The American Journal of Human Genetics, Volume *89*
Supplemental Data

Genome-Wide Comparison of African-Ancestry Populations from CARe and Other Cohorts Reveals Signals of Natural Selection

Gaurav Bhatia, Nick Patterson, Bogdan Pasaniuc, Noah Zaitlen, Giulio Genovese, Samuela Pollack, Swapan Mallick, Simon Myers, Arti Tandon, Chris Spencer, Cameron D. Palmer, Adebowale A. Adeyemo, Ermeg L. Akylbekova, L. Adrienne Cupples, Jasmin Divers, Myriam Fornage, W.H. Linda Kao, Leslie Lange, Mingyao Li, Solomon Musani, Josyf C. Mychaleckyj, Adesola Ogunniyi, George Papanicolaou, Charles N. Rotimi, Jerome I. Rotter, Ingo Ruczinski, Babatunde Salako, David S. Siscovick, Bamidele O. Tayo, Qiong Yang, Steve McCarroll, Pardis Sabeti, Guillaume Lettre, Phil De Jager, Joel Hirschhorn, Xiaofeng Zhu, Richard Cooper, David Reich, James G. Wilson, and Alkes L. Price

Table S1. Inflation of the χ^2 Statistic

$F_{st}\backslash f_s$	0.1	0.2	0.3	0.4	0.5
0.01	1.14E-04	1.33E-04	1.15E-04	1.03E-04	1.03E-04
0.005	1.67E-04	1.15E-04	1.07E-04	1.04E-04	9.85E-05
0.001	1.09E-04	9.84E-05	1.01E-04	1.04E-04	1.04E-04

Cells show the empirical proportion of results observed with $P < 10^{-4}$ over 10 million simulations with the given values of F_{st} and f_s. The reason for this inflation is the biased estimate of the average allele frequency. While the inflation is significant at low frequencies, we believe that our results are trustworthy because the magnitude of the fat tail observed in real data far exceeds that observed in simulations. Additionally, all reported significant markers have large enough minor allele frequencies that we do not expect inflation to play a role.

Table S2. Comparison of Small Differences in Selective Coefficient

10 Generations ($F_{ST} = 0.001$)				
s1/s2	0.00	0.02	0.04	0.06
0.00	0.00	3.81	16.04	37.75
0.02		0.00	4.26	17.82
0.04			0.00	4.70
0.06				0.00
50 Generations ($F_{ST} = 0.005$)				
s1/s2	0.00	0.02	0.04	0.06
0.00	0.00	24.66	118.28	276.60
0.02		0.00	40.03	157.89
0.04			0.00	43.18
0.06				0.00
100 Generations ($F_{ST} = 0.01$)				
s1/s2	0.00	0.02	0.04	0.06
0.00	0.00	60.61	225.97	308.89
0.02		0.00	71.42	136.19
0.04			0.00	19.59
0.06				0.00

We performed forward simulations to analyze the effect of similar selective pressures on two populations. We assume that the selected allele is identical in both populations but the selection coefficients are different. We also assume that genetic drift is small with respect to selection and evaluated the expected χ^2 statistic attainable in the limit of infinite sample size using a 2-population test of differentiation. In these simulations we used

$$p_{t+1} = \frac{p_t(1+s)}{1+p_t s}$$

to advance the selected allele frequency using a start frequency $p_0 = 0.1$. We calculated F_{ST} from the number of generations, τ, using

$$F_{ST} = -\log(1 - \tau/N_e)$$

We assume that $N_e = 10000$. Those simulations that achieve genome-wide significance ($P < 5 \times 10^{-8}$) are highlighted in green. Clearly, this test is sensitive to the number of generations since the start of selection, as well as the difference between the selective coefficients applied to the two populations. Note that for cases of (s1, s2) equal to (0.02, 0), (0.04, 0), (0.06, 0), and (0.04, 0.02), a test has more power after 100 generations than 50 generations. For cases of (s1, s2) equal to (0.06, 0.02), (0.06, 0.04) the reverse is true. In the former cases, the allele frequency difference at 100 generations is greater than twice the difference at 50 generations allowing for a

more powerful test. However, in the latter cases the selected allele approaches fixation and this results in a drop in statistics and power.

Table S3. LSBL Empirical P Values

Chr	Gene or Region	SNP	Position	Rank Based P-values		
				African American	Nigerian	Gambian
6	HLA	rs28366191	32472168	8.20E-01	3.95E-02	2.88E-05
6	HLA	rs6901541	32550239	1.00E+00	2.40E-06	2.28E-05
6	HLA	rs2179915	33173712	3.72E-03	1.29E-04	1.00E+00
7	CD36	rs12721454	79678275	7.22E-04	1.09E-03	1.00E+00
7	CD36	rs513740	79872884	9.97E-01	4.92E-05	8.05E-02
8	PSCA	rs2920283	143754039	2.00E-03	3.07E-03	1.00E+00
11	HBB	rs7936387	5256204	1.00E+00	1.44E-05	2.31E-04

We show the empirical P-values for each of our highly significant SNPs when tested using the LSBL. All of these SNPs have an LSBL in the top 0.5% of the empirical distribution indicating the power of the LSBL to elucidate potential targets of selection. However, the test does not provide a method for assessing genome-wide significance and several other loci rank within this top 0.5% tail.

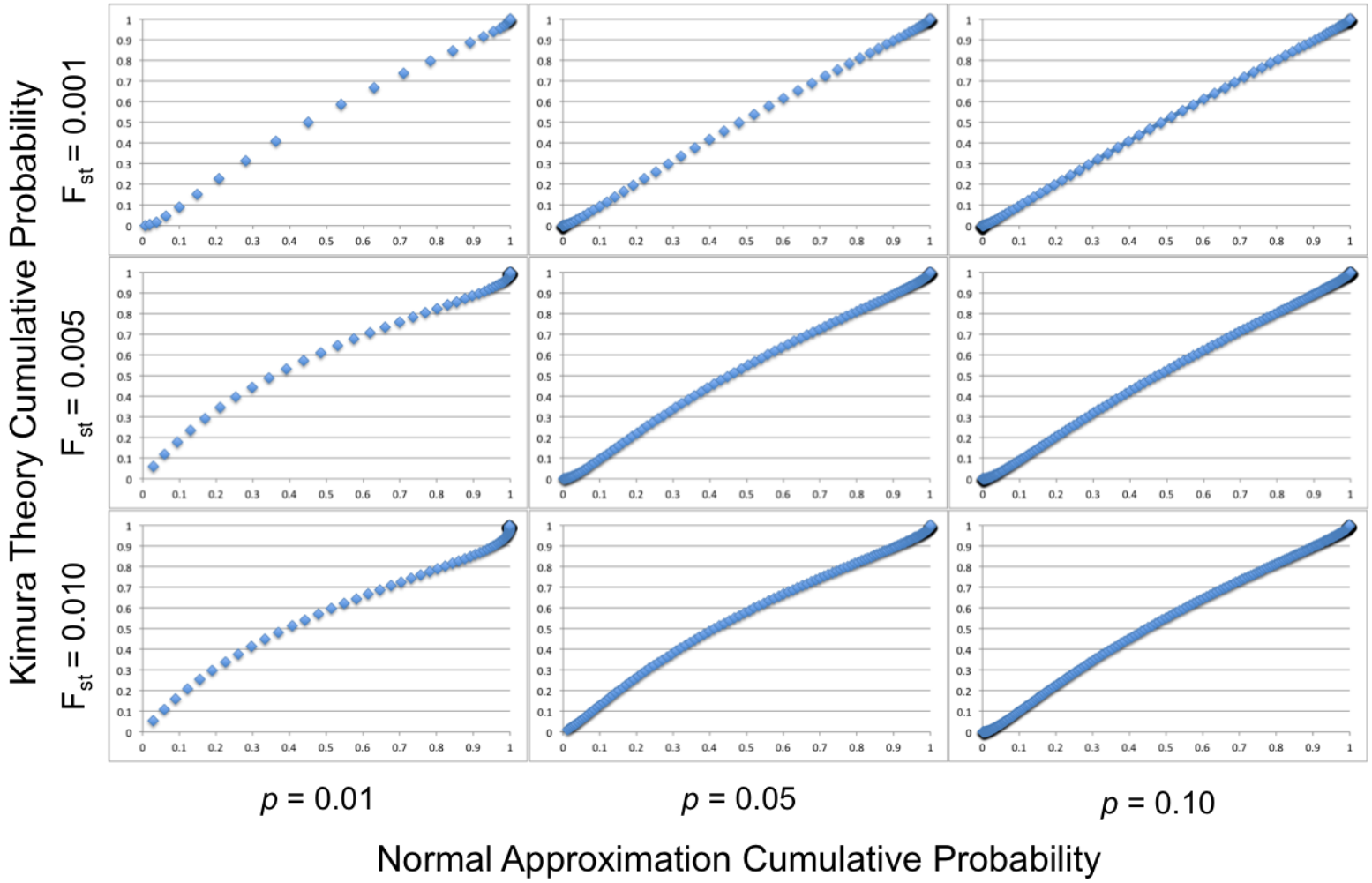

Figure S1. Evaluation of the Normal Approximation to Neutral Genetic Drift
We assume a starting frequency, *p*, and an amount of genetic drift, F_{st}, and evaluate the cumulative distribution function of the frequency following drift. This plot compares the cumulative distribution function under the normal approximation to the true cumulative distribution function evaluated using Kimura theory. It is clear that the approximation holds well under situations with a higher start allele frequency and with lower F_{st} and breaks down for low frequencies. The approximation is worst off in a regime with a low starting frequency and high F_{st}. However, for the range of allele frequencies and F_{st} under consideration, the approximation is reasonable.

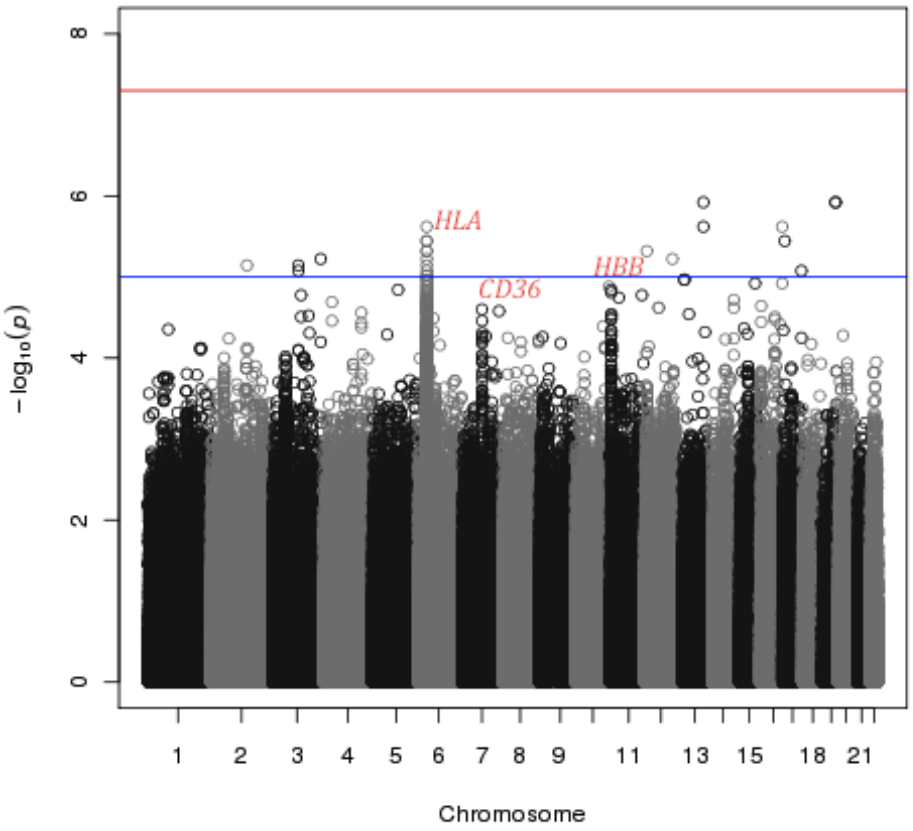

Figure S2. Manhattan Plot of LSBL Results
The statistic is calculated as described above (see Table S2) and then ranked, within results from each branch, to produce the p-value. Peaks at HLA, CD36 and HBB are evident. The result at PSCA is not obviously replicated. These results illustrate the power of the LSBL to uncover loci under selection. Note that because the P-value is simply based on a ranking, it is not possible for any locus to attain genome-wide significance after correction for multiple hypotheses tested.