Supplementary Figures and Tables - Patterns of diversity in modern humans around candidate sites

Fernando Racimo[1,2,*], Martin Kuhlwilm[2], Montgomery Slatkin[1],

1 Department of Integrative Biology, Univeristy of California, Berkeley, CA, USA

2 Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany.

∗ **E-mail: fernandoracimo@gmail.com**

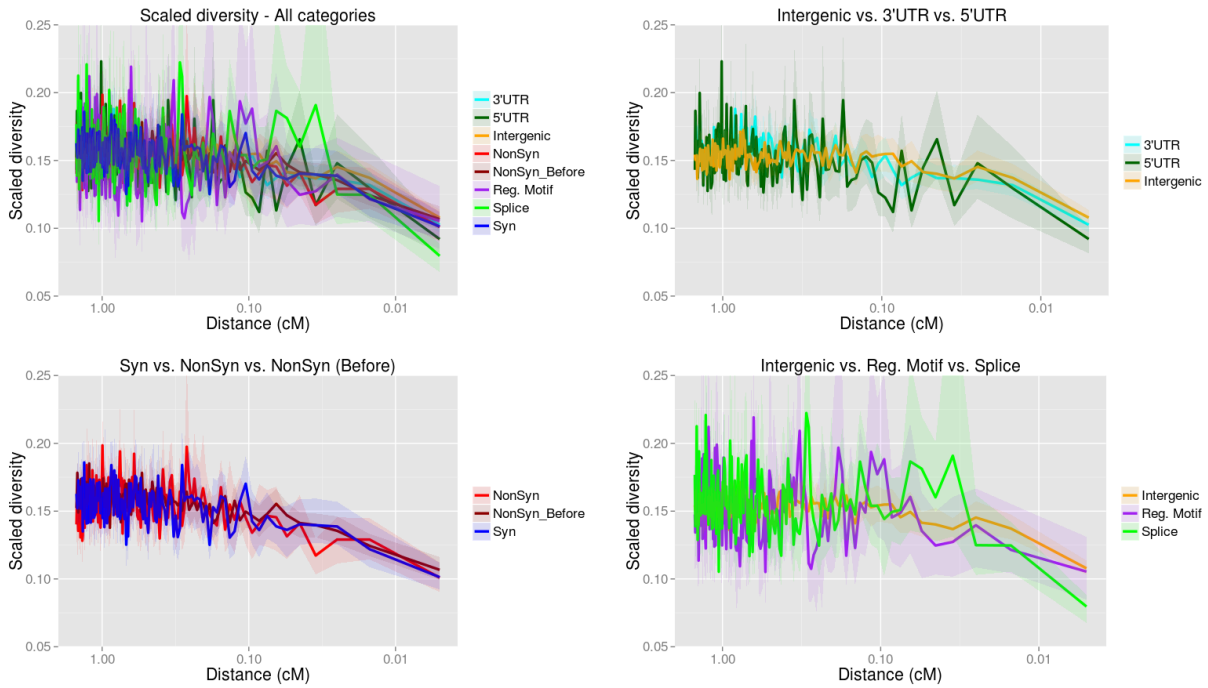

Figure S1. Human diversity per site (calculated in the CG panel) scaled by divergence of the human reference to the human-chimpanzee ancestor around different classes of fixed modern-human-specific single-nucleotied changes where Altai Neanderthal and Denisova are homozygous ancestral. The statistic was calculated in windows of 0.01 cM and the x-axis shows distance of the window midpoint to the fixed change on a log-scale. The upper left panel shows all functional categories tested, while the other panels show different subsets of these for ease of comparison.

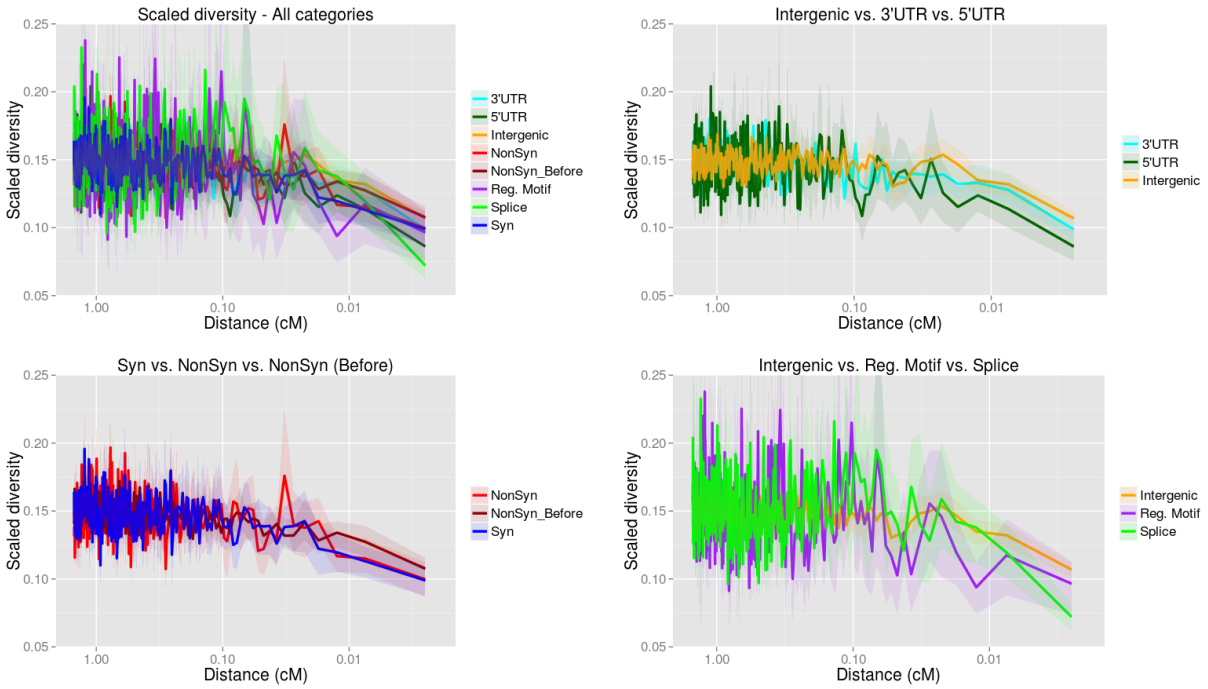

Figure S2. Human diversity per site (calculated in the 1000G panel) scaled by divergence of the human reference to the human-chimpanzee ancestor around different classes of fixed modern-human-specific single-nucleotied changes where Altai Neanderthal and Denisova are homozygous ancestral. The statistic was calculated in windows of 0.005 cM and the x-axis shows distance of the window midpoint to the fixed change on a log-scale. The upper left panel shows all functional categories tested, while the other panels show different subsets of these for ease of comparison.

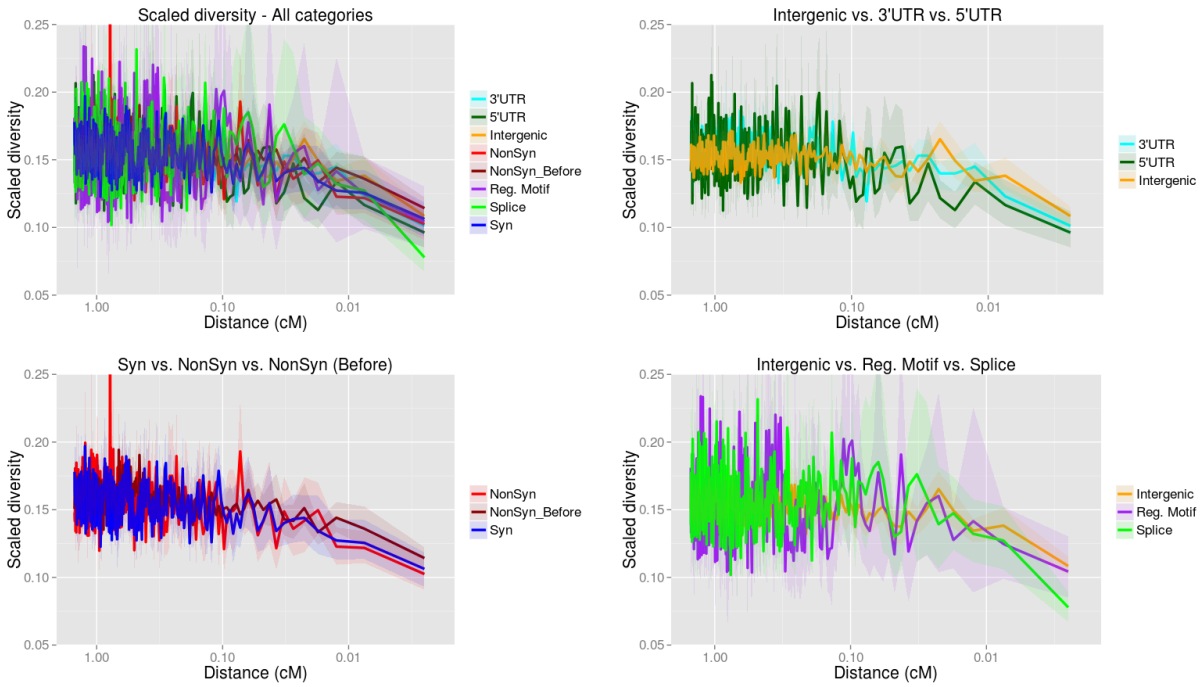

Figure S3. Human diversity per site (calculated in the CG panel) scaled by divergence of the human reference to the human-chimpanzee ancestor around different classes of fixed modern-human-specific single-nucleotied changes where Altai Neanderthal and Denisova are homozygous ancestral. The statistic was calculated in windows of 0.005 cM and the x-axis shows distance of the window midpoint to the fixed change on a log-scale. The upper left panel shows all functional categories tested, while the other panels show different subsets of these for ease of comparison.

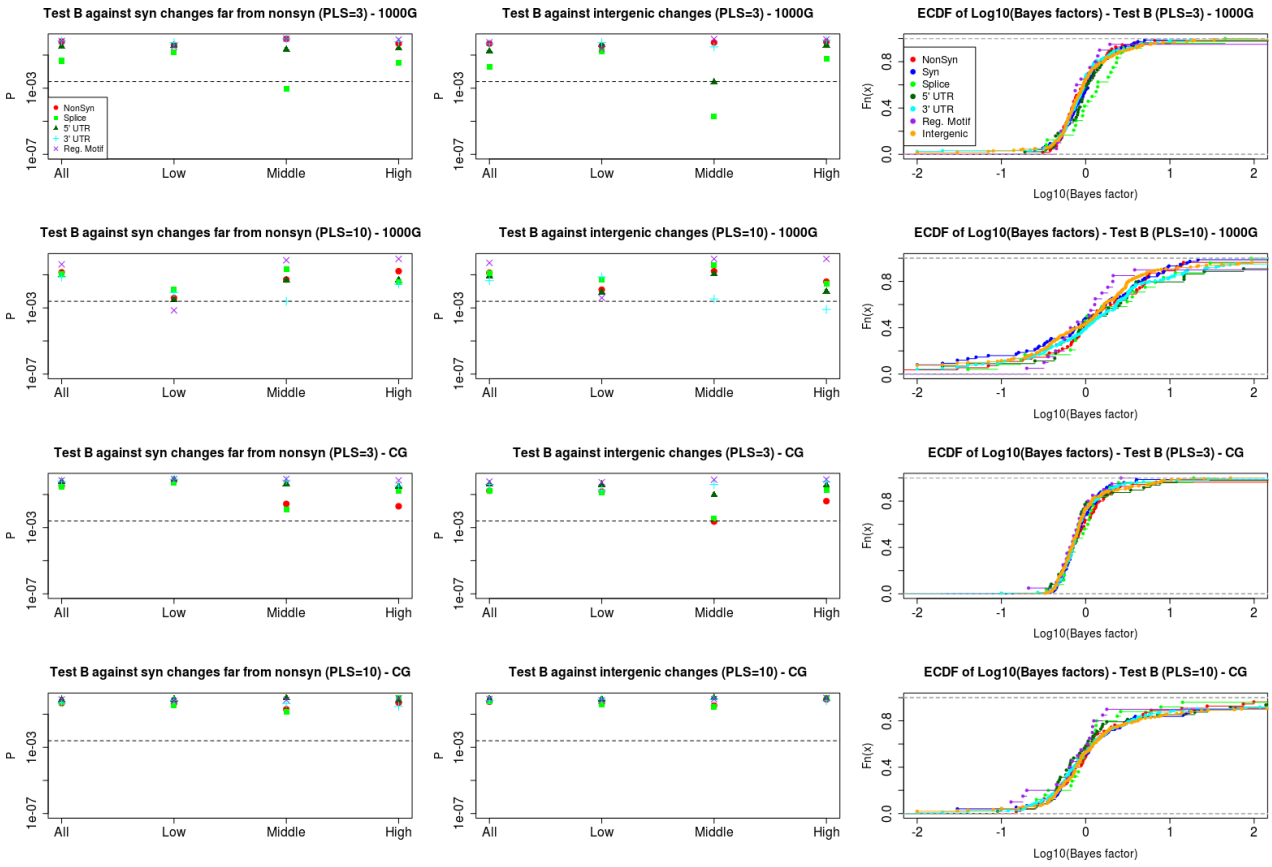

Figure S4. We subs-sampled SNCs within each genomic category so that each SNC was more than 100 kb away from any other. We then tested whether changes in different presumably functional sites have higher Bayes factors in favor of selection relative to synonymous changes that are far ($>$ 1Mb) from any nonsynonymous change (left panels) or relative to intergenic changes (middle panels), using a one-tailed Wilcoxon rank-sum test. The x-axes show different quantile partitions of the data in each of the two categories under comparison. The dashed lines denote the p-values cutoff after correcting for multiple testing ($P = 0.05/20 = 0.0025$). We also show empirical cumulative distribution functions of Bayes factors for each category tested (right panels). First row from top: Test B (including poor model fits) using 1000G data and first 3 PLS-DA components. Second row: Test B using 1000G data and first 10 PLS-DA components. Third row: Test B using CG data and first 3 PLS-DA components. Bottom row: Test B using CG data and first 10 PLS-DA components.

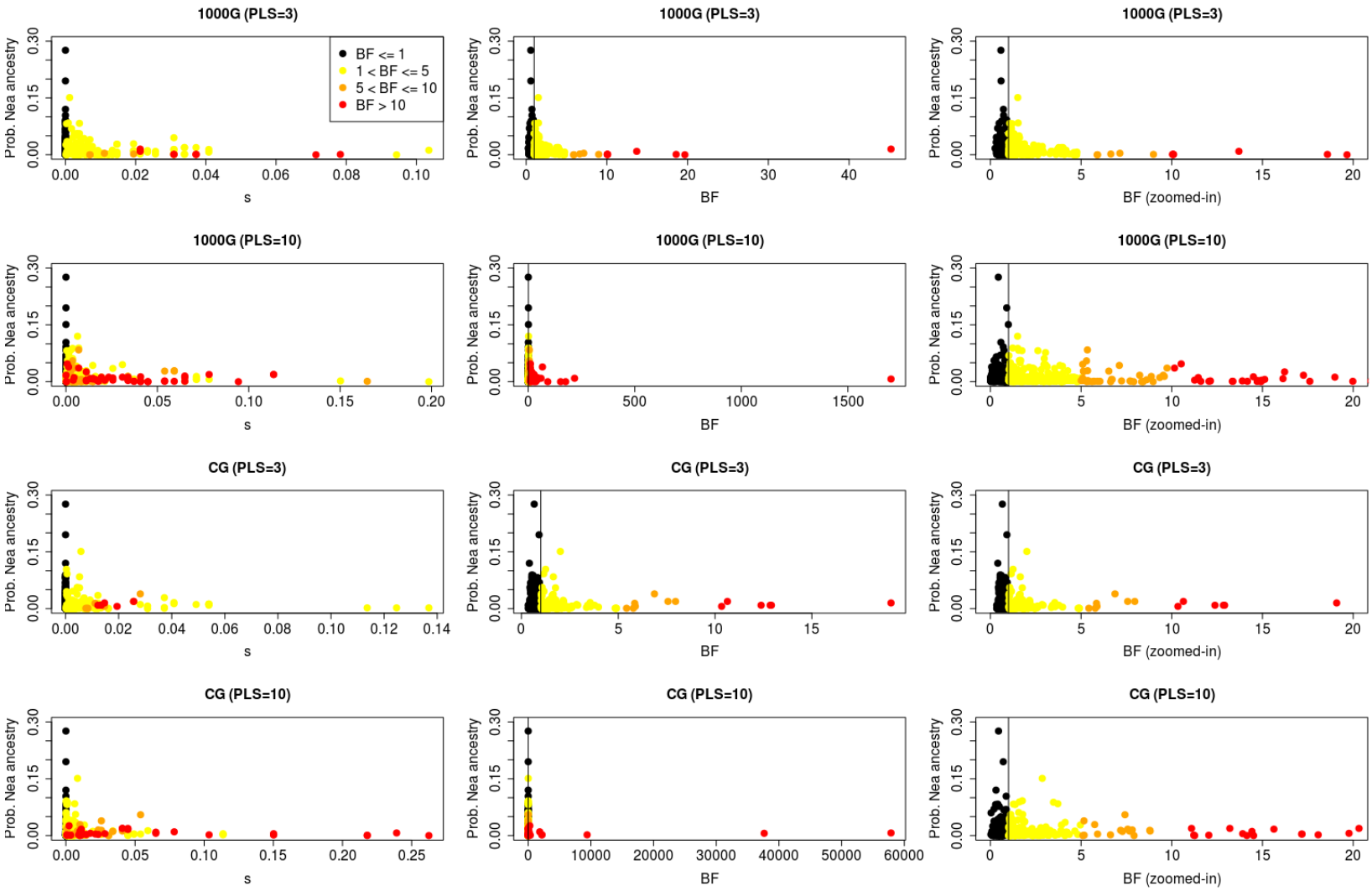

Figure S5. Probability of Neandertal ancestry in Eurasians obtained from Sankararaman et al. (2014) at the nearest informative SNP of each fixed SNC, plotted as a function of each SNC's inferred selection coefficient (left panels) or Bayes factor in favor of selection (middle and right panels). The right panels are zoomed-in versions of the middle panels. First row from top: using 1000G data and first 3 PLS-DA components. Second row: using 1000G data and first 10 PLS-DA components. Third row: using CG data and first 3 PLS-DA components. Bottom row: using CG data and first 10 PLS-DA components.

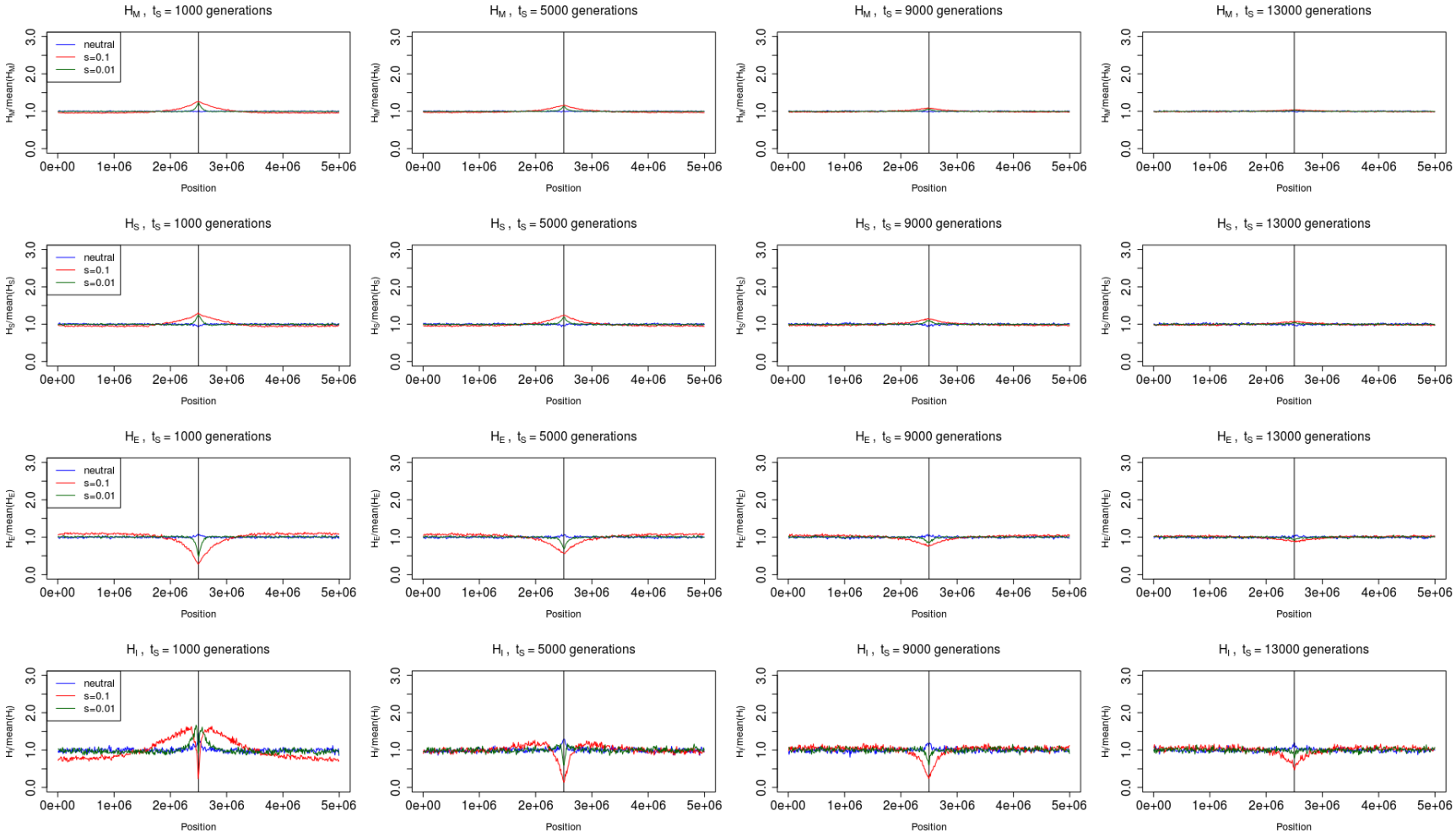

Figure S6. The mean values of the H_E, H_M, H_S and H_I statistics (using 4-SNP blocks) from 200 simulations run under the same parameters were calculated along windows of 100 kb (=0.1 cM) in a 5 Mb region and divided by their mean value along the entire region. For this plot, we sampled 200 present-day human sequences in each simulation.

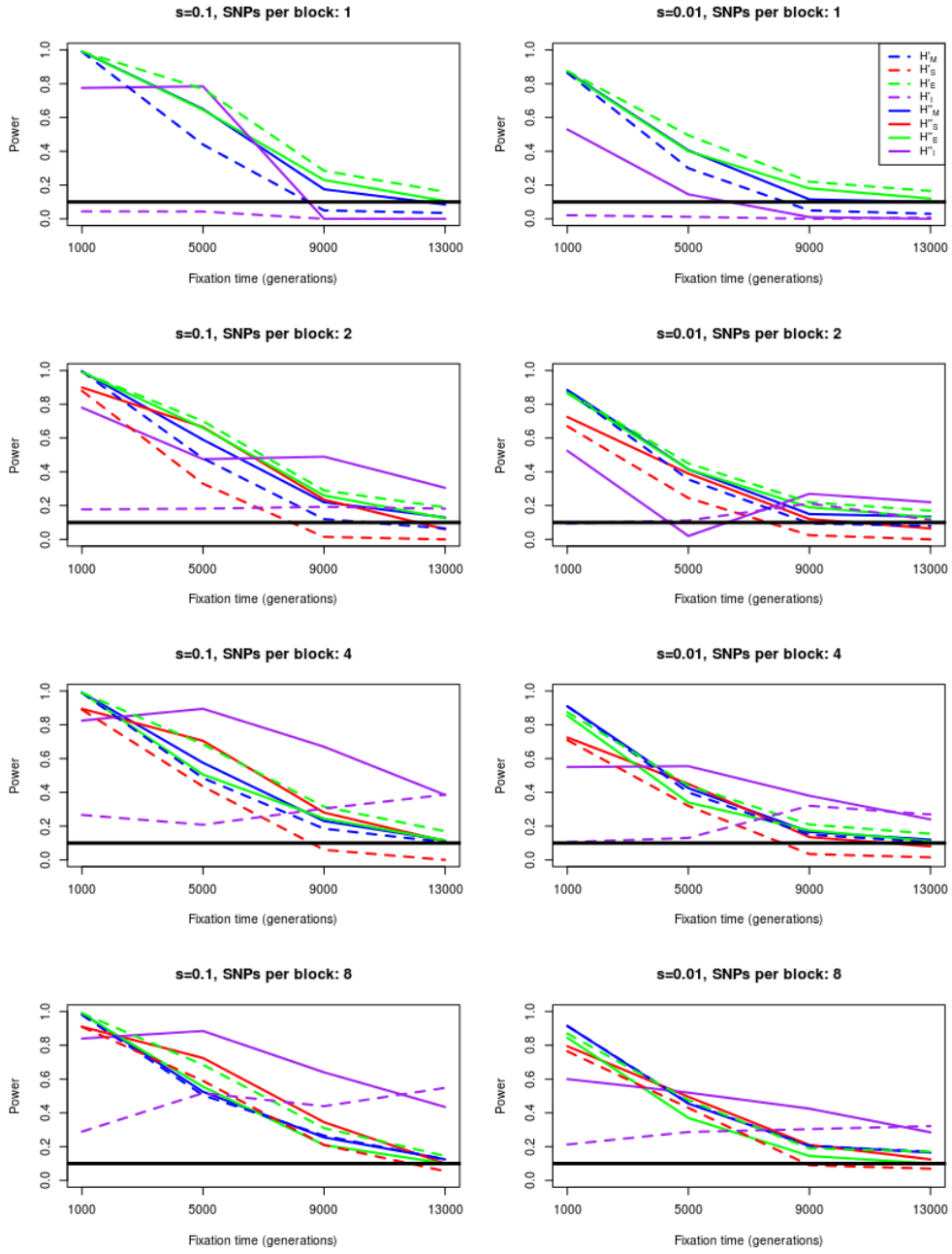

Figure S7. Power to reject neutrality for different statistics and different number of SNPs per block in the case when 200 modern human sequences are available (like in the 1000G data). We tested two different selection regimes: s=0.1 (left column) and s=0.01 (right column). We also tested a range of times since fixation (x-axis). Power was estimated by calculating the proportion of simulations (out of 200) that have a value more extreme (higher for H'_M, H''_M, H'_S and H''_S; lower for H'_E, H''_E, H'_I and H''_I) than 90% of 200 neutral simulations with the same fixation time. Skewness is not shown for blocks of size 1 SNP because the sample third moment of a count vector of size two is always zero, so the statistic is meaningless in that case. The thick black line denotes the 10% rejection level.

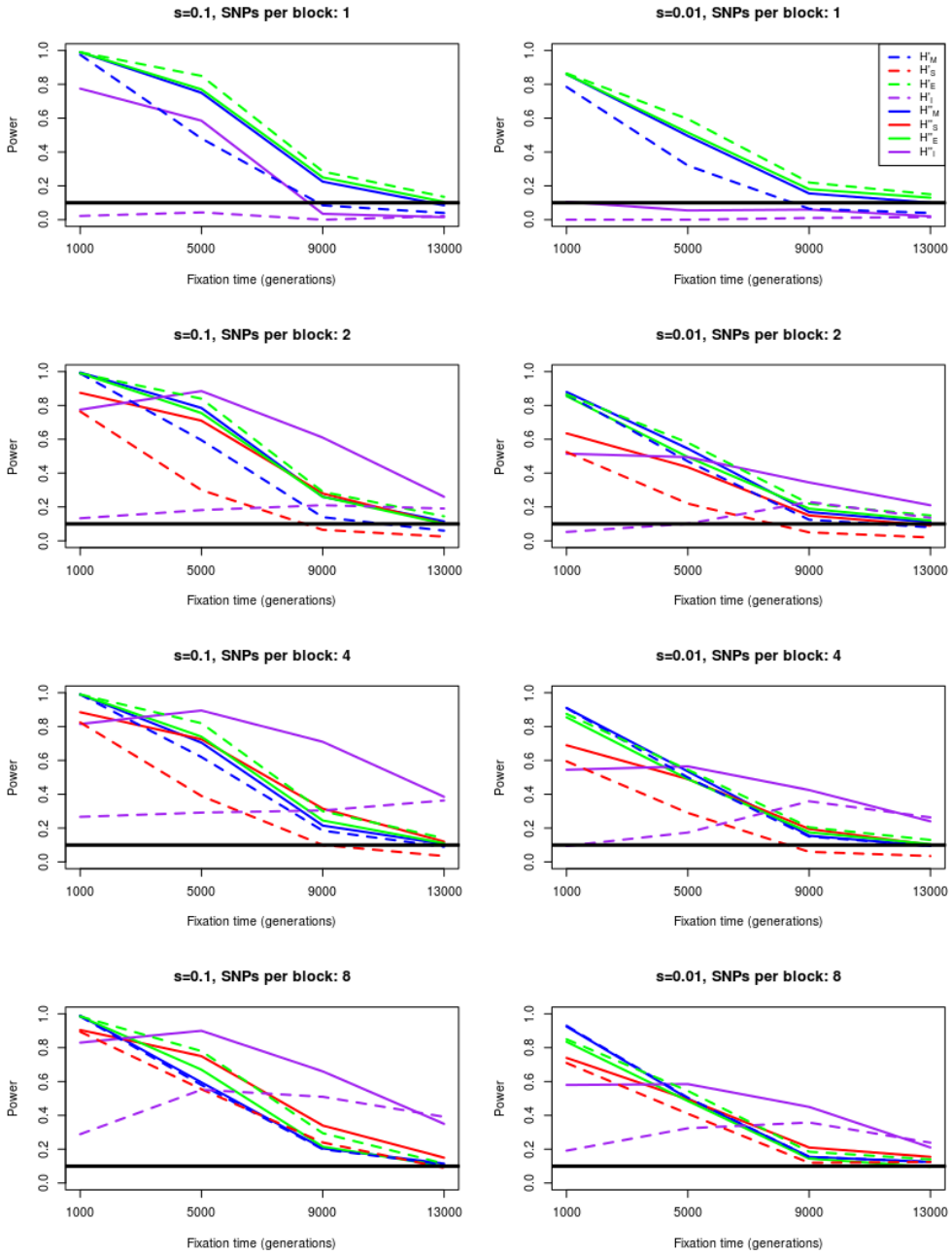

Figure S8. Power to reject neutrality for different statistics and different number of SNPs per block in the case when 200 modern human sequences are available (like in the 1000G data). We tested two different selection regimes: s=0.1 (left column) and s=0.01 (right column). We also tested a range of times since fixation (x-axis). Power was estimated by calculating the proportion (out of 200) of selective simulations with a particular fixation time (x-axis) that have a value more extreme (higher for H'_M, H''_M, H'_S and H''_S; lower for H'_E, H''_E, H'_I and H''_I) than 90% of 800 neutral simulations with different times of fixation (200 with t=1000 gen., 200 with t=5000 gen., 200 with t=9000 gen. and 200 with t=13000 gen.). Skewness is not shown for blocks of size 1 SNP because the sample third moment of a count vector of size two is always zero, so the statistic is meaningless in that case. The thick black line denotes the 10% rejection level.

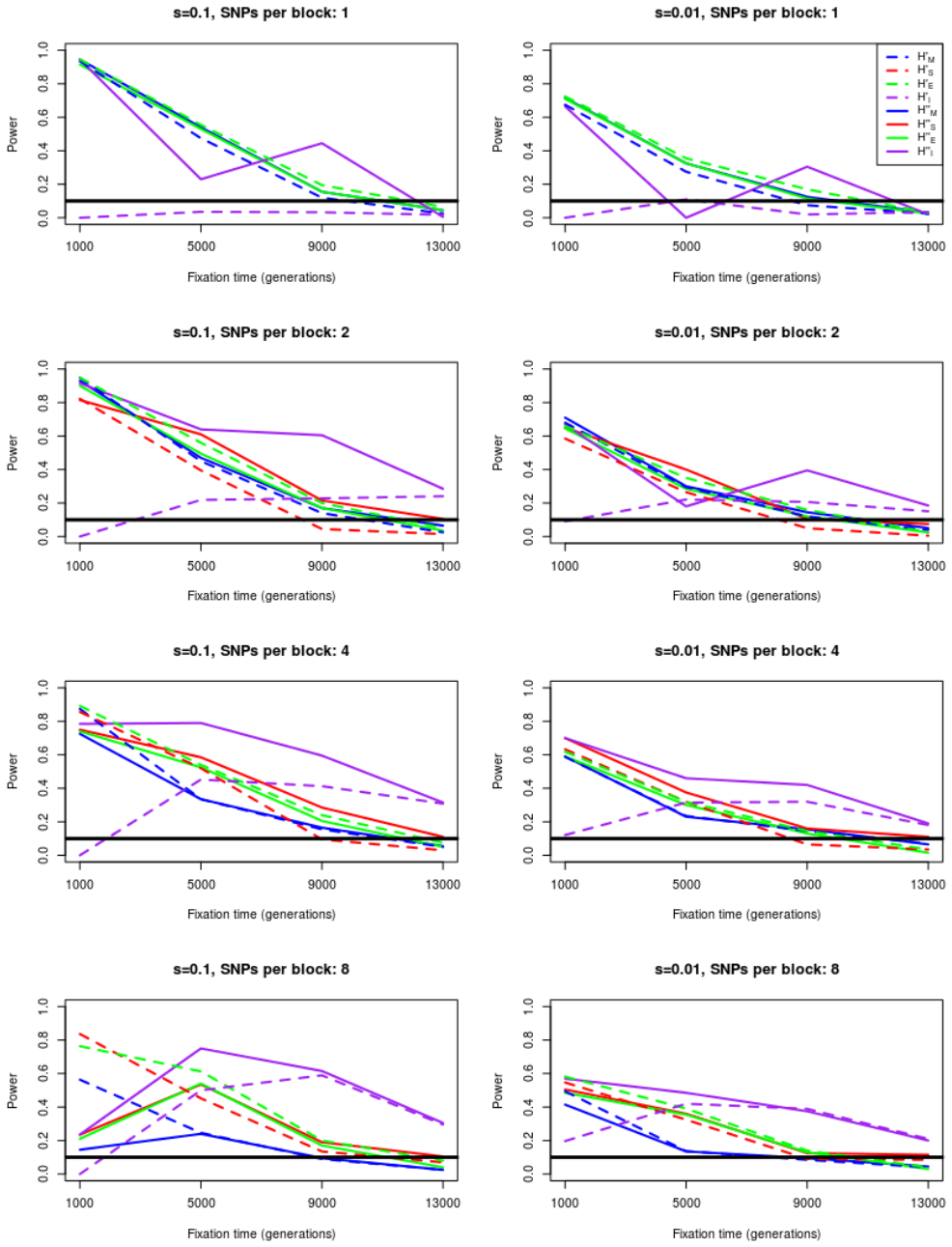

Figure S9. Power to reject neutrality for different statistics and different number of SNPs per block in the case when 26 modern human sequences are available (like in the CG data). We tested two different selection regimes: s=0.1 (left column) and s=0.01 (right column). We also tested a range of times since fixation (x-axis). Power was estimated by calculating the proportion of simulations (out of 200) that have a value more extreme (higher for H'_M, H''_M, H'_S and H''_S; lower for H'_E, H''_E, H'_I and H''_I) than 90% of 200 neutral simulations with the same fixation time. Skewness is not shown for blocks of size 1 SNP because the sample third moment of a count vector of size two is always zero, so the statistic is meaningless in that case. The thick black line denotes the 10% rejection level.

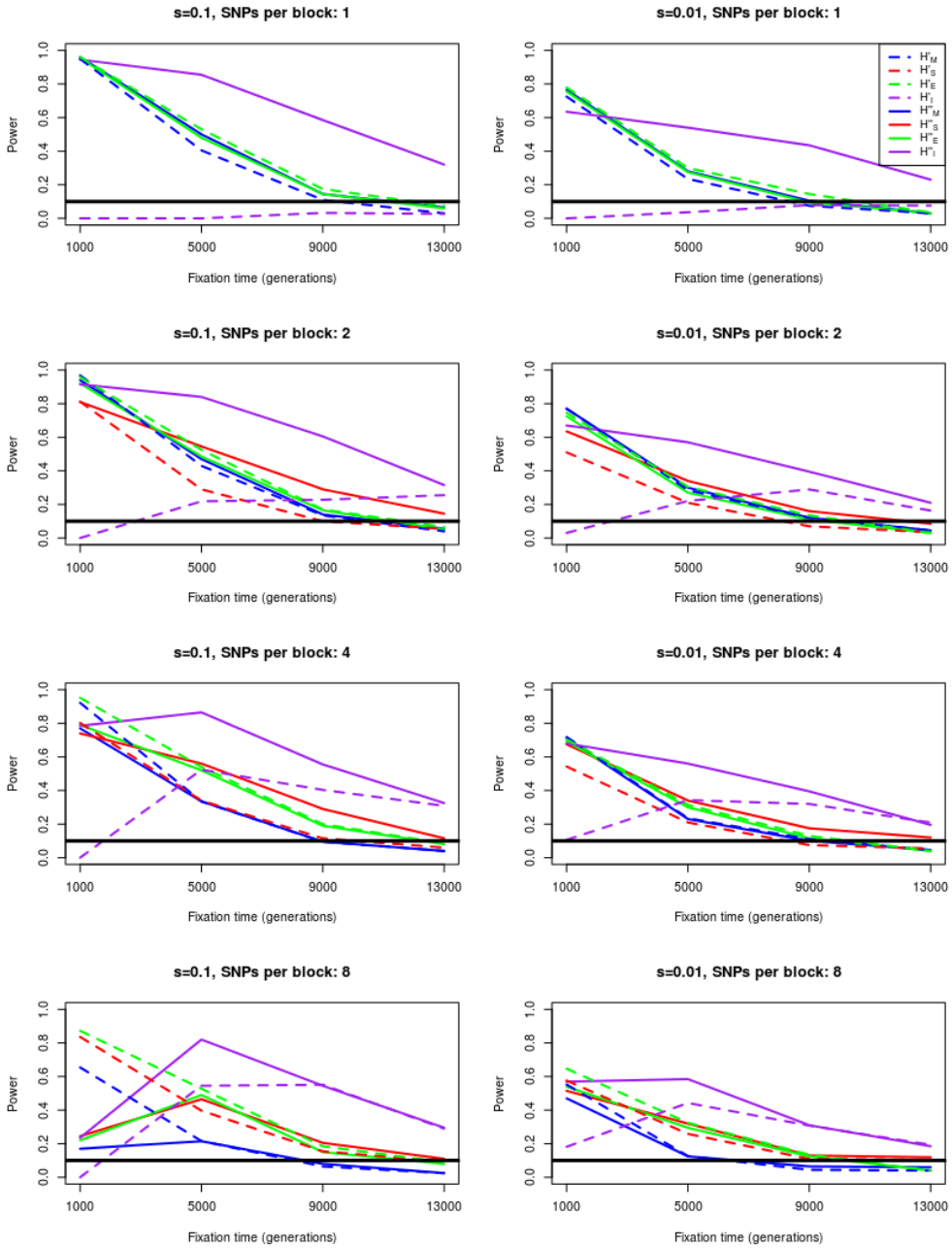

Figure S10. Power to reject neutrality for different statistics and different number of SNPs per block in the case when 26 modern human sequences are available (like in the CG data). We tested two different selection regimes: s=0.1 (left column) and s=0.01 (right column). We also tested a range of times since fixation (x-axis). Power was estimated by calculating the proportion (out of 200) of selective simulations with a particular fixation time (x-axis) that have a value more extreme (higher for H'_M, H''_M, H'_S and H''_S; lower for H'_E, H''_E, H'_I and H''_I) than 90% of 800 neutral simulations with different times of fixation (200 with t=1000 gen., 200 with t=5000 gen., 200 with t=9000 gen. and 200 with t=13000 gen.). Skewness is not shown for blocks of size 1 SNP because the sample third moment of a count vector of size two is always zero, so the statistic is meaningless in that case. The thick black line denotes the 10% rejection level.

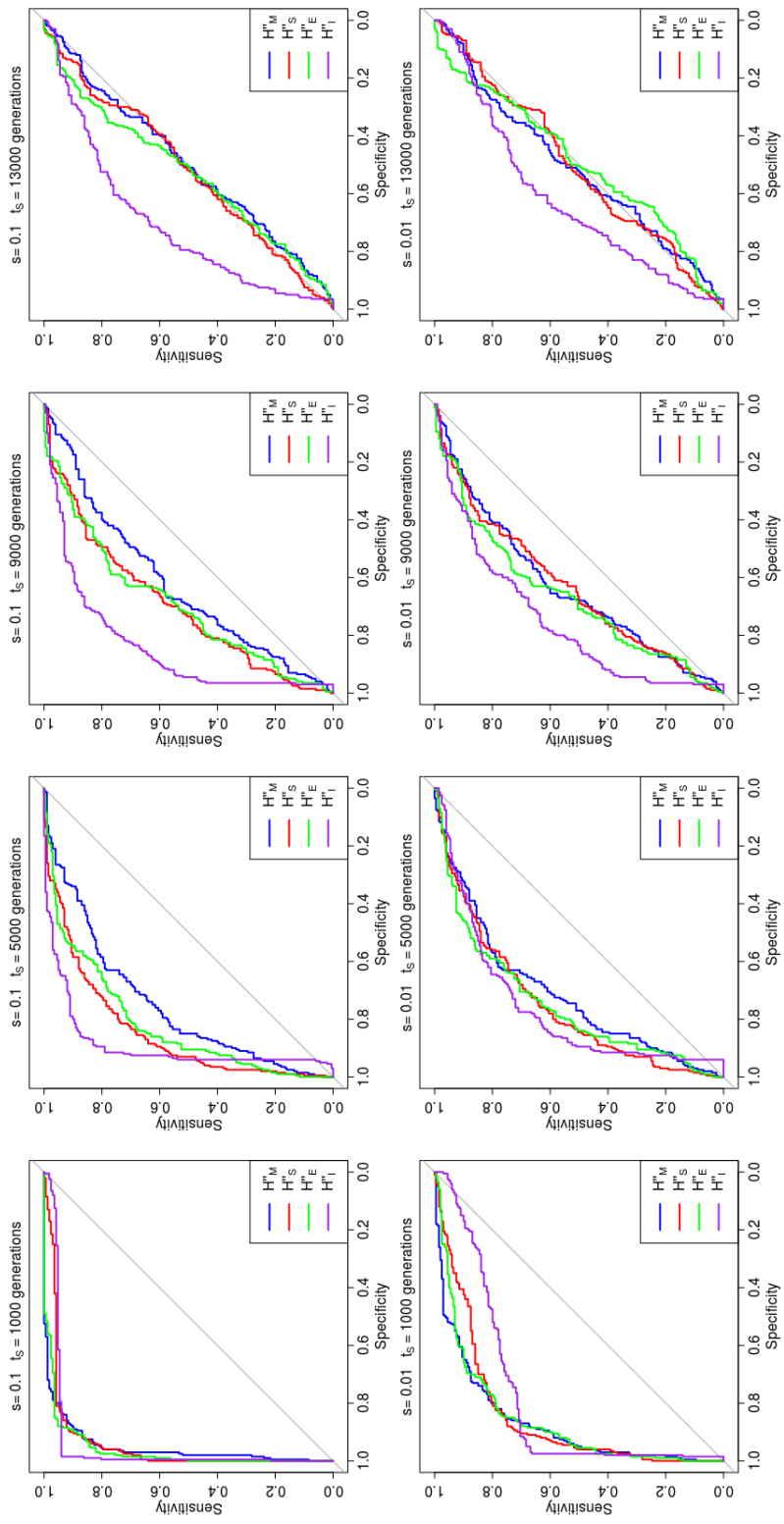

Figure S11. Receiver operating characteristic (ROC) curves showing performance in rejecting neutrality for different statistics (with SNP blocks of size 4) under different selection coefficients and times since fixation, when 26 modern human sequences are available (like in the CG data).

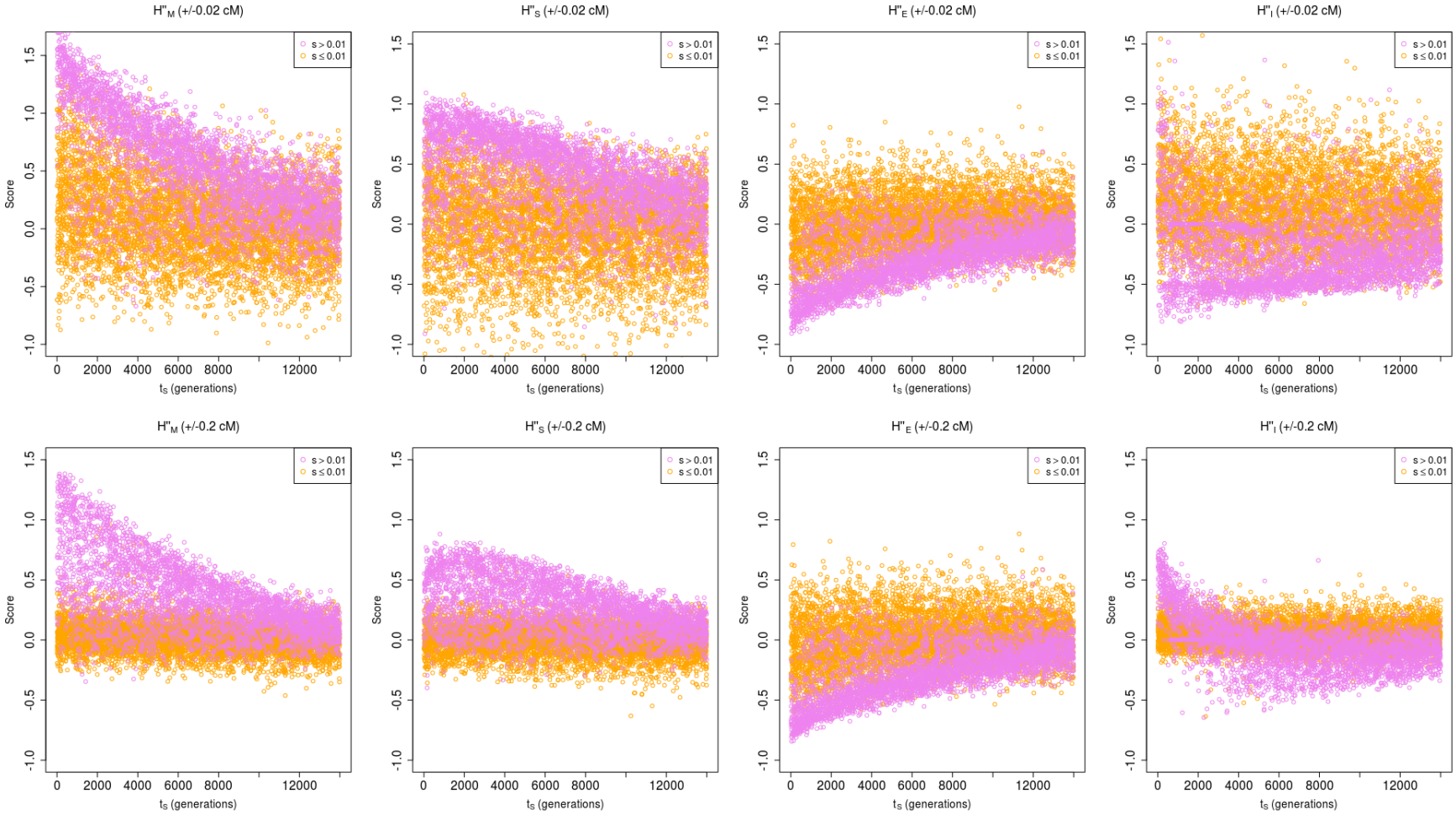

Figure S12. H''_E, H''_M, H''_S and H''_I a a function of t_S (here in units of $4N_e$ generations), computed on 10,000 simulations. The colors correspond to the strength of the sweep event: $s > 0.01$ (orange) or $s < 0.01$ (violet). Top row: using an internal region 0.04 cM long around the candidate site. Bottom row: using an internal region 0.4 cM long around the candidate site.

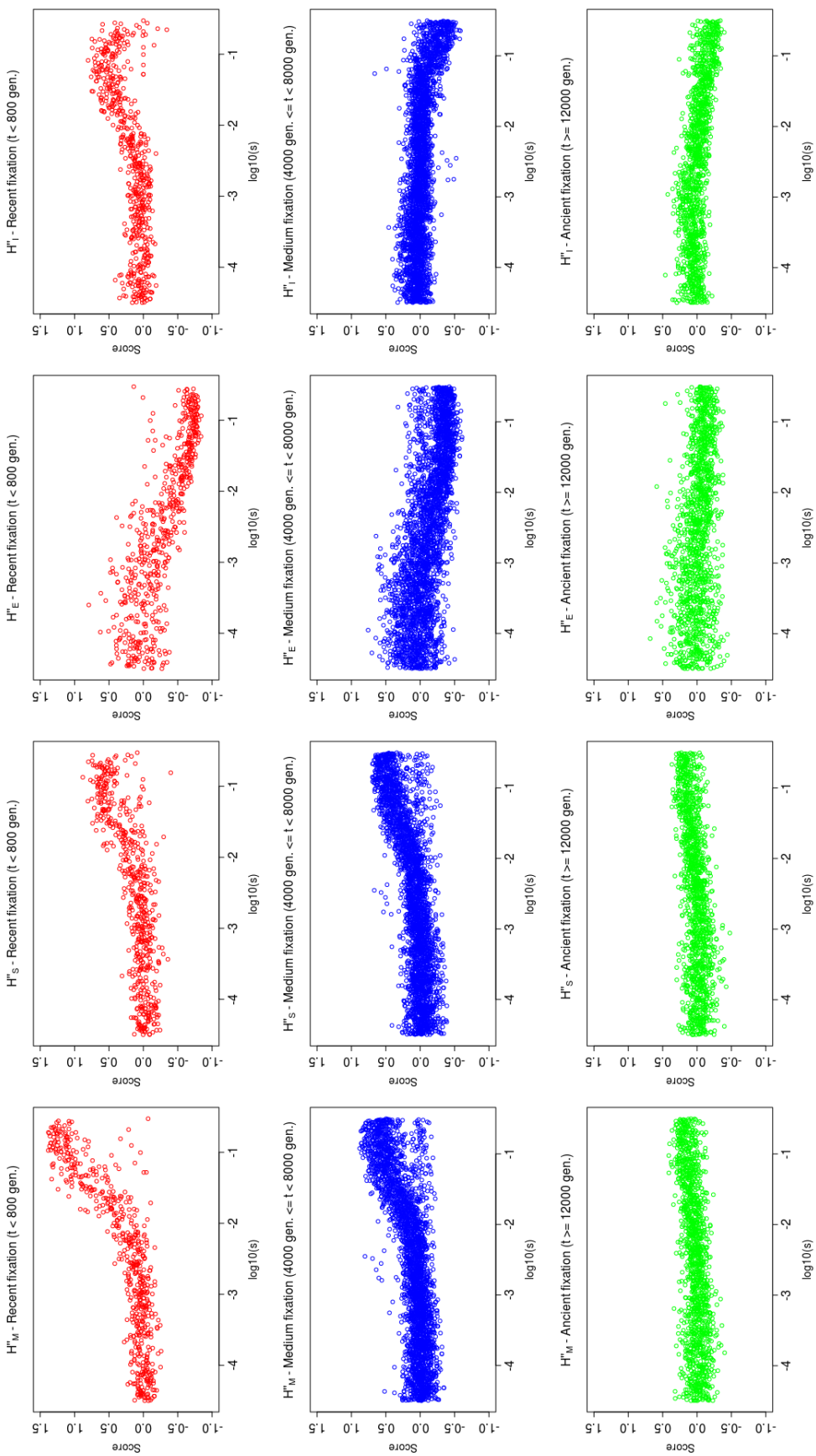

Figure S13. H''_E, H''_M, H''_S and H''_I as a function of log(s). The colors correspond to the timing of the selective event: a recent sweep ($t_S < 800$ gen., red), an older sweep (2000 gen. $< t_S < 8000$ gen., blue) or a very ancient sweep ($t_S > 12000$ gen., green). In all cases, the internal region used to obtain the statistics was 0.4 cM long.

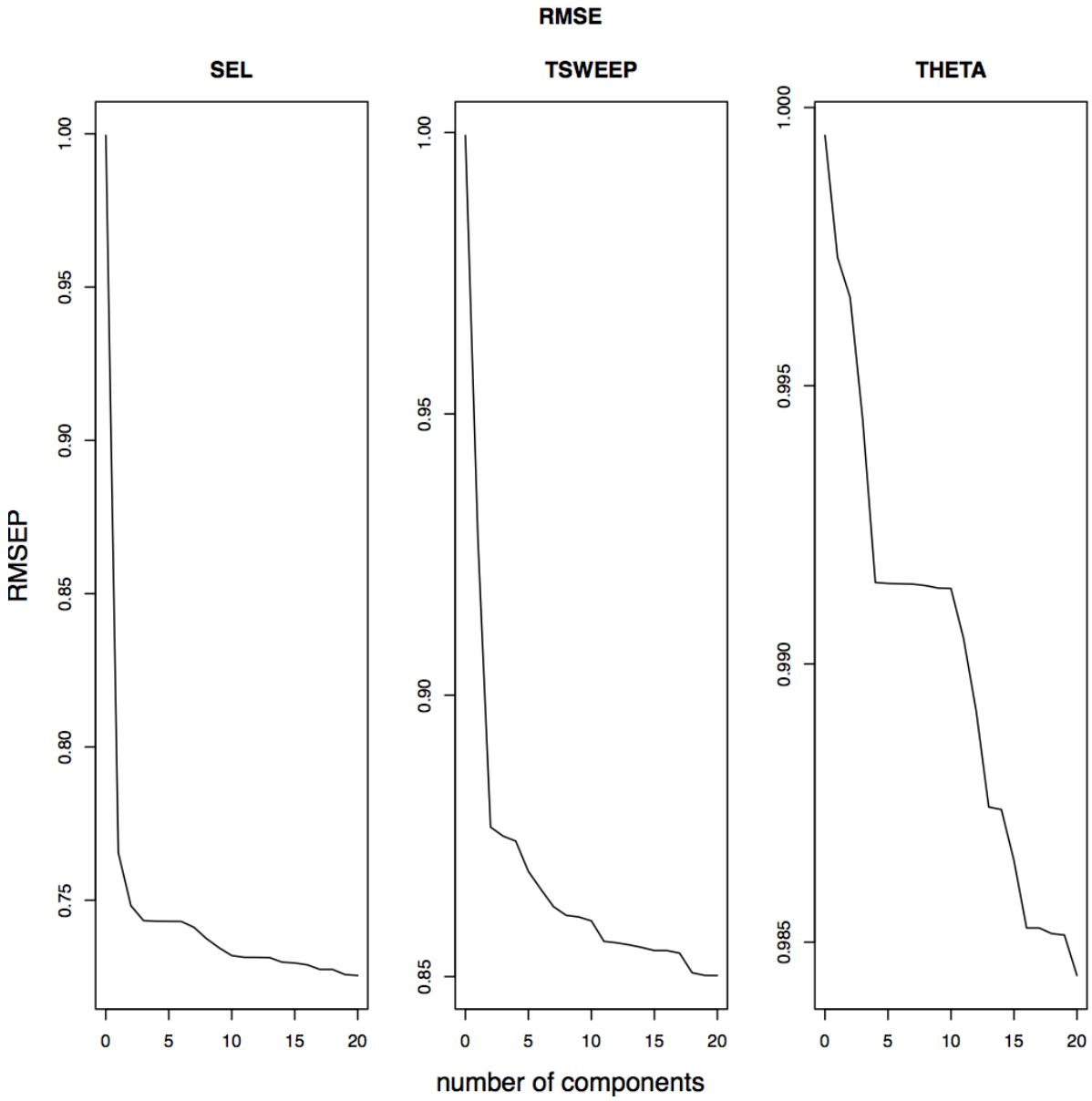

Figure S14. Root mean squared error plots (RMSEP), showing the decrease in RMSE in the first 20 PLS components extracted from the summary statistics, for each parameter on which we placed a prior.

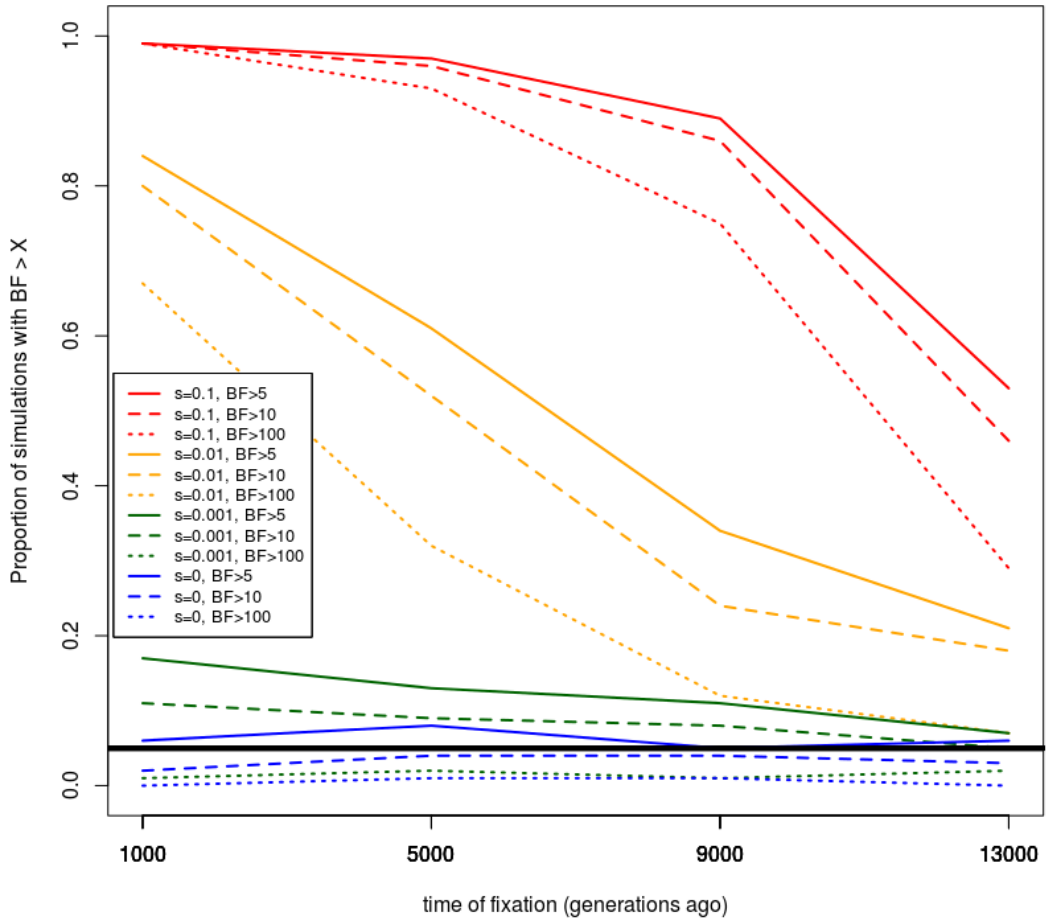

Figure S15. Sets of 100 simulations were run through the ABC pipeline to obtain Bayes factors in favor of selection (versus neutrality) under different known parameters (PLSDA = 10). The colored lines show the proportion of the simulations that have a Bayes factor larger than the specified cutoffs, when 26 present-day human sequences are available. The thick black line denotes the 0.05 significance cutoff. BF = Bayes factor, s=selection coefficient, t=time since derived allele fixation, in generations.

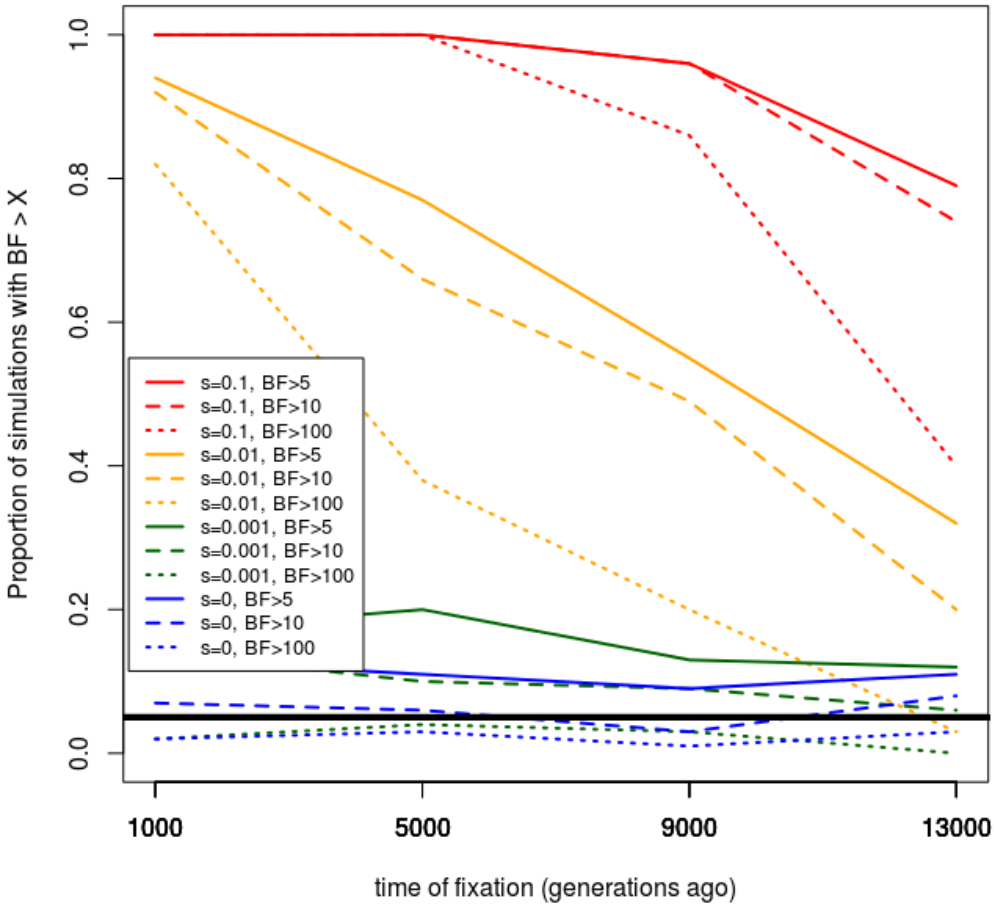

Figure S16. Sets of 100 simulations were run through the ABC pipeline to obtain Bayes factors in favor of selection (versus neutrality) under different known parameters (PLSDA = 10). Here, we simulated the case where two datasets are avaialble: one with 200 sequences (like the 1000G dataset) and one with 26 sequences (like the CG dataset). The colored lines show the proportion of the simulations where the maximum Bayes factor across the two datasets is larger than the specified cutoffs (as in Table 1). The thick black line denotes the 0.05 significance cutoff. BF = Bayes factor, s=selection coefficient, t=time since derived allele fixation, in generations.

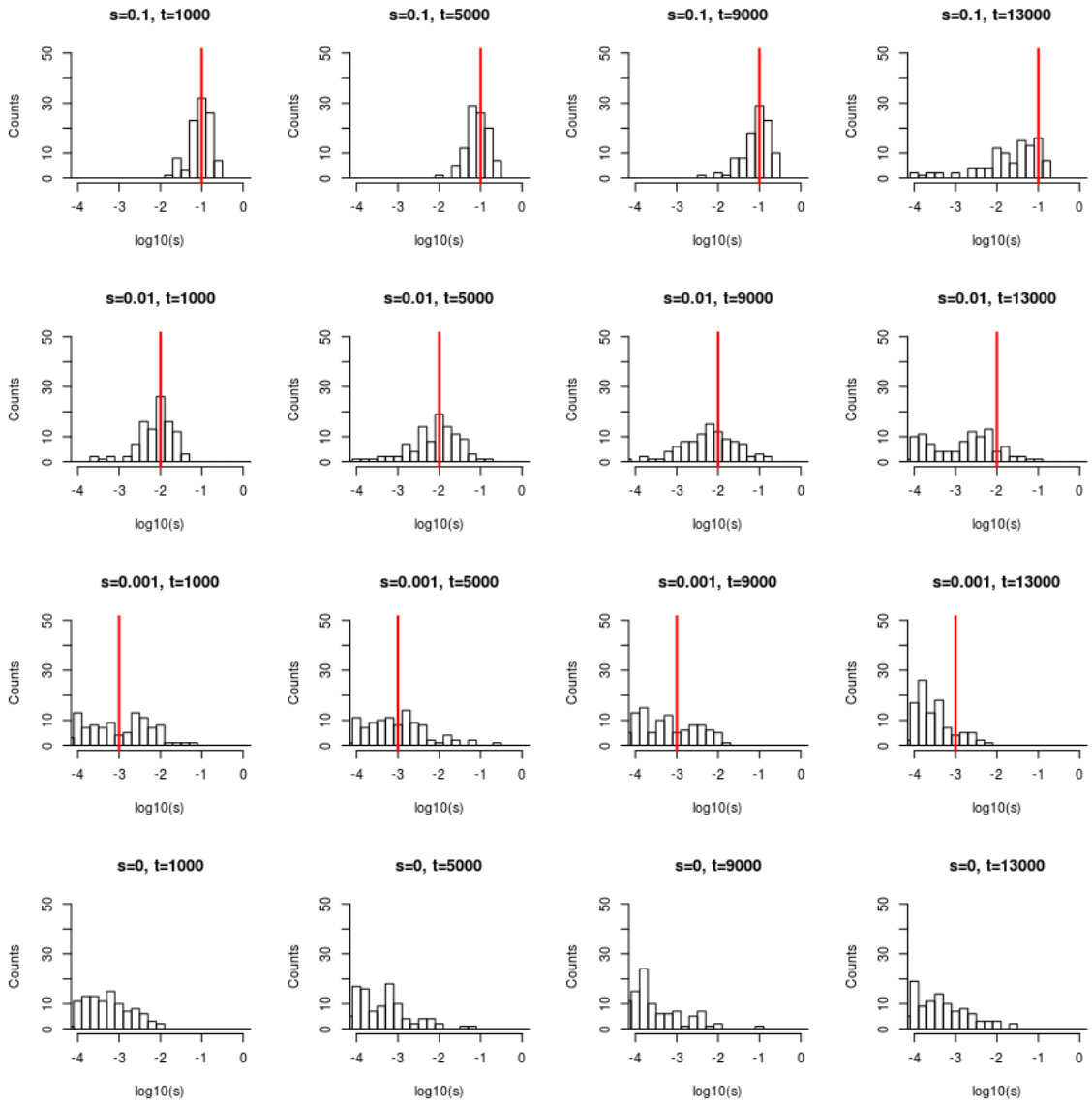

Figure S17. Sets of 100 simulations were run through the ABC pipeline (with 10 PLS components) to infer the selection coefficients under different parameters, assuming 200 sequences were sampled (as in the 1000G data). The red line represents the true value of $log_{10}(s)$, specified in the simulations. The histograms represent the posterior modes inferred for that parameter. s=selection coefficient, t=time since derived allele fixation, in generations.

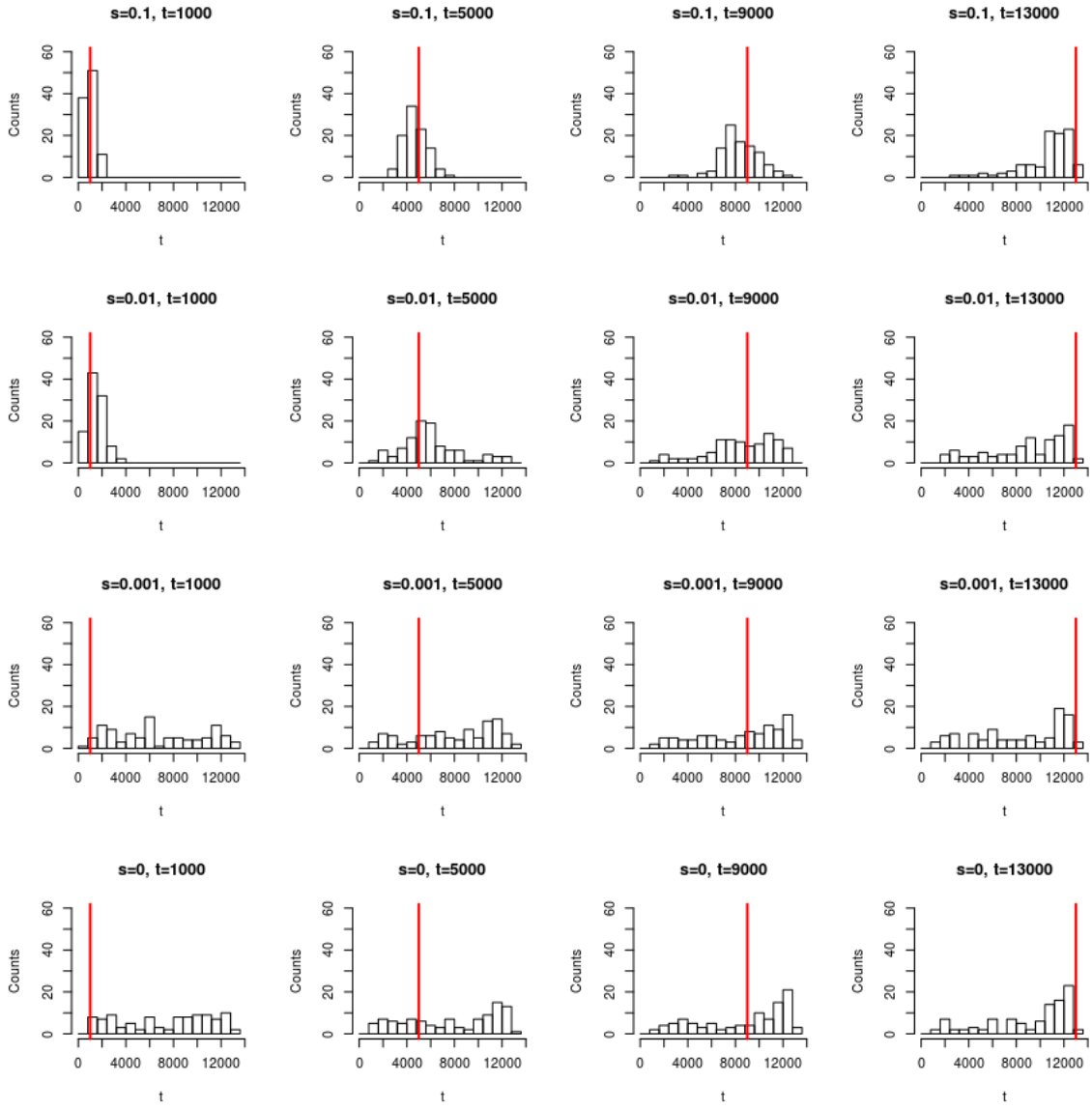

Figure S18. Sets of 100 simulations were run through the ABC pipeline (with 10 PLS components) to infer the selection coefficients under different parameters, assuming 200 sequences were sampled (as in the 1000G data). The red line represents the true value of the time of fixation of the derived allele, specified in the simulations. The histograms represent the posterior modes inferred for that parameter. s=selection coefficient, t=time since derived allele fixation, in generations.

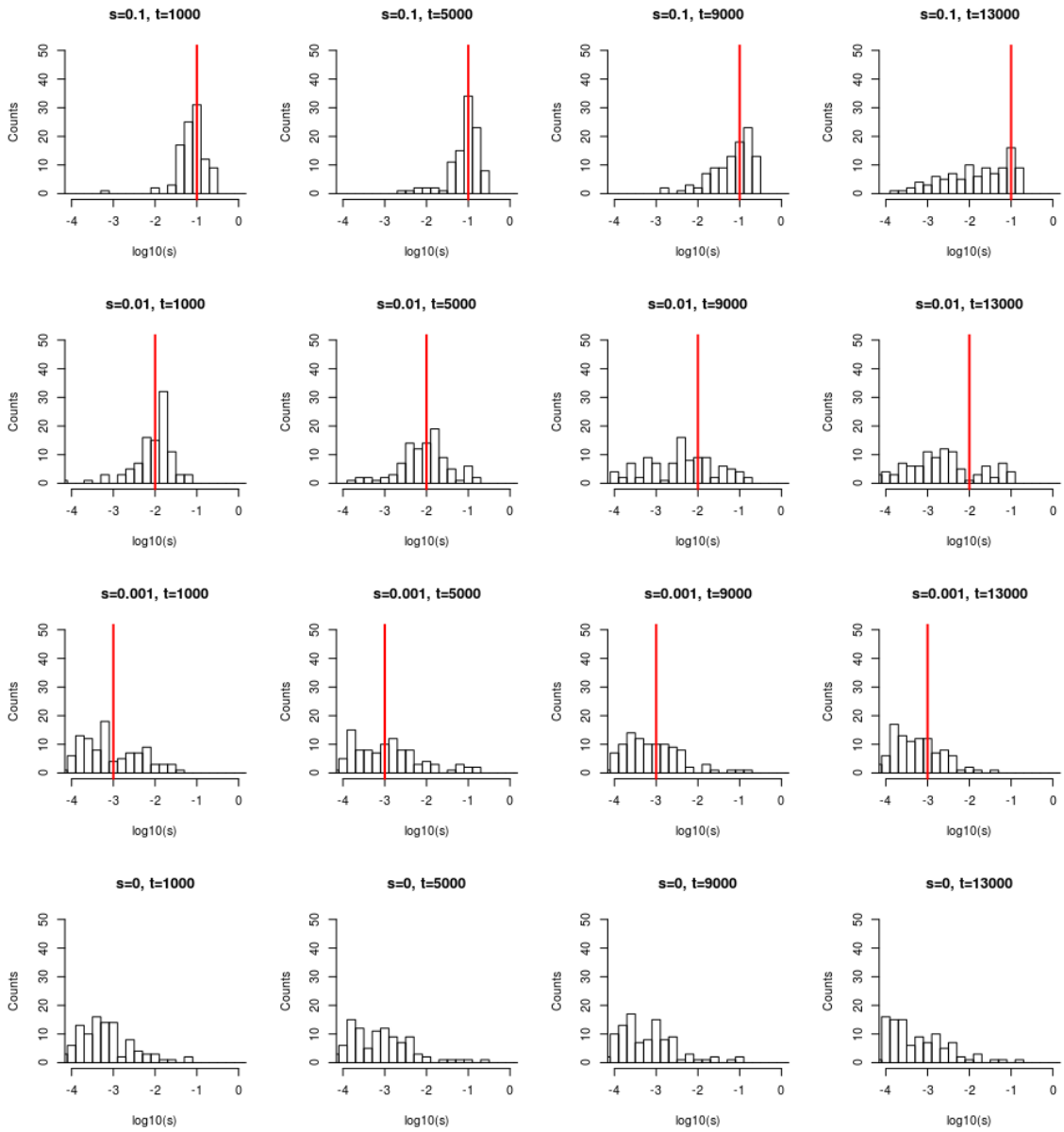

Figure S19. Sets of 100 simulations were run through the ABC pipeline (with 10 PLS components) to infer the selection coefficients under different parameters, assuming 26 sequences were sampled (as in the CG data). The red line represents the true value of $log_{10}(s)$, specified in the simulations. The histograms represent the posterior modes inferred for that parameter. s=selection coefficient, t=time since derived allele fixation, in generations.

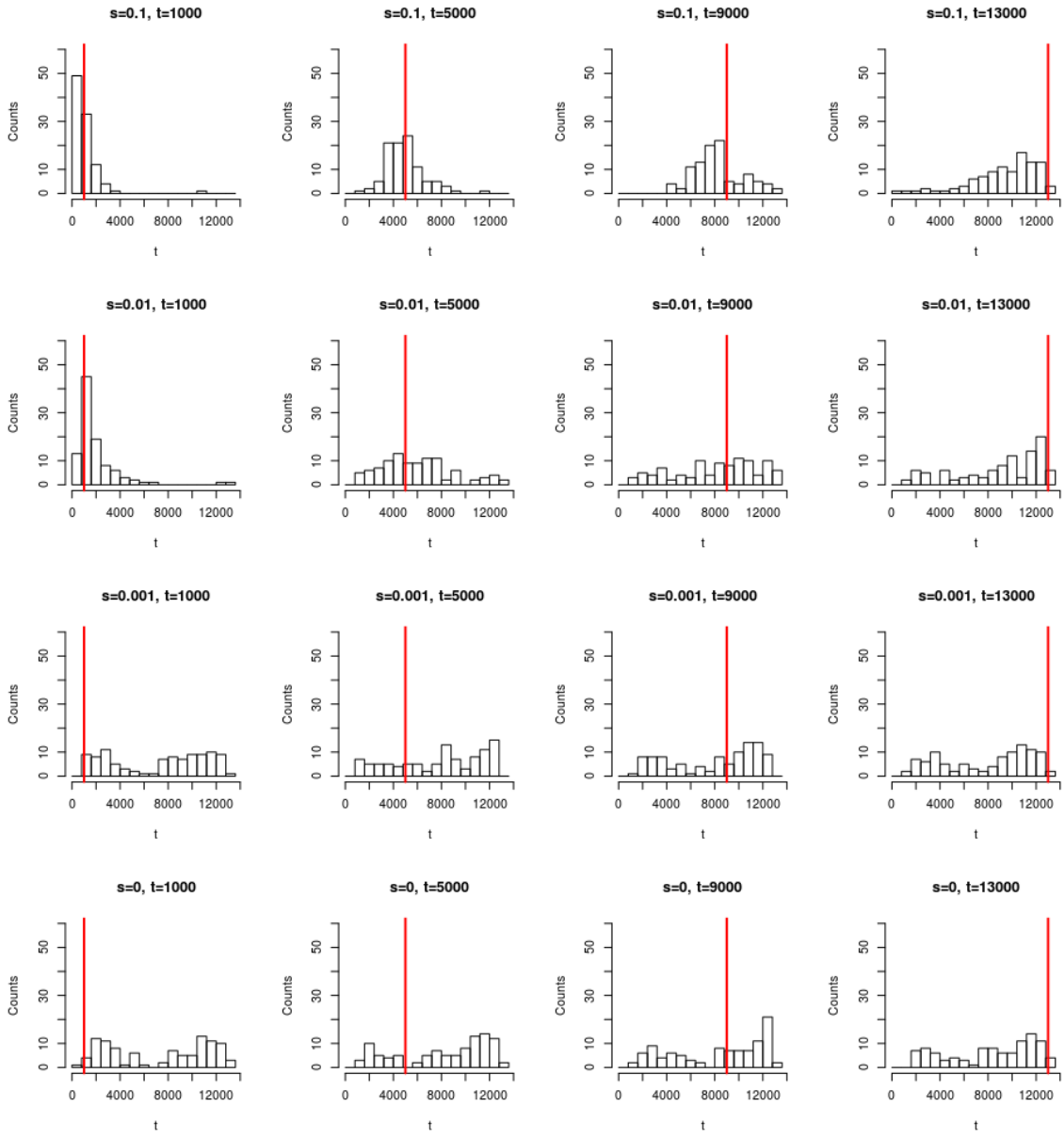

Figure S20. Sets of 100 simulations were run through the ABC pipeline (with 10 PLS components) to infer the selection coefficients under different parameters, assuming 26 sequences were sampled (as in the CG data). The red line represents the true value of the time of fixation of the derived allele, specified in the simulations. The histograms represent the posterior modes inferred for that parameter. s=selection coefficient, t=time since derived allele fixation, in generations.

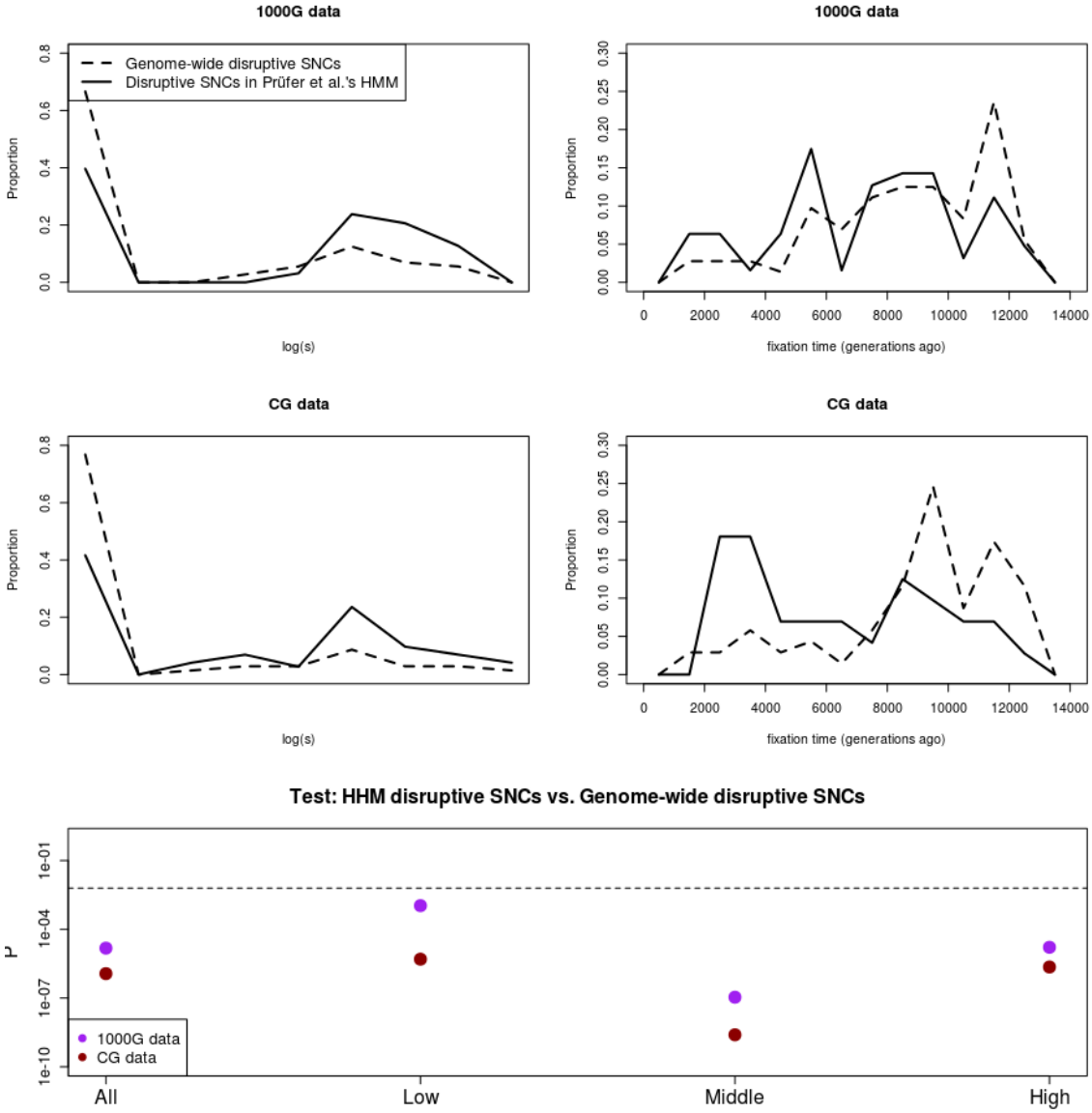

Figure S21. We applied our ABC method (with the first 3 PLS/PLSDA components) to the list of 100 most disruptive SNCs in Prüfer et al. (2014)'s HMM selective sweep screen. We compared the inferred parameters and Bayes factors for this list against the inferred parameters and Bayes factors inferred for the 100 most disruptive SNCs genome-wide. Disruptiveness was determined using the C-score method developed in Kircher et al. (in press) and used in Prüfer et al. (2014). As expected, disruptive SNCs in the HMM regions have larger log(s) and Bayes factors in favor of selection across different quantiles than the genome-wide disruptive SNCs. We have more power to identify the regions as selected using the 1000G data (upper panels and purple dots in lower panel) than when using the CG data (middle panels and dark red dots in lower panel). The dashed black line in the lower panel denotes the P-value cutoff after correcting for multiple testing: P=0.05/8=0.00625.

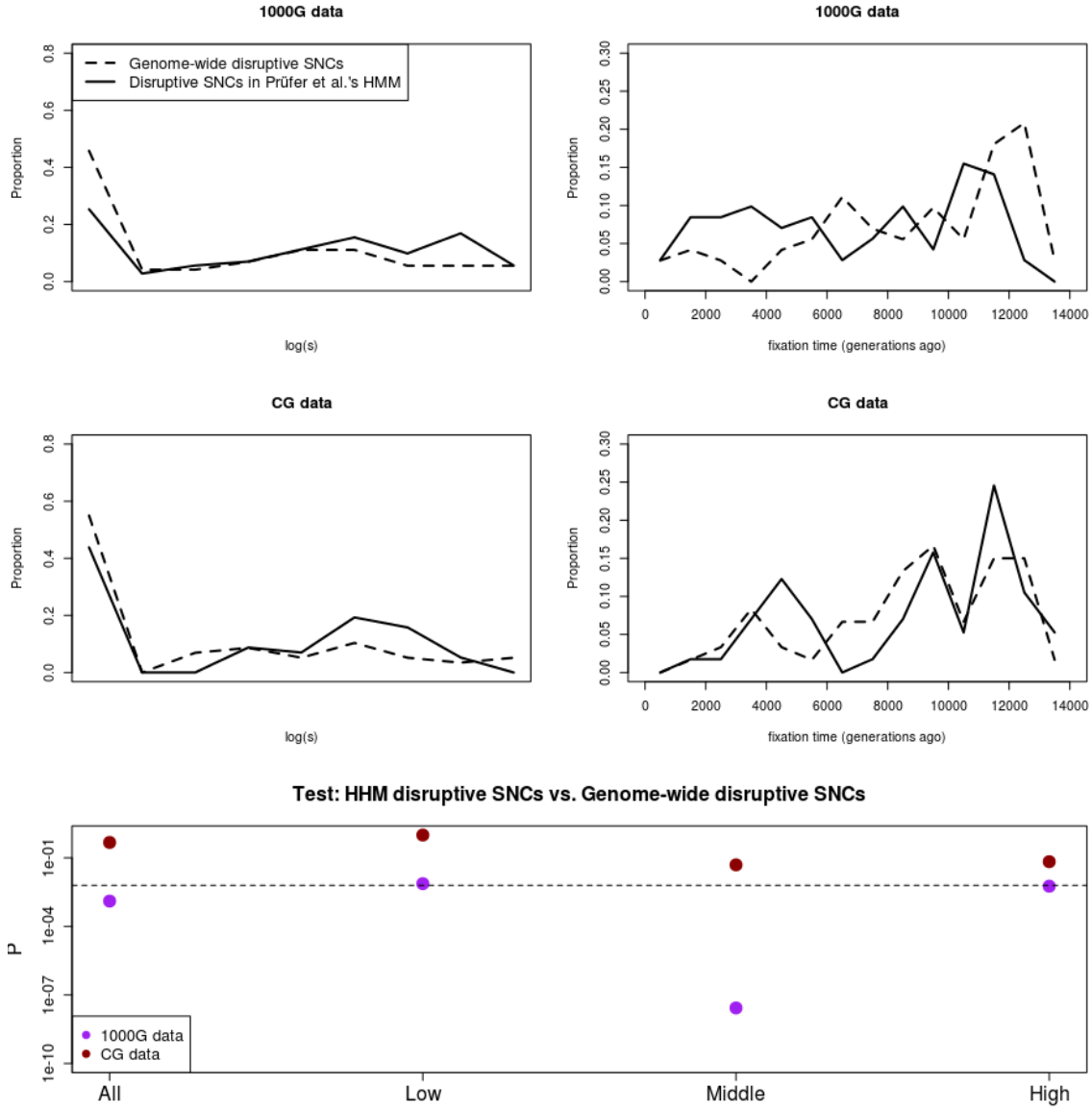

Figure S22. We applied our ABC method (with the first 10 PLS/PLSDA components) to the list of 100 most disruptive SNCs in Prüfer et al. (2014)'s HMM selective sweep screen. We compared the inferred parameters and Bayes factors for this list against the inferred parameters and Bayes factors inferred for the 100 most disruptive SNCs genome-wide. Disruptiveness was determined using the C-score method developed in Kircher et al. (in press) and used in Prüfer et al. (2014). As expected, disruptive SNCs in the HMM regions have larger log(s) and Bayes factors in favor of selection across different quantiles than the genome-wide disruptive SNCs. We have more power to identify the regions as selected using the 1000G data (upper panels and purple dots in lower panel) than when using the CG data (middle panels and dark red dots in lower panel). The dashed black line in the lower panel denotes the P-value cutoff after correcting for multiple testing: P=0.05/8=0.00625.

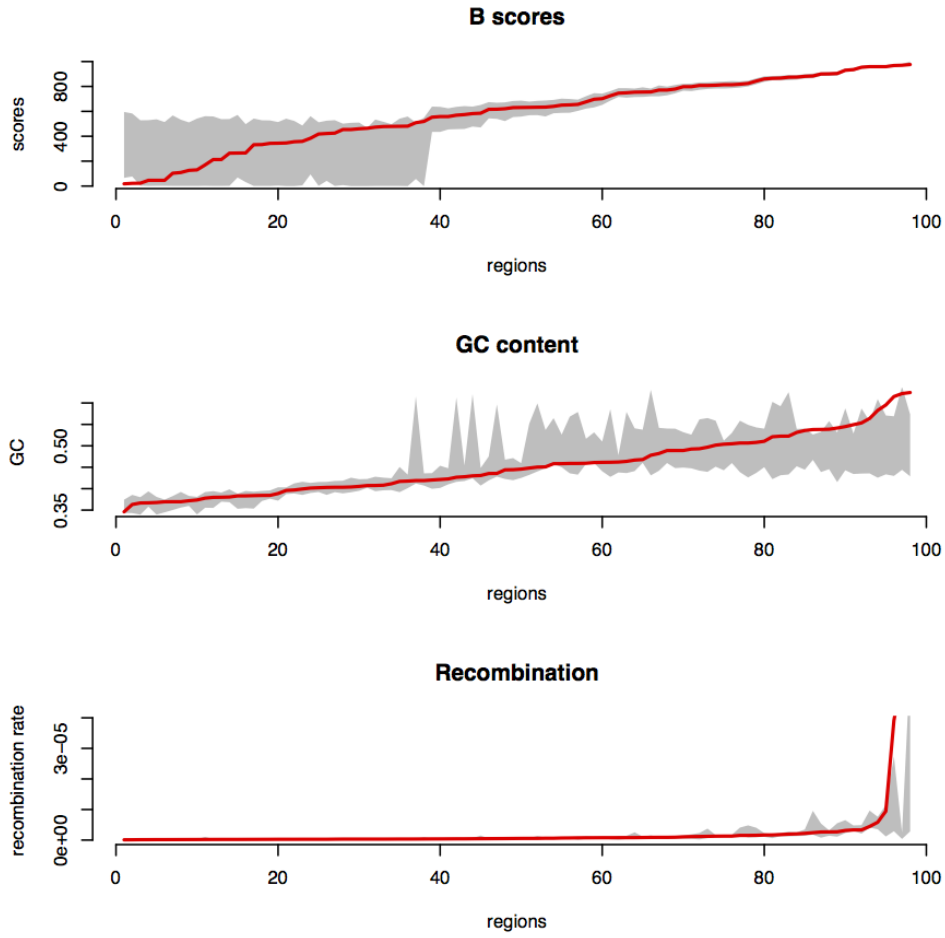

Figure S23. Distribution of scores for nonsynonymous-SNC-matched region filters in which we used "top best-matching" criteria. The red line shows the real value of the region containing a nonsynonymous SNC. The grey shade shows the distribution of the top X% best-matching regions that we were able to sample from the genome. X=10 for B-scores, while X=25 for GC content and recombination rates.

Table S1. Modern-human specific changes that lead to an amino acid replacement, affect a splice site or are located in a UTR, and that: 1) have Bayes factors > 10 in favor of selection using either the 1000G and CG datasets and 2) are a good fit ($P > 0.05$) to the selection model using both the 1000G and CG datasets. Parameters listed are the posterior modes inferred using ABC. The Bayes factor shown for each site is the maximum across the two datasets. t_S is in generations. All logs are base 10. 1K: 1000 Genomes. CG: Complete Genomics. BF: Bayes factor.

Position	log(BF)	log(s) (1K)	log(s) (CG)	t_S (1K)	t_S (CG)	Class	Gene
chr1:38423232	1.05	-1.95	-2.6	9476	10322	3' UTR	SF3A3
chr1:78183739	1.96	-1.03	-1.99	11596	7777	Splice	USP33
chr1:114516356	4.76	-1.47	-0.62	5094	11878	3' UTR	HIPK1
chr1:162750208	1.21	-1.95	-3.85	11737	9333	3' UTR	DDR2
chr3:9428211	1.44	-1.59	-3.77	6648	8767	3' UTR	THUMPD3
chr3:28476768	1.43	-1.35	-1.99	11596	4525	Splice	ZCWPW2
chr3:28503157	1.55	-1.35	-2.04	11596	4384	3' UTR	ZCWPW2
chr3:47316797	1.16	-1.99	-0.58	11313	12302	3' UTR	KIF9
chr3:47386060	1.05	-2.08	-0.66	12303	11029	3' UTR	KLHL18
chr3:52009091	1.41	-1.59	-2.48	11737	3535	5' UTR	ABHD14B
chr3:52109349	1.21	-1.87	-2.36	11879	11171	3' UTR	POC1A
chr4:103936040	1.17	-1.39	-3.37	8486	2828	5' UTR	SLC9B1
chr4:139983298	2.52	-2.28	-0.66	10182	11736	5' UTR	ELF2
chr4:73930626	1.06	-1.23	-3.45	10041	7212	Splice	COX18
chr5:86564477	1.14	-1.27	-1.99	10748	11171	NonSyn	RASA1
chr7:73113999	2.18	-1.47	-1.19	7638	9474	3' UTR	STX1A
chr9:127282609	1.23	-1.71	-1.91	10324	10888	3' UTR	NR6A1
chr10:102724515	1.17	-2.4	-3.81	11879	12160	3' UTR	FAM178A
chr10:15254162	1.01	-2.16	-3.77	9900	12302	3' UTR	FAM171A1
chr11:64900743	1.17	-1.47	-2.32	11455	9333	5' UTR	SYVN1
chr11:66406503	1.17	-1.39	-1.91	8345	4667	5' UTR	RBM4
chr11:66406696	1.17	-1.39	-1.91	8345	4667	5' UTR	RBM4
chr11:66407111	1.13	-1.43	-1.91	8203	4667	5' UTR	RBM4
chr11:66407983	1.15	-1.39	-1.91	8345	4667	3' UTR	RBM4
chr11:66453702	1.3	-1.27	-1.95	7073	8060	3' UTR	SPTBN2
chr11:129769974	1.64	-1.19	-1.47	12161	10322	3' UTR	PRDM10
chr11:129771185	1.44	-1.47	-2.08	12303	11312	3' UTR	PRDM10
chr11:129771376	1.37	-1.51	-2.08	12727	11312	3' UTR	PRDM10
chr11:129771773	1.28	-1.63	-1.39	12444	12019	3' UTR	PRDM10
chr11:129772293	1.16	-1.39	-1.23	12727	11453	NonSyn	PRDM10
chr13:41132149	1.06	-2.36	-3.13	12161	9191	3' UTR	FOXO1
chr13:52301811	1.48	-1.19	-1.63	6083	9757	Splice	WDFY2
chr16:66947064	1.14	-3.09	-1.55	10465	7212	NonSyn	CDH16
chr16:66968760	1.3	-2.48	-1.75	8062	7212	5' UTR	CES2
chr17:27955042	1.9	-3.85	-1.83	11737	8201	3' UTR	SSH2
chr17:27959258	2	-4.01	-2.04	11879	8060	NonSyn	SSH2
chr20:33337529	2.24	-3.69	-2.76	6648	11029	NonSyn	NCOA6
chr20:35412163	1.41	-0.94	-1.39	9193	6363	3' UTR	SOGA1
chr20:35412323	1.43	-1.11	-1.39	8203	6505	3' UTR	SOGA1
chr20:35413846	1.34	-0.94	-1.35	9193	7212	3' UTR	SOGA1
chr22:40723118	1.18	-1.79	-1.43	6931	4384	3' UTR	TNRC6B
chr22:40724058	2.34	-1.83	-1.99	9052	6646	3' UTR	TNRC6B
chr22:40760978	1.08	-1.95	-0.94	6790	6080	NonSyn	ADSL