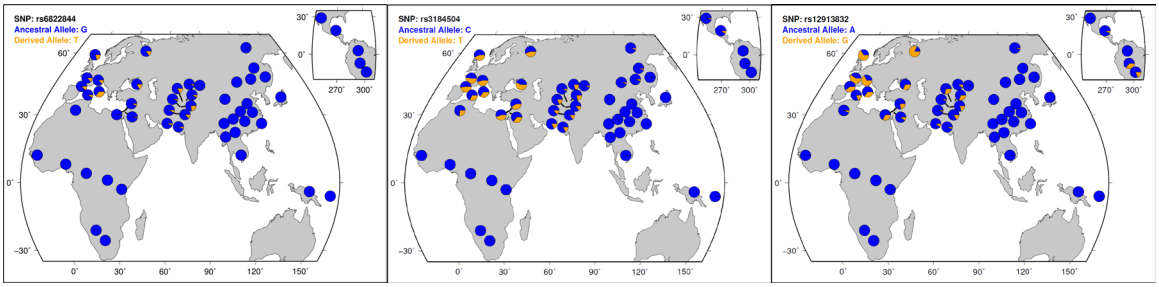
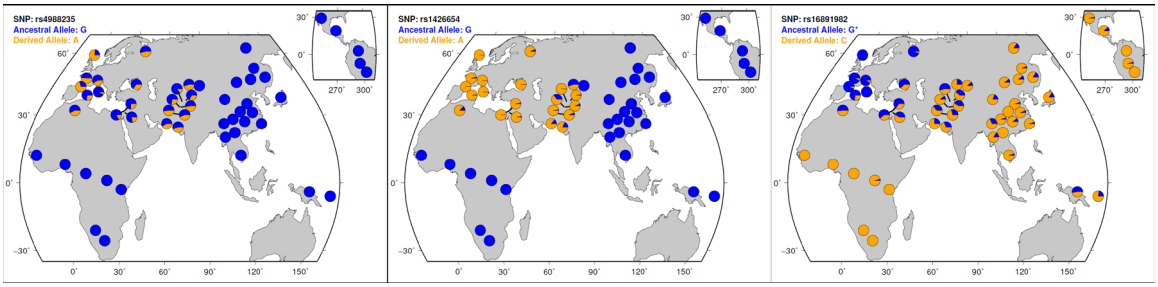
SUPPLEMENTARY FIGURE LEGENDS

Figure S1. Geographic distribution for eight SNPs in the Human Genome Diversity Project (HGDP) populations. The derived (and selected) allele is represented in orange (except for rs16891982 for which the derived and selected allele is represented in blue). (A) rs6822844, (B) rs3184504, (C) rs12913832, (D) rs4988235, (E) rs1426654, (F) rs16891982, (G) rs17810546, and (H) rs2188962. The maps were obtained from the HGDP Selection Bowser.

(A) (B) (C)



(D) (E) (F)



(G) (H)

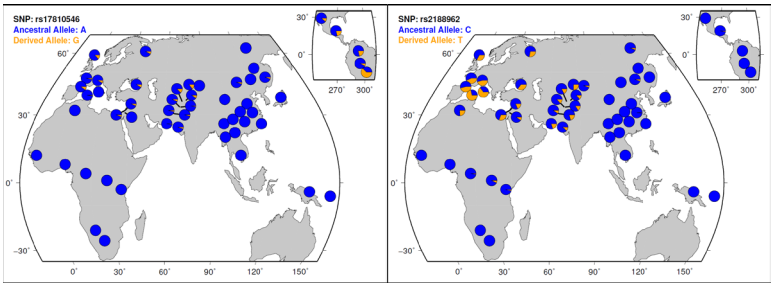


Figure S2. Genetic map of the eight studied regions. The genomic position and the recombination rate (cM/Mb) are represented on the x- and y-axis, respectively. The regions for which the capture array was designed is shown for SNPs rs6822844, rs3184504 and rs12913832 while a 1Mb region centered on the focal SNP is shown for the 5 SNPs analyzed using only the GC data. The blue area represents the subregions chosen for simulations and for allele age estimation. The red solid line marks the position of the SNPs associated with the selection signal.

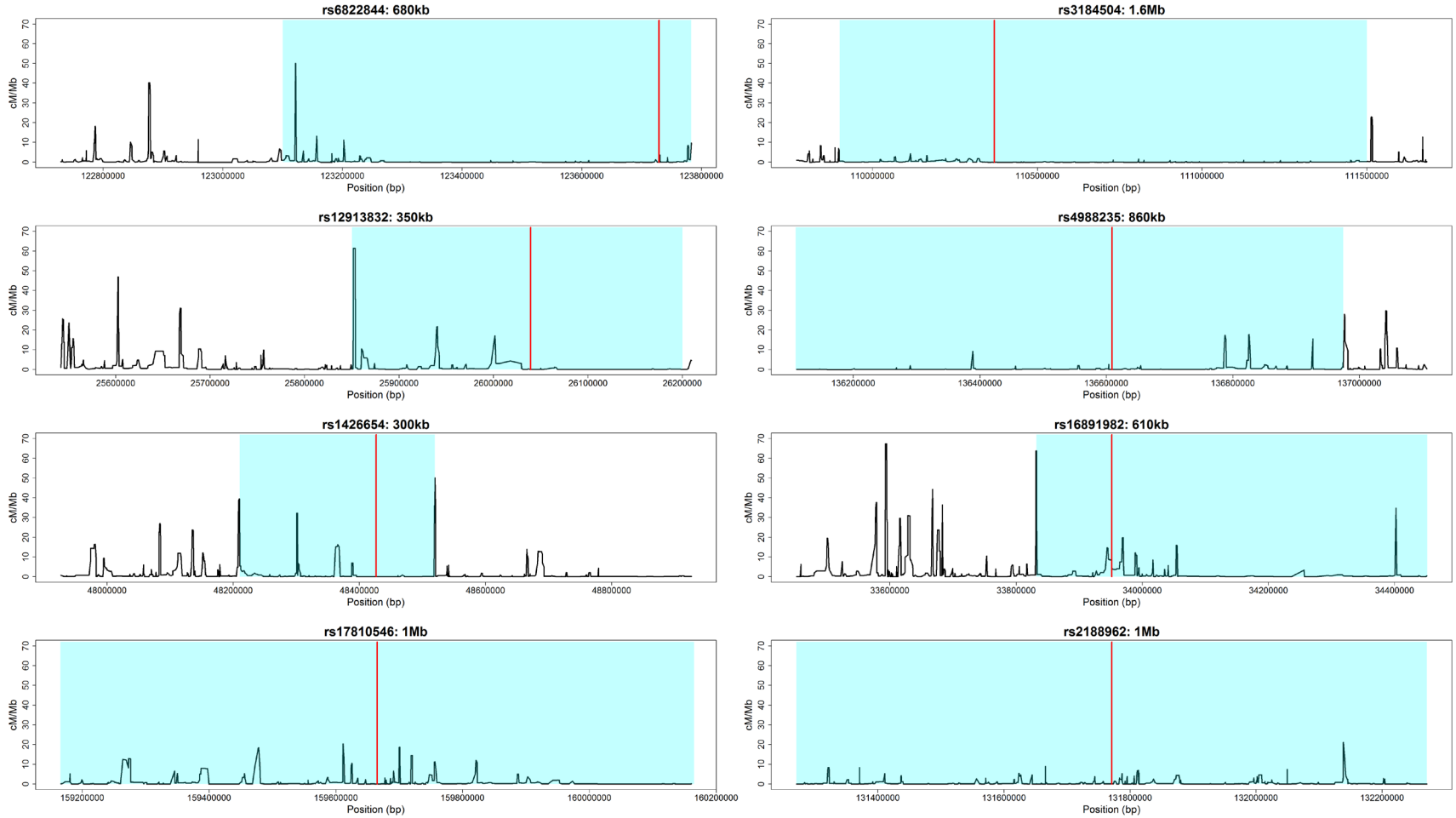


Figure S3. Coverage plots of 14 samples within the regions including rs6822844 (A), rs3184504 (B) and rs12913832 (C) targeted regions. Per region, two plots are shown. In both cases, the x-axis represents the positions within the targeted region while the y-axis represents the median read coverage and the number of genotypes called within the top and bottom plots, respectively. Lines of different colors represent the median coverage for each continuous sequenced region – separated by gaps – for each sample. In the bottom portion, the black and red lines represent all observed variants and variants that passed the filters, respectively.

(A) rs6822844 (B) rs3184504 (C) rs12913832

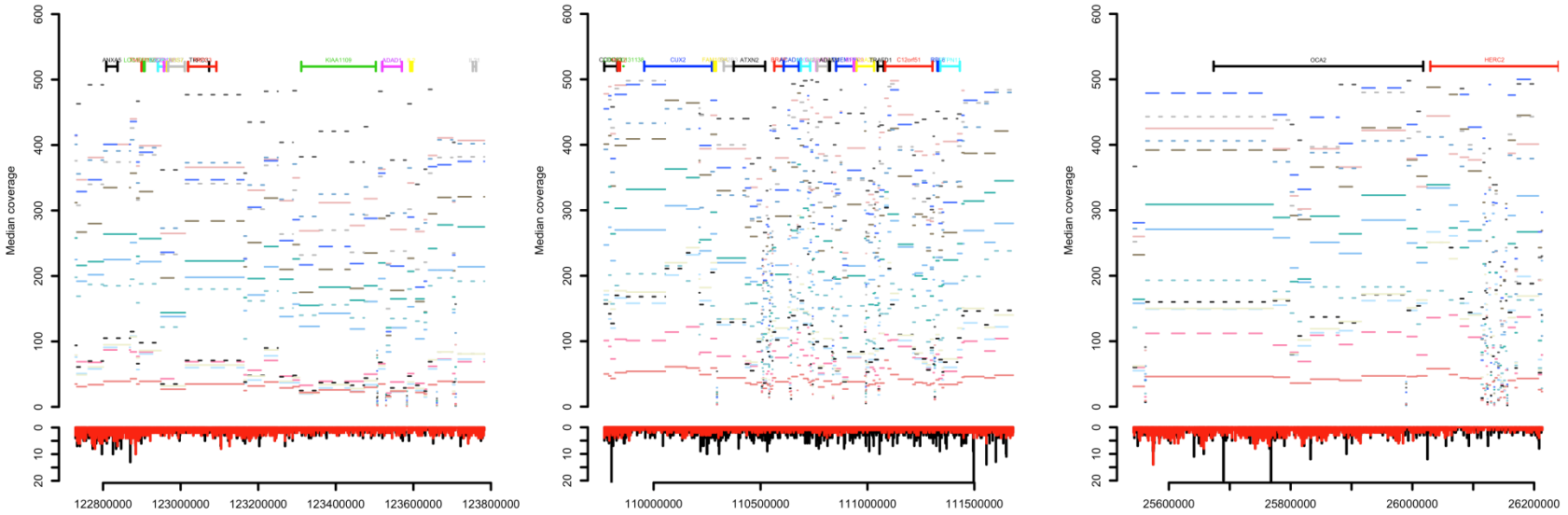


Figure S4. Histogram representing the allele imbalance at heterozygous positions in the regions spanning rs6822844 (A), rs3184504 (B) and rs12913832 (C). The reference genome was used to design the captured array oligos and to align the reads; as a result, reads carrying the reference allele are favored causing a slight allelic imbalance at heterozygous positions (~53% of the reads carried the reference allele at heterozygous positions). All observed heterozygous positions are represented in blue while the ones in orange represent the positions that were called heterozygous in the HapMap data.

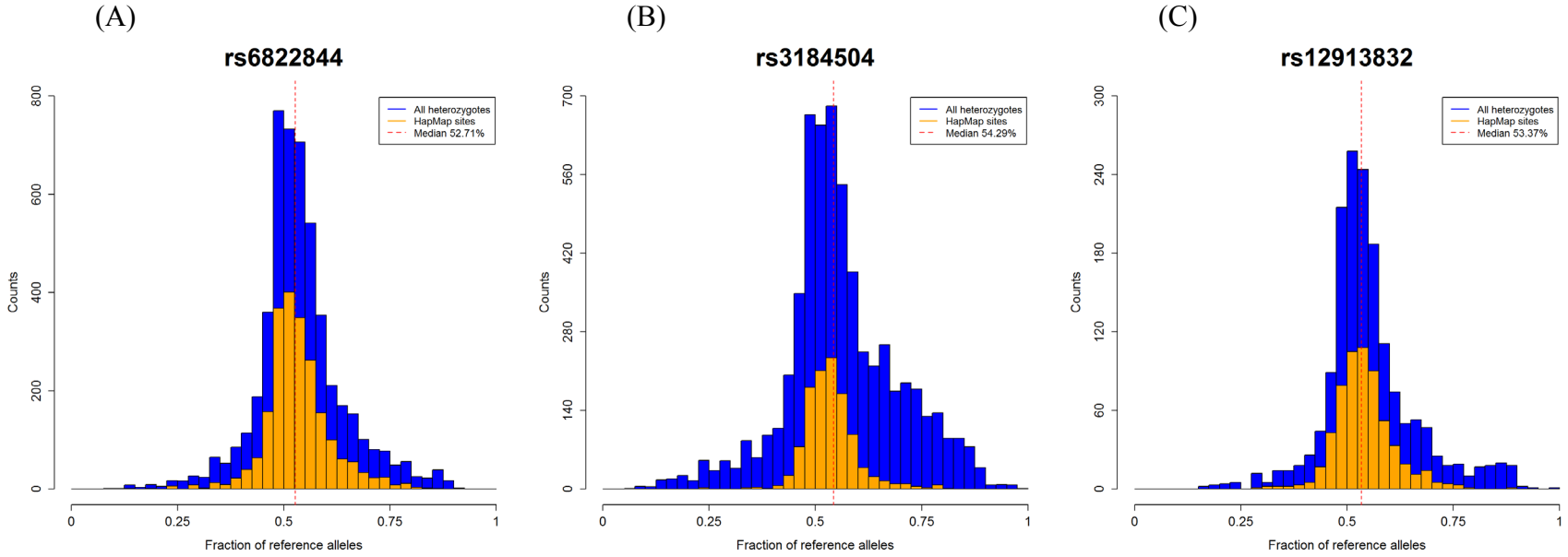
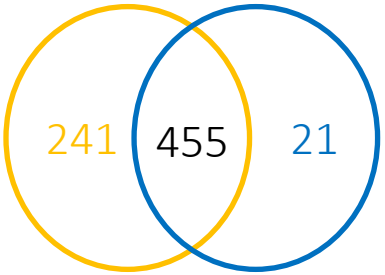


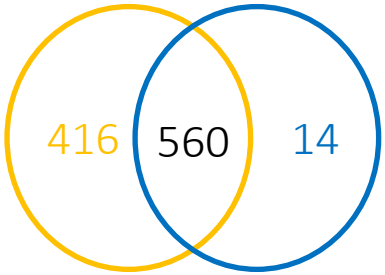
Figure S5. Venn diagrams of segregating sites among the six individuals typed in the CapSeq (orange) and the CG (blue) data sets. These counts were obtained from the targeted subregions spanning rs6822844 (A), rs3184504 (B) and rs12913832 (C). Numbers outside the circles show the total number of segregating sites in each data set.

(A) rs6822844



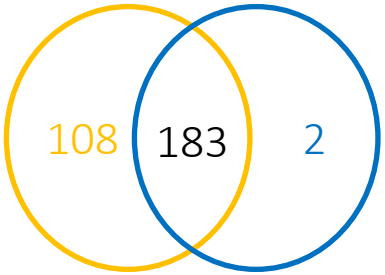
696 476

(B) rs3184504



976 574

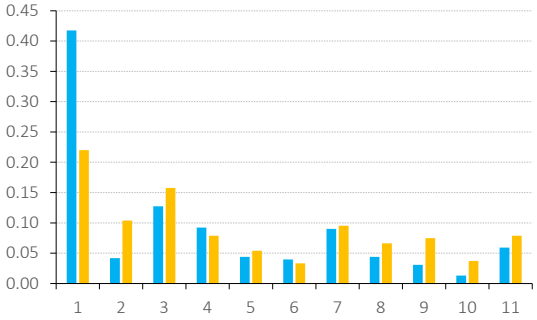
(C) rs12913832



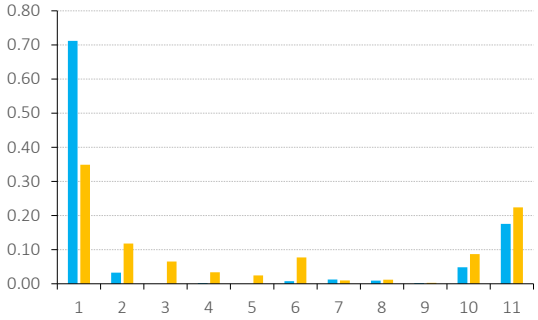
291 185

Figure S6. Site-frequency spectra (SFS) for the six individuals typed in the CapSeq and CG data sets. Blue bars represent the SFS of variants detected by both CapSeq and CG data while orange bars those detected by CapSeq but not CG data (Figure S5) for rs6822504 (A), rs3184504 (B) and rs12913832 (C). The *x*-axis shows the number of copies of alleles observed at a site and the *y*-axis the fraction of sites in each allele frequency class.

(A) rs6822844



(B) rs3184504



(C) rs12913832

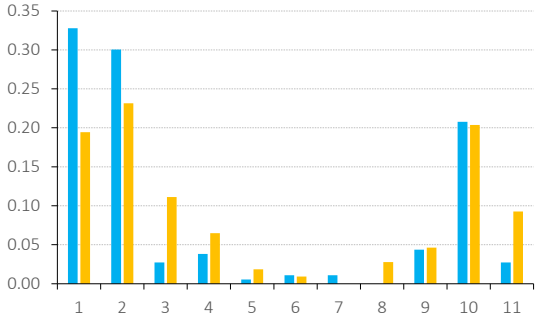


Figure S7. Boxplot of $\log_2$ ratio of the estimated to the true age of the allele to assess the sensitivity of the age estimation to demographic history, sample size, unsurveyed subregions and phasing uncertainty. The performance of the ABC method of simulated selection events at four time points (t) 2000 (D, H, L), 1200 (C, G, K), 800 (B, F, J) and 400 (A, E, I) generations ago representing pre-bottleneck, bottleneck, recovery and expansion periods are shown for the three SNPs for which we collected ultra-high depth sequence data. Blue boxplots refer to the actual data while orange boxplots refer to data with sample size n = 128, green boxplots refer to data without gaps and red boxplots refer to data without phasing uncertainty (see Text S4 for more details).

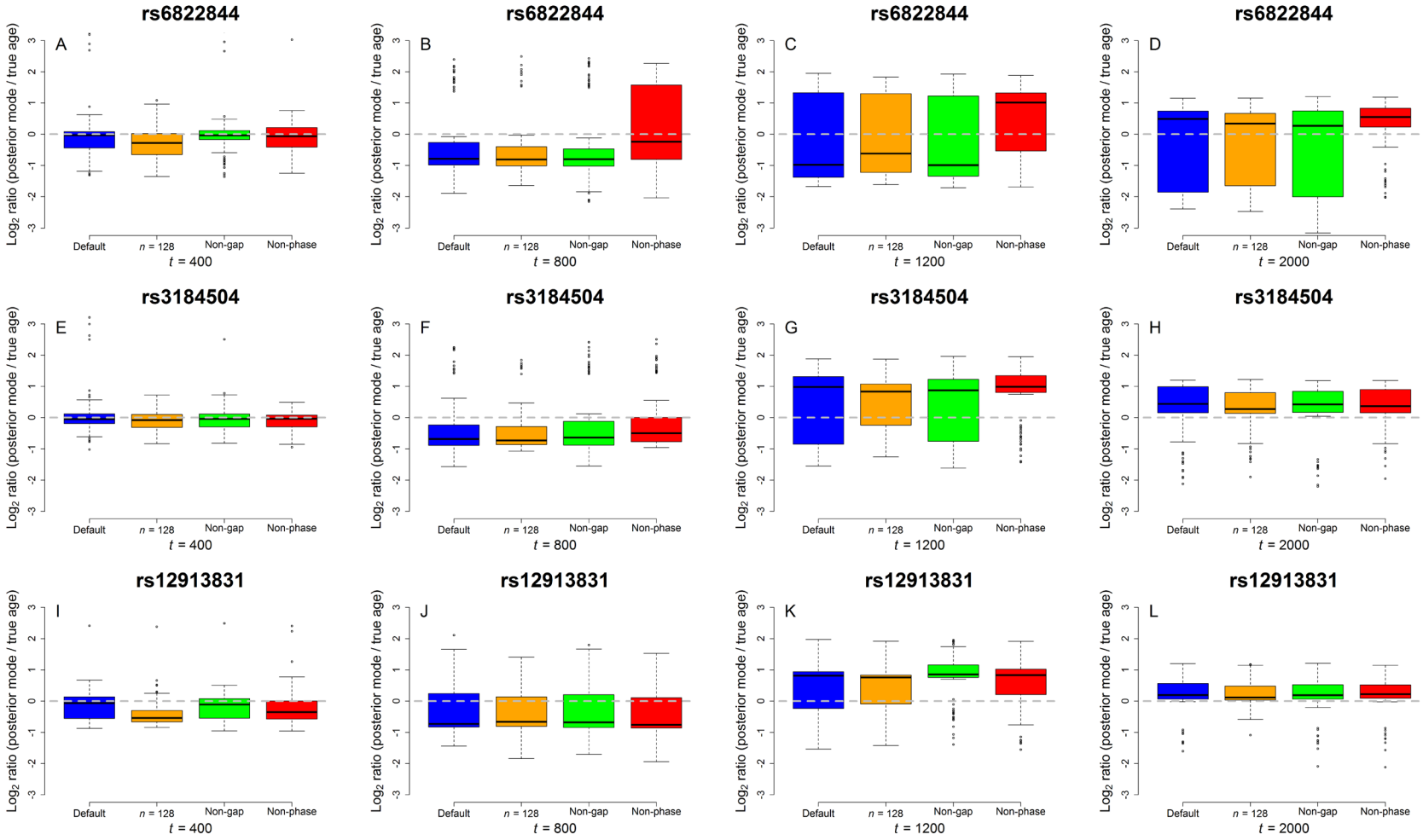


Figure S8. Posterior density distributions of the selection coefficients and the age of the selected alleles with the joint density plot for these two parameters for the CapSeq (orange) and the CG (blue) data sets. From left to right, the plots show the results for rs6822844, rs3184504 and rs12913832, respectively.

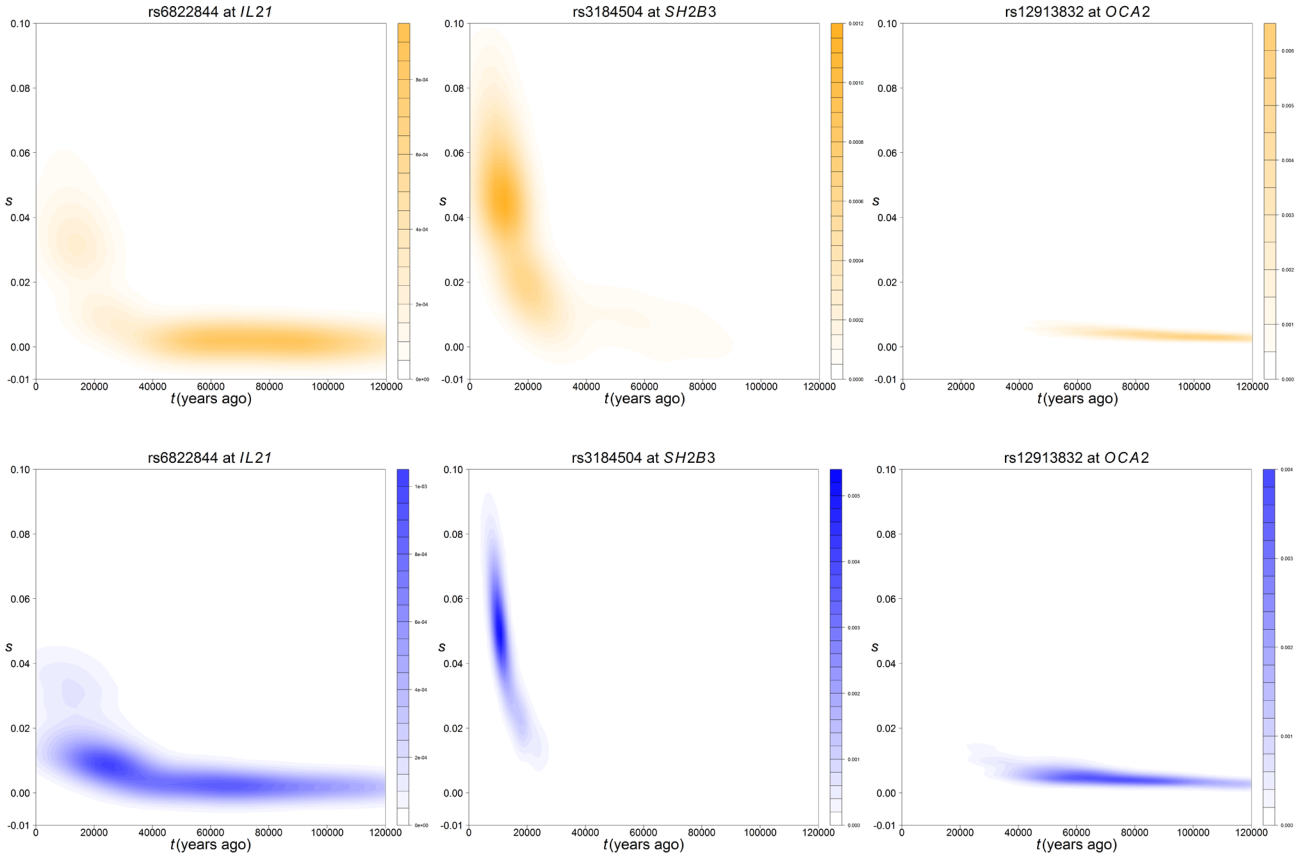


Figure S9. Posterior predictive checks for the ABC method for the 3 CapSeq regions. This figure compares the observed and simulated summary statistics (*SSs*) for the targeted subregions for the CapSeq (orange) and the CG (blue) data sets. The distribution of the *SSs* estimated in the simulated data is shown for (i) the inverse of the genetic distance in the selected region ($1/L_H$), (ii) the average number of mutations accumulated in the haplotypes carrying the selected alleles divided by the physical distance in the selected region (M_H) and (iii) the number of singleton variants divided by the total number of segregating sites in these haplotypes (R_H), respectively for the rs6822844 (upper), rs3184504 (middle) and rs12913832 (lower). The horizontal red line indicates the value of the corresponding *SS* observed in the CapSeq data, and the horizontal dashed black lines represent the 95% confidence interval (CI).

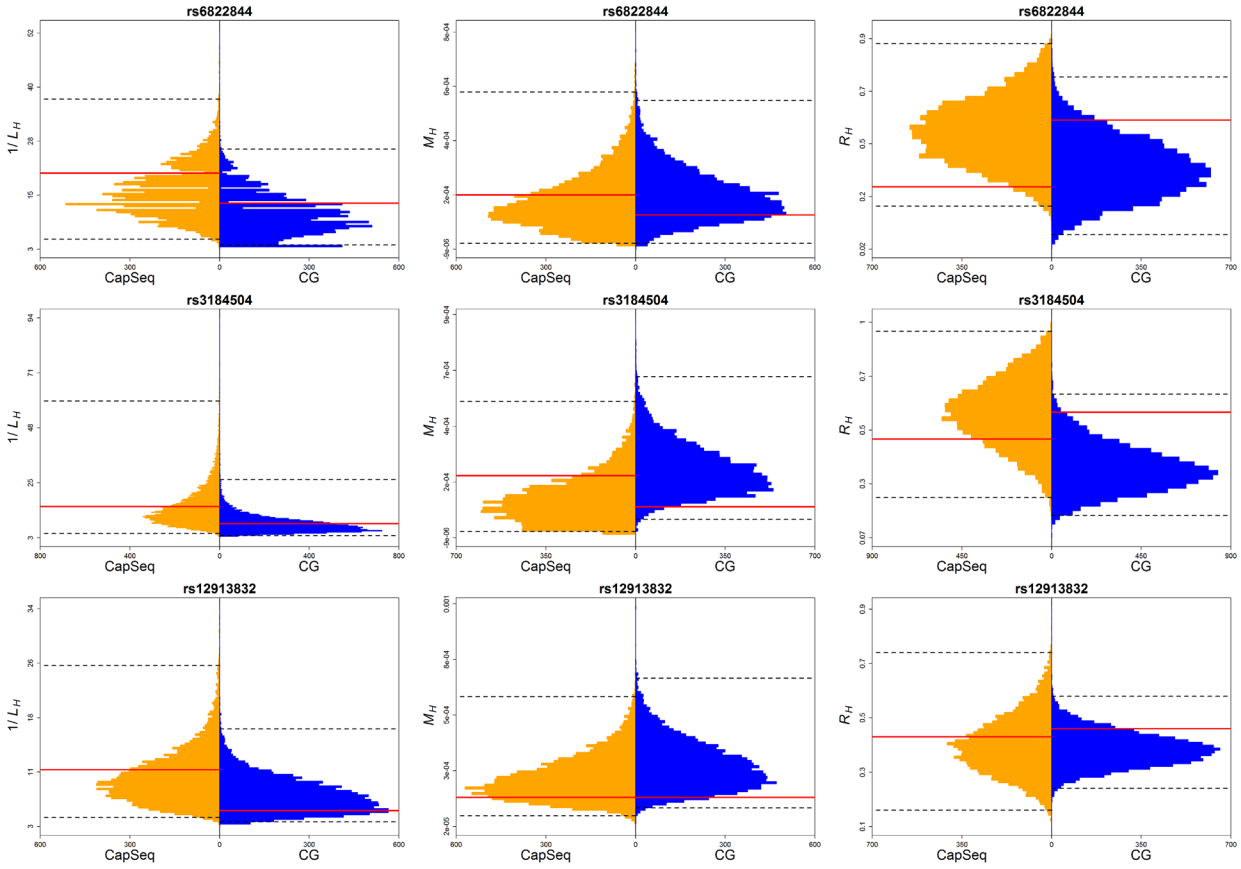


Figure S10. Posterior predictive checks for the ABC method. This figure compares the observed and simulated summary statistics (*SS*s) for the five SNPs analyzed using only the CG data. Each row represents the distributions of three *SS*s ($1/L_H$, M_H, and R_H) from one SNP. The red line indicates the value of the corresponding observed *SS*, and the dashed black lines represent the 95% confidence interval (CI).

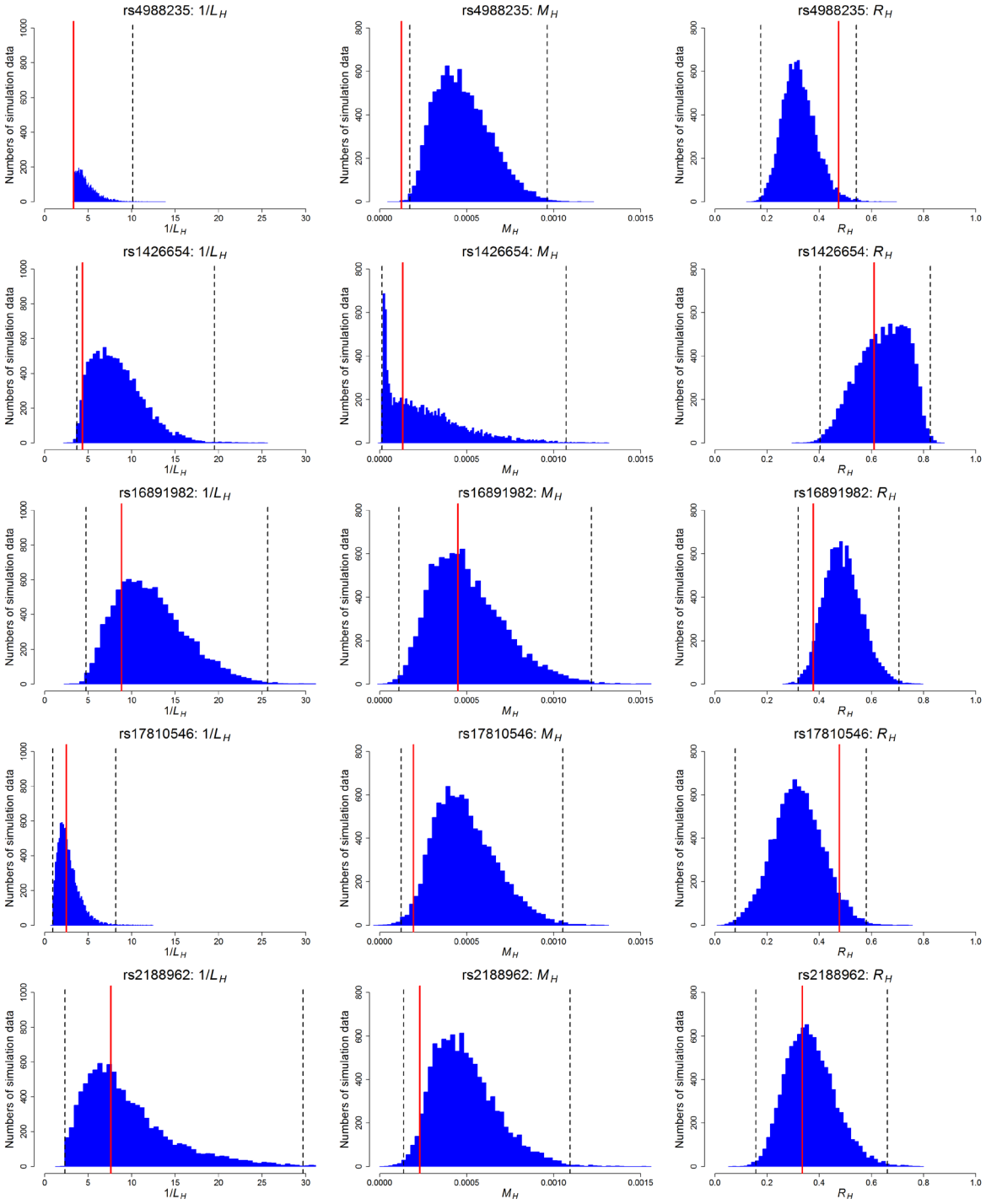


Figure S11. Comparison of the posterior probability distributions for *t* of rs4988235, rs17810546 and rs3181504 obtained using a prior distribution with density proportional *to N vs.* a uniform prior between 10 and 5000 generations ago.

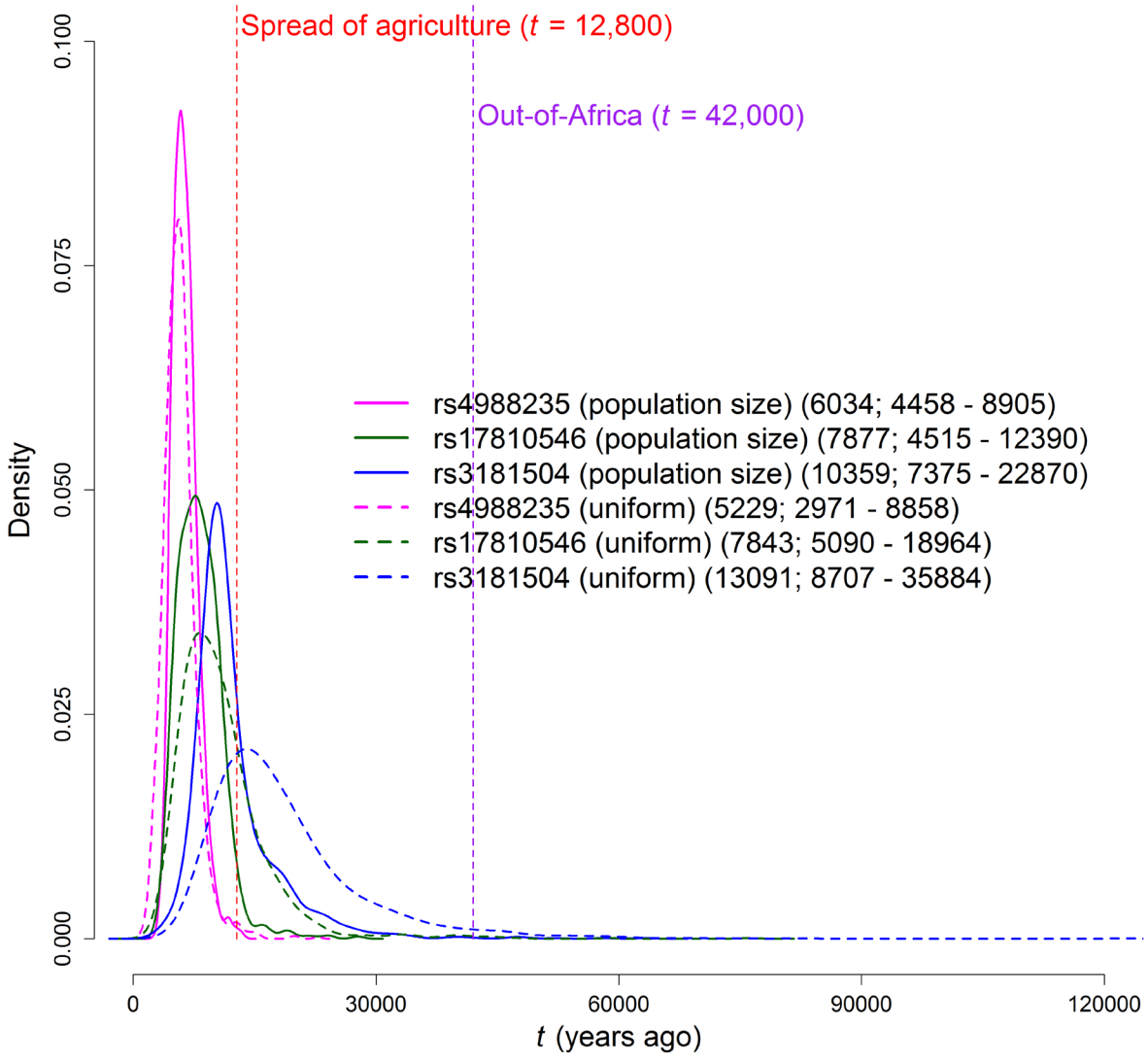
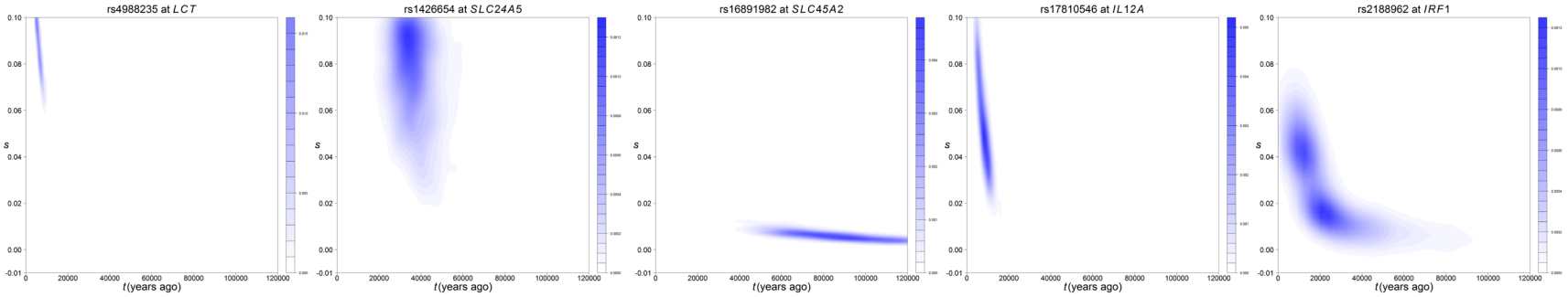


Figure S12. Joint posterior distributions of *t* (x-axis) and *s* (y-axis) for the five SNPs analyzed using only the CG data (rs4988235, rs1426654, rs16891982, rs17810546 and rs2188962).



SUPPLEMENTARY TABLES

Table S1. Summary information of the sequenced CEU samples. The Coriell ID and the genotypes of the corresponding targeted SNPs are indicated. The targeted allele is T for rs6822844 and rs3184504 and G for rs12913832, respectively.

Coriell_ID	rs6822844	rs3184504	rs12913832
NA10839	TT	CT	GG
NA10864	GT	TT	AG
NA11881	GT	TT	GG
NA12144	GT	TT	AG
NA12248	TT	TT	AG
NA12748	TT	CT	GG
NA12762	GT	TT	GG
NA07031	GT	CT	GG
NA10855	GT	CT	GG
NA11830	GT	CT	GG
NA11839	GT	CT	GG
NA12489	GT	CT	GG
NA12874	GT	CT	GG
NA12891	GT	CT	GG

Table S2. Summary information of the array captured targeted regions spanning rs6822844, rs3184504 and rs12913832. The start and end positions of the targeted regions are given with reference to build 36.

Parameters	rs6822844	rs3184504	rs12913832
Start position	122729299	109769404	25542017
End position	123782528	111681303	26213429
Size of the targeted region (bp)	1053229	1911899	671412
Size of the region covered by oligos (bp)	836589	1283878	503616
Size of the area not covered by oligos (i.e., gap) (bp)	216715	628086	167861
gap size/targeted region size	0.21	0.33	0.25
Number of small gaps (i.e., <400bp)	1274	2598	908
Average size of continuous segment targeted by oligos (bp)	657	494	555
Average gap size (bp)	170	242	185
%A+T	0.64	0.56	0.57

Table S3. Statistics of sequence reads per sample. For each sample the total number of reads is reported as well as the percentage of the reads left when duplicates were removed (non-redundant reads). In addition, the observed coverage per base averaged across individuals is given on the last row.

			% of non-redundant reads aligned to targeted region			Average coverage per sample		
Coriell_ID	No. of reads	% non-redundant reads	rs6822844	rs3184504	rs12913832	rs6822844	rs3184504	rs12913832
NA10864	66,612,726	34	2.2	5.6	2.2	31.2	44.8	44.1
NA12144	67,897,518	53	3.3	9.3	4.4	83.3	127.5	162.7
NA11830	70,014,962	46	3.5	13.5	6.0	80.8	168.2	209.9
NA11839	71,423,260	42	4.3	15.2	6.5	91.5	174.9	210.3
NA12489	67,098,136	32	5.1	19.9	8.0	78.3	168.2	190.8
NA10839	64,011,962	85	5.2	10.8	4.0	192.6	216.5	218.1
NA10855	73,994,098	62	7.0	16.7	6.3	227.1	293.4	310.6
NA12748	65,299,828	58	7.5	19.5	7.9	198.0	281.0	312.4
NA11881	71,681,848	71	9.0	22.0	8.0	321.1	425.4	426.9
NA07031	71,389,744	76	9.3	19.5	7.2	355.9	402.2	404.7
NA12891	72,597,604	49	10.7	27.0	10.3	269.8	368.8	392.2
NA12874	73,636,302	51	12.3	30.3	11.4	326.4	432.6	458.5
NA12248	71,136,622	58	12.5	31.5	10.0	362.7	482.2	431.4
NA12762	73,460,858	56	15.5	33.0	12.5	448.7	507.8	527.1
	70,018,248	55	7.7	19.6	7.5	219.1	292.39	307.12

Table S4. Comparison of genotype calls between the CapSeq data and the HapMap data. The genotype calls for all overlapping SNPs between the CapSeq data and the HapMap data are shown. RR denotes homozygote for reference allele; RA: heterozygote; and AA homozygote for the alternative allele.

	HapMap RR			HapMap RA			HapMap AA		
Sample	RR	RA	AA	RR	RA	AA	RR	RA	AA
NA10839	1659	0	2	2	117	1	12	0	263
NA07031	403	0	0	0	210	0	0	1	96
NA10855	1457	5	1	0	376	0	9	4	219
NA11830	1621	2	1	0	297	1	11	3	141
NA11839	1501	3	0	1	354	0	7	2	181
NA12489	479	0	0	0	140	0	0	0	96
NA12874	1640	1	1	1	206	0	10	3	208
NA12891	1472	1	1	2	378	0	8	5	204
NA10864	532	0	0	0	83	0	0	0	96
NA11881	1619	3	1	0	193	0	11	0	226
NA12144	1521	2	1	1	376	0	12	1	146
NA12248	1607	4	1	0	173	0	8	4	275
NA12748	534	0	0	0	48	0	0	0	133
NA12762	1648	1	0	0	239	0	11	0	159

Table S5. Detailed information on the mismatches observed when the CapSeq genotype calls were compared to HapMap genotypes. Among the 162 mismatches observed, genotype calls for 68 mismatches shown in the table were available in the Phase 3 1KG data. R denotes reference allele and A alternative allele.

					Numbers of reads		Genotype calls			
SAMPLE	SNP	R	A	Total coverage	R	A	CapSeq	HapMap	1KG	ERROR
NA11830	rs10018569	G	C	71	0	71	CC	GG	CC	HapMap
NA11830	rs1880865	G	A	394	204	190	GA	GG	GA	HapMap
NA11830	rs955710	G	T	20	11	9	GT	GG	GT	HapMap
NA12874	rs10018569	G	C	219	0	219	CC	GG	CC	HapMap
NA12874	rs17005630	G	A	300	173	127	GA	GG	GA	HapMap
NA11881	rs10018569	G	C	163	0	163	CC	GG	CC	HapMap
NA11881	rs1880865	G	A	878	414	464	GA	GG	GA	HapMap
NA12144	rs10018569	G	C	62	0	62	CC	GG	CC	HapMap
NA12144	rs17005630	G	A	23	12	11	GA	GG	GA	HapMap
NA12762	rs10018569	G	C	232	105	127	GC	GG	GC	HapMap
NA11830	rs11065857	A	G	273	273	0	AA	GG	AA	HapMap
NA11830	rs7978923	G	T	129	129	0	GG	TT	GG	HapMap
NA11830	rs4572196	A	G	385	385	0	AA	GG	AA	HapMap
NA11830	rs10774624	G	A	35	23	12	GA	AA	GA	HapMap
NA11830	rs11065961	G	A	95	68	27	GA	AA	GA	HapMap
NA11830	rs668774	G	C	122	121	1	GG	CC	GG	HapMap
NA11830	rs7953257	A	T	24	10	14	AT	TT	AT	HapMap
NA11830	rs6489845	G	C	135	1	134	CC	CG	CC	HapMap
NA11830	rs7136443	A	C	80	80	0	AA	CC	AA	HapMap
NA11830	rs7136494	T	A	179	179	0	TT	AA	TT	HapMap
NA11830	rs11065844	A	C	202	202	0	AA	CC	AA	HapMap
NA11830	rs10160956	G	A	132	132	0	GG	AA	GG	HapMap
NA11830	rs7976102	C	G	42	42	0	CC	GG	CC	HapMap

NA12874	rs12099707	C	G	457	457	0	CC	CG	CC	HapMap
					Numbers of reads		Genotype calls			
SAMPLE	SNP	R	A	Total coverage	R	A	CapSeq	HapMap	1KG	ERROR
NA12874	rs11065857	A	G	558	558	0	AA	GG	AA	HapMap
NA12874	rs7978923	G	T	301	157	144	GT	TT	GT	HapMap
NA12874	rs4572196	A	G	588	588	0	AA	GG	AA	HapMap
NA12874	rs10774624	G	A	210	111	99	GA	AA	GA	HapMap
NA12874	rs668774	G	C	287	287	0	GG	CC	GG	HapMap
NA12874	rs7953257	A	T	116	53	63	AT	TT	AT	HapMap
NA12874	rs7136443	A	C	396	396	0	AA	CC	AA	HapMap
NA12874	rs7136494	T	A	355	355	0	TT	AA	TT	HapMap
NA12874	rs11065844	A	C	397	397	0	AA	CC	AA	HapMap
NA12874	rs10160956	G	A	89	89	0	GG	AA	GG	HapMap
NA12874	rs7976102	C	G	160	160	0	CC	GG	CC	HapMap
NA11881	rs4766551	C	T	774	410	364	CT	CC	CT	HapMap
NA11881	rs4766452	T	C	56	28	28	TC	TT	TC	HapMap
NA11881	rs11065857	A	G	569	569	0	AA	GG	AA	HapMap
NA11881	rs7978923	G	T	199	198	1	GG	TT	GG	HapMap
NA11881	rs10774624	G	A	148	148	0	GG	AA	GG	HapMap
NA11881	rs668774	G	C	189	189	0	GG	CC	GG	HapMap
NA11881	rs7953257	A	T	52	52	0	AA	TT	AA	HapMap
NA11881	rs7136443	A	C	148	148	0	AA	CC	AA	HapMap
NA11881	rs11065844	A	C	323	323	0	AA	CC	AA	HapMap
NA11881	rs10160956	G	A	102	102	0	GG	AA	GG	HapMap
NA11881	rs7976102	C	G	79	79	0	CC	GG	CC	HapMap
NA12144	rs11065857	A	G	211	211	0	AA	GG	AA	HapMap
NA12144	rs7978923	G	T	99	97	2	GG	TT	GG	HapMap
NA12144	rs4572196	A	G	254	254	0	AA	GG	AA	HapMap
NA12144	rs10774624	G	A	29	29	0	GG	AA	GG	HapMap

NA12144	rs12371484	T	C	143	71	72	TC	TT	TC	HapMap
					Numbers of reads		Genotype calls			
SAMPLE	SNP	R	A	Total coverage	R	A	CapSeq	HapMap	1KG	ERROR
NA12144	rs668774	G	C	38	37	1	GG	CC	GG	HapMap
NA12144	rs7953257	A	T	19	10	9	AT	TT	AT	HapMap
NA12144	rs10850052	C	A	99	99	0	CC	AC	CC	HapMap
NA12144	rs7136443	A	C	39	39	0	AA	CC	AA	HapMap
NA12144	rs7136494	T	A	95	95	0	TT	AA	TT	HapMap
NA12144	rs11065844	A	C	59	59	0	AA	CC	AA	HapMap
NA12144	rs10160956	G	A	32	32	0	GG	AA	GG	HapMap
NA12144	rs7976102	C	G	65	65	0	CC	GG	CC	HapMap
NA12762	rs7978923	G	T	274	273	1	GG	TT	GG	HapMap
NA12762	rs4572196	A	G	944	944	0	AA	GG	AA	HapMap
NA12762	rs10774624	G	A	253	253	0	GG	AA	GG	HapMap
NA12762	rs668774	G	C	352	351	1	GG	CC	GG	HapMap
NA12762	rs7953257	A	T	144	144	0	AA	TT	AA	HapMap
NA12762	rs7136494	T	A	392	392	0	TT	AA	TT	HapMap
NA12762	rs11065844	A	C	222	222	0	AA	CC	AA	HapMap
NA12762	rs10160956	G	A	118	118	0	GG	AA	GG	HapMap
NA12762	rs7976102	C	G	204	204	0	CC	GG	CC	HapMap

Table S6. Summary information of the parameters used on the simulations. Additive model (h = 0.5) was used in all eight regions in addition to the same effective population size (N = 120,000). The derived allele of each targeted SNP showed evidence of selection; however, frequency differences were observed. Therefore, different derived allele counts were simulated per region. Finally, the size of the locus with uniform recombination rate as well as the relative location of the targeted SNPs were also different within the observed data, so these differences were also accounted in the simulations.

Parameters	rs6822844	rs3184504	rs12913832	rs4988235	rs1426654	rs16891982	rs17810546	rs2188962
μ: mean mutation rate (per bp per generation) (see Methods and Text S2)	1.2E-08	1.2E-08	1.1E-08	1.5E-08	1.9E-8	1.7E-8	2.0E-8	1.6E-8
r: recombination rate (cM/bp)	4.0E-07	2.0E-07	8.0E-07	3.5E-7	11.6E-7	10.4E-7	10.8E-7	4.4E-7
Derived alleles in the HapMap or CG sample	33	119	179	97	128	124	15	52
HapMap or CG sample size	226	220	226	128	128	128	128	128
Number of chromosomes in CapSeq and CG data	28 and 128	28 and 128	28 and 128	128	128	128	128	128
Number of derived alleles in our sample and CG data	17 and 17	19 and 53	25 and 99	97	128	124	15	52
Locus size of the subregion (kb) (see Figure S2)	680	1,600	350	860	300	610	1,000	1,000
Relative position of selected SNP	90%	30%	55%	58%	70%	19%	50%	50%

Table S7. Correlation among summary statistics (*SS*s) used for age (t) estimation. Simulation conditions are different among the 3 CapSeq regions studied (see Table S6), so 3 tables summarize the correlation among the three *SS*s used for allele age estimation (corrected for the age in order to exclude spurious correlations) as well as the correlation between each of them and the log(t) for rs6822844, rs3184504 and rs12913832, respectively.

rs6822844

Spearman's rho	$1/L_H$	M_H	R_H
log(t)	0.56	0.3	-0.27
L_H		0.53	0.16
M_H			-0.29

rs3184504

Spearman's rho	$1/L_H$	M_H	R_H
log(t)	0.65	0.13	-0.1
L_H		0.39	0.31
M_H			-0.28

rs12913832

Spearman's rho	$1/L_H$	M_H	R_H
log(t)	0.78	0.08	0.02
L_H		0.38	0.18
M_H			-0.02