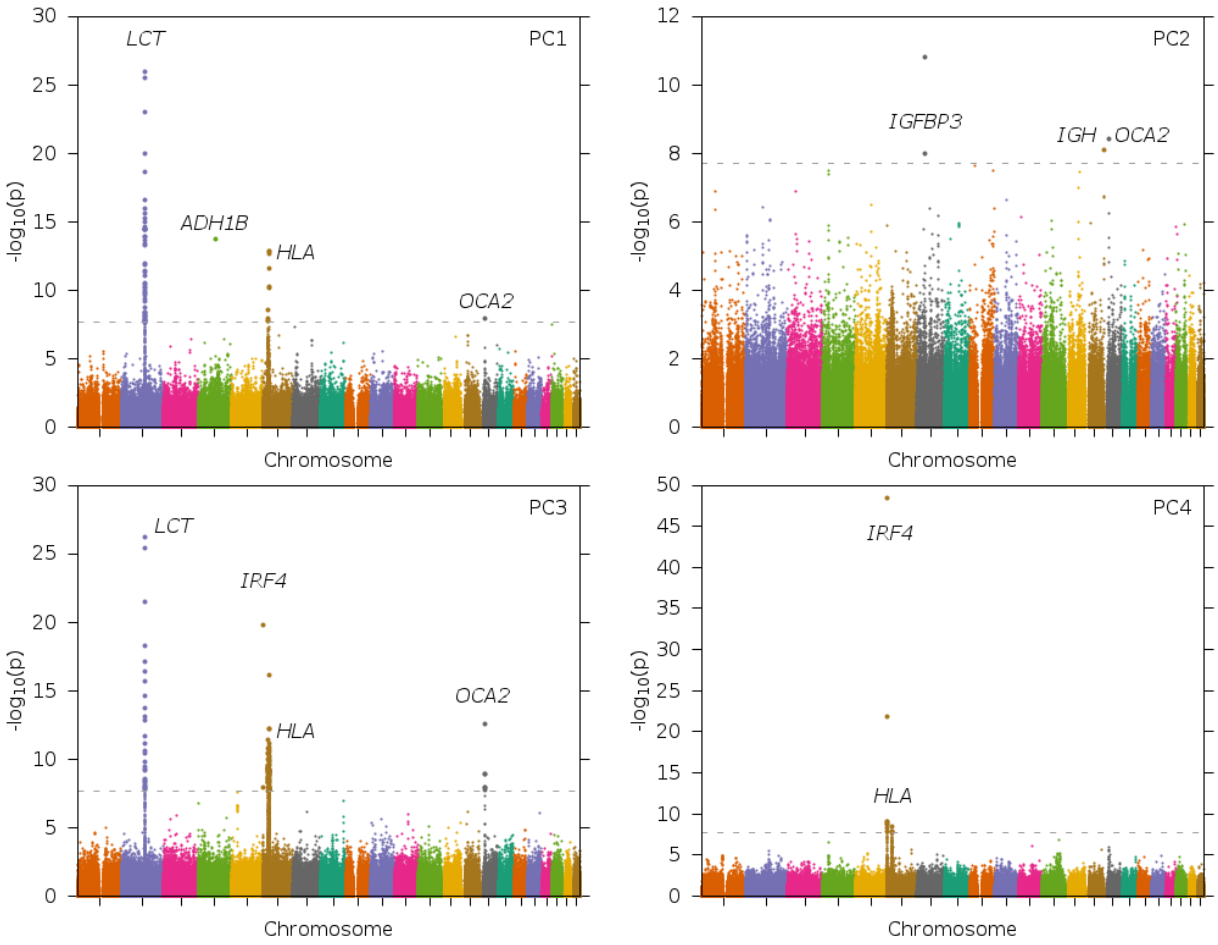

Figure S7. Selection statistics for PCs 1-4 in GERA data after removing significant regions.

We removed the genome-wide significant regions listed in Table 1, reran FastPCA and calculated the selection statistic across the genome. The significant hits in PCs 1-4 remain largely unchanged (Figure 6). The notable exception is the removal of the inversion on chromosome 8 spanning from 8-12 Mb. This indicates that the signal in that region was artifactual.

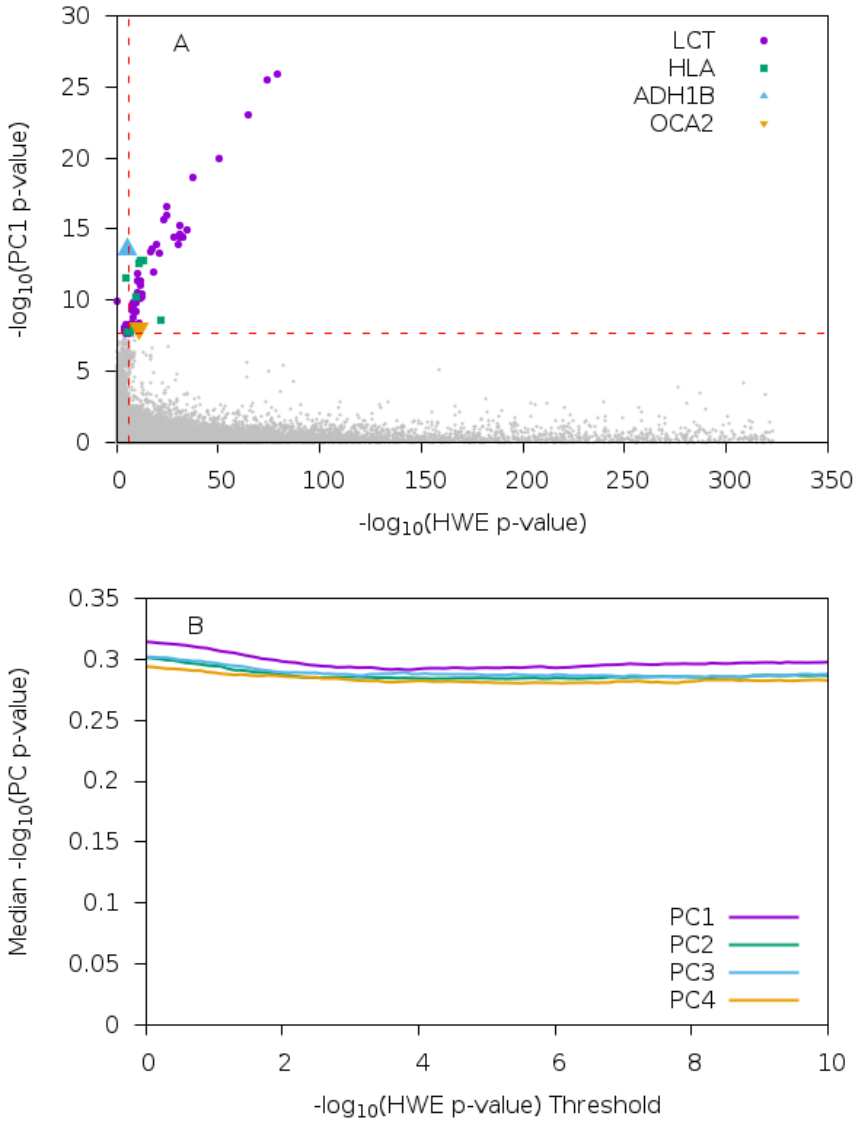

Figure S8. Comparison of selection statistic and Hardy-Weinberg disequilibrium p-values

Removing SNPs with a Hardy-Weinberg p-value less than 10^{-6} (those to the right of the vertical red line) removes many significant signals of selection. (a) For PC1, 51/63 significant SNPs have low Hardy-Weinberg p-values (for PCs 2-4 those numbers are 1/4, 39/116 and 2/12), compared with 3.9% of overall QC SNPs having HW p-value less than 10^{-6}. (b) We found no evidence of more significant selection statistics across PCs 1-4 for SNPs with strongly significant Hardy-Weinberg p-values.

Samples	FastPCA			flashpca			PLINK2-PCA			smartpca		
(x1000)	CPU		Memory	CPU		Memory	CPU		Memory	CPU		Memory
1	0:01:42	(0:06)	0.54	0:00:55	(0:01)	1.25	0:00:19	(0:01)	0.02	0:02:10	(0:10)	0.17
1.5	0:02:00	(0:04)	0.55	0:01:41	(0:01)	1.64	0:00:42	(0:01)	0.03	0:05:39	(0:33)	0.25
2	0:02:18	(0:06)	0.57	0:02:44	(0:01)	2.03	0:01:15	(0:01)	0.05	0:10:11	(0:48)	0.35
3	0:02:53	(0:07)	0.59	0:05:38	(0:02)	2.82	0:02:53	(0:04)	0.09	0:23:38	(1:18)	0.58
5	0:03:58	(0:08)	0.64	0:14:31	(0:06)	4.44	0:08:20	(0:17)	0.25	1:11:21	(7:09)	1.19
7	0:05:08	(0:07)	0.69	0:27:24	(0:04)	6.13	0:17:13	(0:19)	0.47	2:21:24	(8:13)	2.02
10	0:06:56	(0:05)	0.77	0:54:37	(0:16)	9.11	0:39:15	(1:08)	0.94	5:15:58	(16:59)	3.64
15	0:09:50	(0:08)	0.89	2:01:16	(0:42)	14.71	1:45:43	(3:51)	2.10	14:13:13	(38:46)	7.39
20	0:13:05	(0:09)	0.98	3:32:55	(0:55)	21.04	3:41:55	(10:06)	3.70	29:34:22	(41:27)	12.44
30	0:19:36	(0:10)	1.22	7:53:56	(2:00)	35.96	11:41:39	(12:20)	8.27			
50	0:29:57	(0:36)	1.69				47:16:16	(50:39)	0.02			
70	0:41:18	(1:16)	2.30									
100	0:56:00	(1:25)	3.20									

Table S1. 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. The standard deviation of memory usage was 0.00 GB for all runs. 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.

F_{ST}	N	a	Inflation	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
0.001	50k	0.0	1.01	9.98e-2	9.80e-3	1.02e-3	9.83e-5	1.33e-5
		0.5	1.00	9.98e-2	9.79e-3	9.55e-4	9.83e-5	1.00e-5
		1.0	1.00	9.99e-2	9.78e-3	9.97e-4	9.33e-5	1.17e-5
		2.0	1.01	9.97e-2	9.88e-3	1.01e-3	1.12e-4	1.17e-5
	5k	0.0	1.01	9.97e-2	9.83e-3	1.01e-3	1.02e-4	1.00e-5
		0.5	1.00	9.99e-2	9.88e-3	1.07e-3	9.33e-5	5.00e-6
		1.0	1.00	1.00e-1	9.85e-3	9.47e-4	9.50e-5	1.00e-5
		2.0	1.00	1.00e-1	9.98e-3	1.06e-3	1.18e-4	5.00e-6
	500	0.0	1.01	9.99e-2	9.58e-3	9.03e-4	9.67e-5	6.67e-6
		0.5	1.00	1.00e-1	9.92e-3	9.75e-4	8.50e-5	3.33e-6
		1.0	1.01	1.00e-1	9.82e-3	9.73e-4	7.33e-5	8.33e-6
		2.0	1.00	1.00e-1	1.00e-2	9.63e-4	1.00e-4	1.17e-5
0.01	50k	0.0	1.02	9.95e-2	8.95e-3	8.30e-4	5.83e-5	3.33e-6
		0.5	1.02	9.95e-2	9.00e-3	8.43e-4	6.00e-5	3.33e-6
		1.0	1.02	9.97e-2	8.92e-3	8.37e-4	5.83e-5	3.33e-6
		2.0	1.02	9.96e-2	9.07e-3	8.52e-4	5.67e-5	5.00e-6
	5k	0.0	1.02	9.96e-2	8.87e-3	8.28e-4	6.17e-5	0
		0.5	1.02	9.96e-2	8.99e-3	8.13e-4	5.67e-5	3.33e-6
		1.0	1.02	9.96e-2	9.10e-3	7.78e-4	7.67e-5	3.33e-6
		2.0	1.02	9.99e-2	9.07e-3	8.43e-4	5.50e-5	3.33e-6
	500	0.0	1.03	9.94e-2	8.76e-3	7.72e-4	6.17e-5	1.67e-6
		0.5	1.02	9.94e-2	9.28e-3	8.42e-4	7.00e-5	8.33e-6
		1.0	1.02	1.00e-1	9.24e-3	8.27e-4	8.17e-5	3.33e-6
		2.0	1.01	1.00e-1	9.45e-3	9.55e-4	8.67e-5	1.00e-5
0.1	50k	0.0	1.18	9.32e-2	5.65e-3	2.62e-4	8.33e-6	0
		0.5	1.18	9.32e-2	5.66e-3	2.65e-4	6.67e-6	0
		1.0	1.18	9.32e-2	5.63e-3	2.58e-4	6.67e-6	0
		2.0	1.18	9.32e-2	5.64e-3	2.67e-4	6.67e-6	0
	5k	0.0	1.18	9.32e-2	5.64e-3	2.52e-4	8.33e-6	0
		0.5	1.18	9.34e-2	5.65e-3	2.55e-4	8.33e-6	0
		1.0	1.18	9.33e-2	5.64e-3	2.50e-4	6.67e-6	0
		2.0	1.18	9.34e-2	5.69e-3	2.53e-4	8.33e-6	0
	500	0.0	1.18	9.35e-2	5.61e-3	2.62e-4	1.67e-6	0
		0.5	1.18	9.39e-2	5.78e-3	2.53e-4	8.33e-6	0
		1.0	1.16	9.46e-2	5.87e-3	2.72e-4	3.33e-6	0
		2.0	1.15	9.47e-2	6.23e-3	2.77e-4	5.00e-6	0

F_{ST}	N_e	τ	N	Inflation	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
0.001	100k	200	50k	1.01	9.98e-2	9.80e-3	1.02e-3	9.83e-5	1.33e-5
			5k	1.01	9.97e-2	9.83e-3	1.01e-3	1.02e-4	1.00e-5
			500	1.01	9.99e-2	9.58e-3	9.03e-4	9.67e-5	6.67e-6
	10k	20	50k	1.00	1.00e-1	9.90e-3	9.75e-4	1.00e-4	8.33e-6
			5k	1.00	1.00e-1	1.01e-2	1.09e-3	1.00e-4	8.33e-6
			500	1.01	1.00e-1	9.61e-3	8.88e-4	1.04e-4	1.25e-5
0.01	10k	200	50k	1.02	9.95e-2	8.95e-3	8.30e-4	5.83e-5	3.33e-6
			5k	1.02	9.96e-2	8.87e-3	8.28e-4	6.17e-5	0
			500	1.03	9.94e-2	8.76e-3	7.72e-4	6.17e-5	1.67e-6
	1k	20	50k	1.02	1.00e-1	9.06e-3	8.22e-4	7.78e-5	1.67e-5
			5k	1.02	1.01e-1	9.17e-3	7.33e-4	6.11e-5	1.11e-5
			500	1.02	1.00e-1	9.07e-3	7.78e-4	7.78e-5	5.56e-6
0.1	1k	200	50k	1.18	9.32e-2	5.65e-3	2.62e-4	8.33e-6	0
			5k	1.18	9.32e-2	5.64e-3	2.52e-4	8.33e-6	0
			500	1.18	9.35e-2	5.61e-3	2.62e-4	1.67e-6	0
	100	20	50k	1.18	9.33e-2	5.76e-3	2.37e-4	0	0
			5k	1.18	9.34e-2	5.75e-3	2.30e-4	0	0
			500	1.18	9.33e-2	5.88e-3	2.07e-4	3.33e-6	0

N	PC	Inflation	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
500k	1	1.02	9.96e-2	9.53e-3	8.53e-4	7.5e-5	5.00e-6
	2	1.00	9.94e-2	1.04e-2	1.12e-3	1.02e-4	6.67e-6
	3	0.99	9.95e-2	1.04e-2	1.12e-3	1.32e-4	1.83e-5
	4	0.99	1.00e-2	1.03e-2	1.04e-3	1.10e-4	1.83e-5
50k	1	1.02	9.94e-2	9.55e-3	8.80e-4	7.33e-5	3.33e-6
	2	1.00	9.93e-2	1.06e-2	1.11e-3	9.33e-5	1.17e-5
	3	0.99	1.00e-1	1.05e-2	1.17e-3	1.38e-4	1.50e-5
	4	0.99	1.00e-1	1.03e-2	1.11e-3	1.27e-4	2.33e-5
500	1	1.02	9.95e-2	9.33e-3	8.42e-4	8.00e-5	3.33e-6
	2	0.99	1.00e-1	1.02e-2	1.03e-3	8.50e-5	8.33e-6
	3	1.00	1.00e-1	1.01e-2	1.00e-3	9.17e-5	1.00e-5
	4	1.00	9.98e-2	1.02e-2	1.06e-3	1.20e-4	1.33e-5

Table S2. Inflation of the selection statistic in simulated data.

We ran 10 simulations containing 60k SNPs and various numbers of simulated individuals (N) in two populations under different levels of admixture and calculated the selection statistic under the null. Admixture was sampled from a $Beta(a, a)$ where an increase in the admixture parameter (a) represents a greater probability of fractional ancestry. When $a = 0$ there is no admixture and when $a = 1$,

fractional ancestry follows a $Uniform(0,1)$ distribution. We report the inflation of the median selection statistic (median divided by the theoretical value of 0.455 under the null) and the proportion of SNPs that attain significance at different thresholds. The median selection statistic was inflated for simulations with large F_{ST} (at large F_{ST} it is impossible for the selection statistic to be extremely significant, and this deficiency in the tail implies a higher ratio of median to average; see Figure S1), but well behaved at the small values of F_{ST} that correspond to our analyses of real data. The proportion of SNPs that attain significance was well-calibrated in all experiments.

We additionally investigated the effect of population bottlenecks on the selection statistic. For a fixed F_{ST}, we would generate two simulated datasets differing in the effective population size (N_e) and number of generations (τ). The statistic remained well calibrated under tighter population bottlenecks.

Lastly, we considered the effect of a more complicated population structure on the selection statistic. We simulated five populations with a phylogenetic structure where three of the populations are more closely related than the other two (see Methods). We again did not see inflation in the median selection statistic nor the proportion of SNPs that attain different significance thresholds.

PCA SNP Set	LD-pruned				LD-pruned, Table 1 Removed			
Selection SNP Set	LD-pruned		Full		LD-pruned, Table 1 removed		Full, Table 1 removed	
	Mean	Med	Mean	Med	Mean	Med	Mean	Med
PC1	1.00	1.02	1.07	1.06	1.00	1.03	1.05	1.06
PC2	1.00	0.98	1.03	1.00	1.00	0.98	1.01	1.00
PC3	1.00	0.95	1.07	0.99	1.00	0.96	1.02	0.99
PC4	1.00	0.96	1.03	0.96	1.00	0.97	0.99	0.96
PC5	1.00	0.12	2.81	0.21	1.00	0.90	0.97	0.89
PC6	1.00	0.89	1.02	0.88	1.00	0.96	0.99	0.96
PC7	1.00	0.92	1.26	0.94	1.00	0.50	0.86	0.47
PC8	1.00	0.34	8.12	0.33	1.00	0.86	0.93	0.81
PC9	1.00	0.40	5.56	0.39	1.00	0.95	0.95	0.89
PC10	1.00	0.49	0.94	0.46	1.00	0.78	0.97	0.70

Table S3. Inflation of the selection statistic in GERA data.

This table indicates the average value of the selection statistic as well as the median selection statistic divided by the theoretical median (0.455) in GERA data. PCA was run on the set of 162,335 LD-pruned SNPs, and the selection statistic was applied to either the set of 162,335 LD-pruned SNPs or the full set of 608,981 SNPs passing QC. Additional analyses were performed with the significant regions from Table 1 removed from all SNP sets. When computing selection statistics using the full set of SNPs passing QC, inflation can occur if SNPs with higher differentiation tend to have higher LD, which can occur as a consequence of true selection. PCs 2-4 show moderate inflation when examining the means, but no inflation when looking at the median chi-squared (1 d.o.f) statistic, indicating that inflation is driven by outliers in the distribution. Removing Table 1 regions decreased the mean for these PCs, without affecting the median value. For PC1, a qualitatively similar reduction was observed, although a slight inflation in the mean remained. However, after conservatively correcting selection statistics for inflation in the mean and/or median, all SNPs in Table 1 remained genome-wide significant except for the OCA2 locus (a known signal of selection) on PC1. For PCs 5-10, the unusual mean and/or median values are consistent with the fact that these PCs are caused by PC artifacts and do not represent true population structure (Figure S5).

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[74]	4	38.5 - 38.5	2	rs5743611	5.42e-8
			4	rs4833095	6.52e-7
SLC45A2[76]	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

Table S4. Suggestive signals of selection in GERA data.

We report the regions with suggestive ($10^{-6} < p < 2.05 \text{ x } 10^{-8}$) evidence of selection (analogous to Table 1).

Locus	Chromosome	Region (Mb)	PC	Best Hit	p-value
LCT	2	135.0 – 137.1	1	rs6754311	1.23×10^{-26}
			3	rs4988235	5.65×10^{-27}
ADH1B	4	100.5	1	rs1229984	1.76×10^{-14}
IRF4	6	0.3 – 0.5	3	rs12203592	1.61×10^{-20}
			4		3.29×10^{-49}
HLA	6	31.1 – 32.8	1	rs382259	1.47×10^{-13}
			3	rs9268628	7.15×10^{-17}
			4	rs1265103	2.84×10^{-9}
IGFBP3	7	45.3-45.9	2	rs150353309	1.53×10^{-11}
IGH	14	106.0	2	rs34614900	7.86×10^{-9}
OCA2	15	25.9 – 26.2	1	rs12916300	1.26×10^{-8}
			2		3.76×10^{-9}
			3		2.67×10^{-13}

Table S5. Top signals of selection in GERA data using PCs computed from SNPs in other regions.

After removing Table 1 regions from the set of SNPs used to compute PCs, the selected loci remained the same except for the inversion on chromosome 8.

		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

Table S6. Allele frequencies for highlighted 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 highlighted loci. We also report F_{ST} between each pair of subpopulations.

rs1229984	AJ	EE	IR	NE	SE
AJ	1.47e-6				
EE	4.15e-5	0.556			
IR	8.31e-7	0.00731	1.83e-8		
NE	1.04e-6	0.00932	0.293	2.61e-10	
SE	0.000121	0.0126	4.98e-6	8.84e-6	0.00012

Table S7. Natural selection at *ADH1B* between discrete subpopulations.

The discrete-population selection statistic[19] (see 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 $\alpha = 0.05$).

Haplotype	rs1693439	rs3811801	rs1159918	rs1229984	rs4147536	rs2075633	rs2066701	rs17033	rs1042026	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%

Table S8. *ADH1B* haplotypes in 1000 genomes.

We computed frequencies of known haplotypes in 1000 genomes Asian, European and African populations. 9 SNPs were used to determine haplotype and haplotypes not described in Li *et al*.[55] were excluded from the analysis. 98% of the European haplotypes did not contain rs1229984*T (above line) compared to 20.8% of Asian haplotypes. The "A" allele of 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.

rs150353309	AJ	EE	IR	NE	SE
AJ	0.755				
EE	0.178	4.07e-7			
IR	0.48	4.38e-7	0.00441		
NE	0.678	4.62e-7	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-7			
IR	0.909	5.99e-7	0.0703		
NE	0.757	2.33e-7	0.207	0.451	
SE	0.827	0.000332	0.614	0.379	0.233

Table S9. Natural selection at *IGFBP3* between discrete subpopulations.

As in Table S7, 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 $\alpha = 0.05$).

rs34614900	AJ	EE	IR	NE	SE
AJ	0.23				
EE	0.00557	8.17e-8			
IR	0.386	4.43e-7	0.12		
NE	0.214	2.65e-6	0.0165	0.173	
SE	0.754	6.35e-7	0.0437	0.00577	0.00347

rs35237072	AJ	EE	IR	NE	SE
AJ	0.378				
EE	0.0151	2.76e-7			
IR	0.554	1.37e-6	0.151		
NE	0.373	3.21e-6	0.0569	0.432	
SE	0.771	1.13e-5	0.139	0.0384	0.0245

rs34479337	AJ	EE	IR	NE	SE
AJ	0.616				
EE	0.0472	1.52e-6			
IR	0.745	1.39e-5	0.371		
NE	0.613	9.15e-6	0.247	0.655	
SE	0.305	6.72e-6	0.0489	0.0183	0.0079

Table S10. Natural selection at *IGH* between discrete subpopulations.

As in Table S7, 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 $\alpha = 0.05$).

	LD Strategy			
Rare Variant Strategy	**Standard PCA**	**LD Pruning**	**LD Shrinkage**	**LD Regression**
Include all variants	0.023	0.012	0.007	0.006
Exclude MAF < 0.02	0.287	0.441	0.442	0.463
Exclude Singletons	0.341	0.541	0.567	0.504
Reweight FST = 0.01	0.352	0.534	0.563	0.528

Table S11. Evaluation of LD and rare variant strategies for running PCA in POPRES targeted sequencing data.

We evaluated four methods for dealing with LD, and four methods for dealing with rare variants. We report the total variance explained by the top PCs in distinguishing Northern and Southern Europeans in POPRES targeted sequencing data.