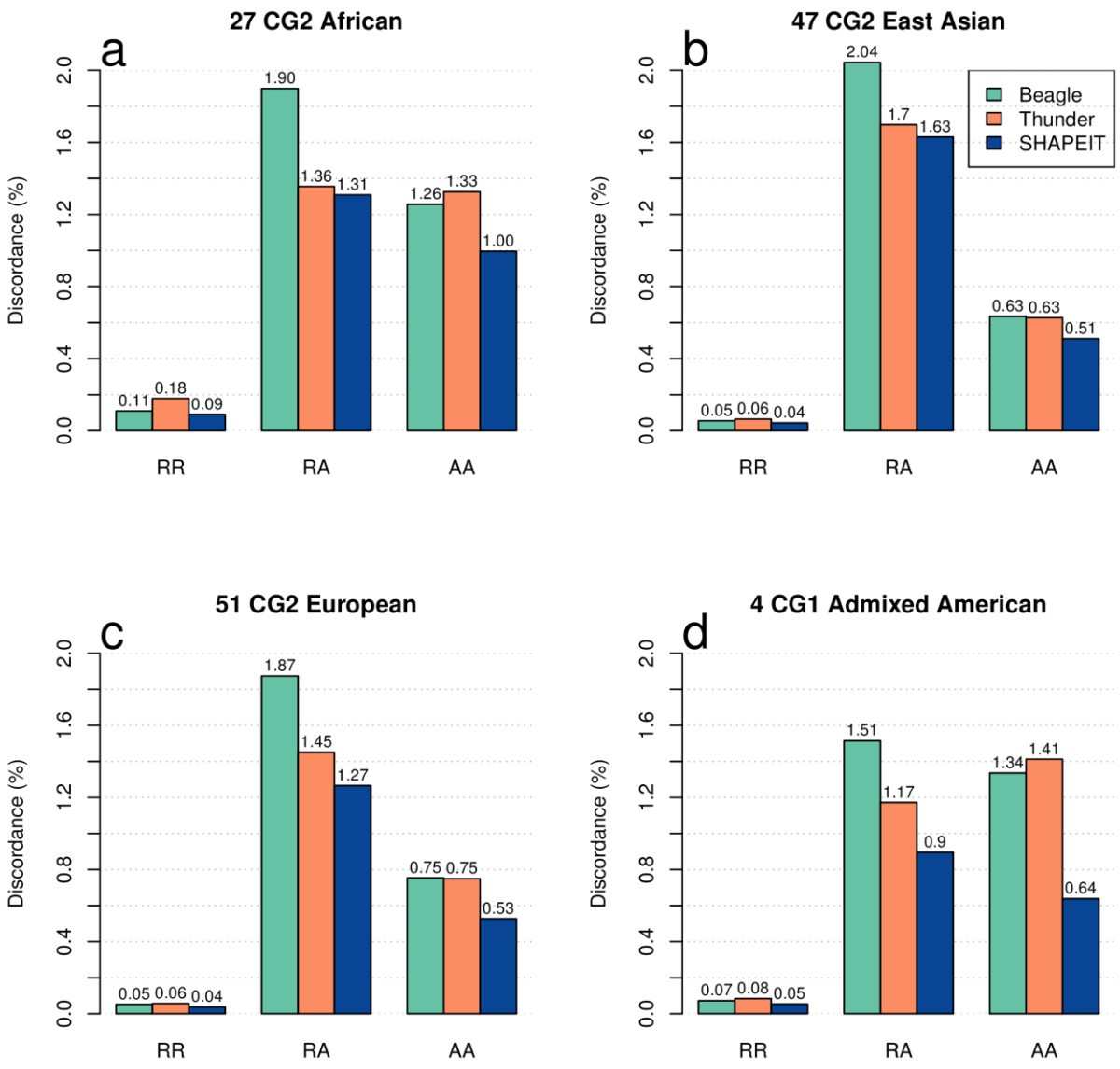
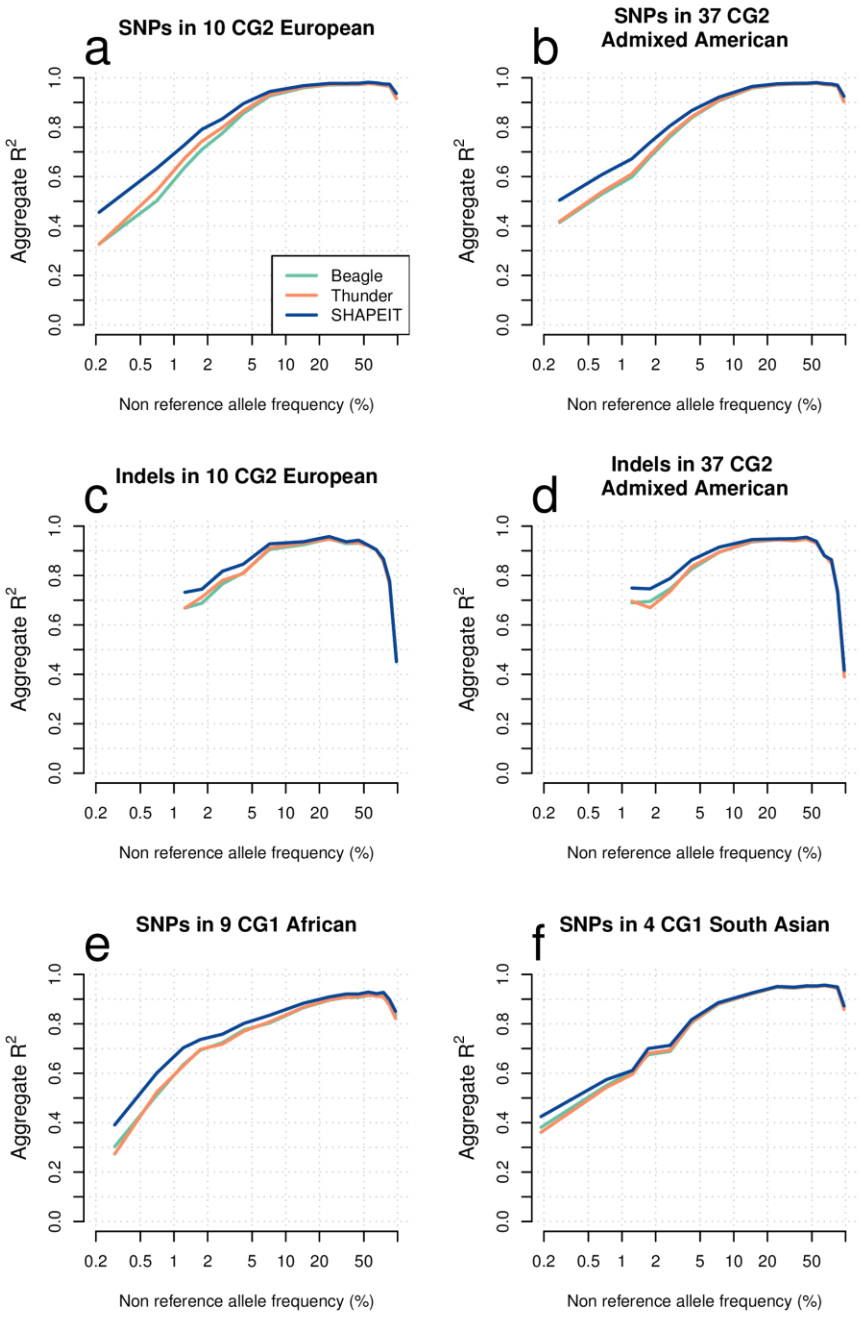
Supplementary Figure 1



Genotype discordance by continental group. Genotype discordance of the Beagle, Thunder and SHAPEIT2 call sets at SNP genotypes in (a) 27 CG2 African samples, (b) 47 CG2 East Asian samples, (c) 51 CG2 European samples and (d) 4 CG1 Admixed American samples. RR, RA and AA stand for Reference/Reference, Reference/Alternative and Alternative/Alternative genotypes, respectively.

Supplementary Figure 2



Imputation performance by continental group. Imputation performance of the Beagle, Thunder and SHAPEIT2 call sets to impute SNPs and indels in various continental groups genotyped on Illumina 1M: (a) SNPs in 10 CG2 European samples, (b) SNPs in 37 CG2 Admixed American samples, (c) Indels in 10 CG2 European samples, (d) Indels in 37 CG2 Admixed American samples, (e) SNPs in 9 CG1 African samples and (f) SNPs in 4 CG1 South Asian samples. Imputation accuracy is measured as the aggregate squared correlation coefficient and plotted against the non-reference allele frequency.

Supplementary Figure 3

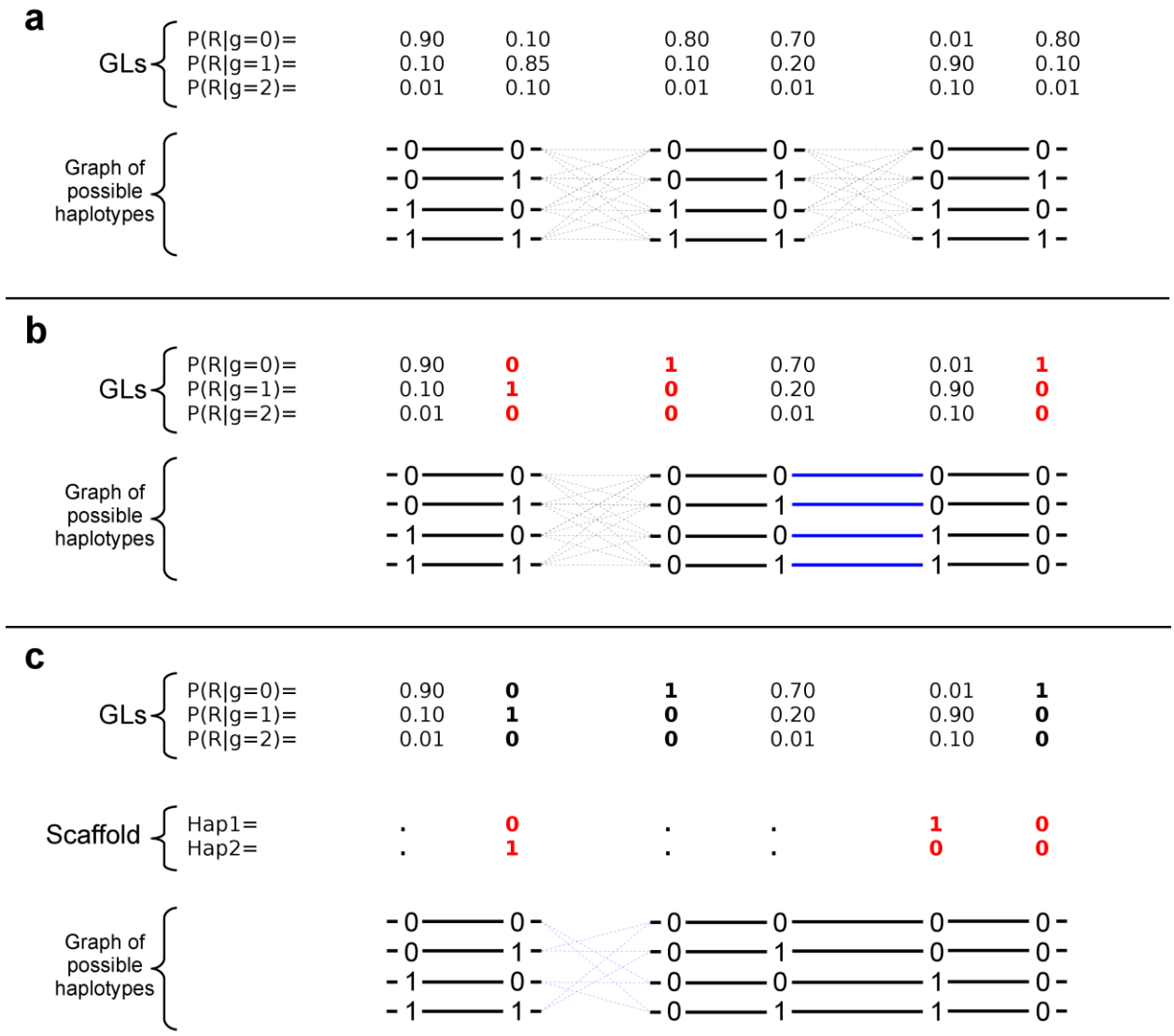
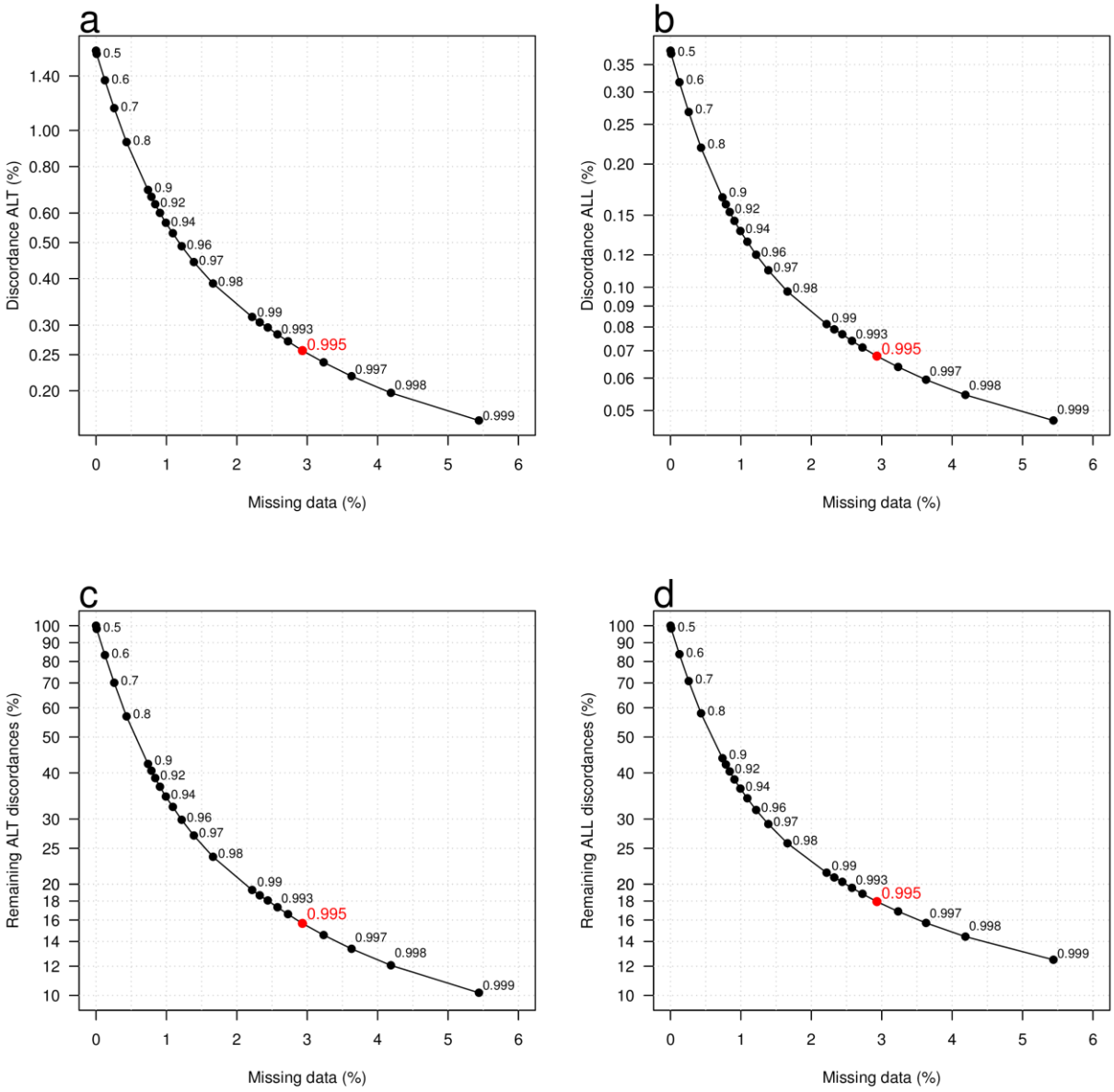


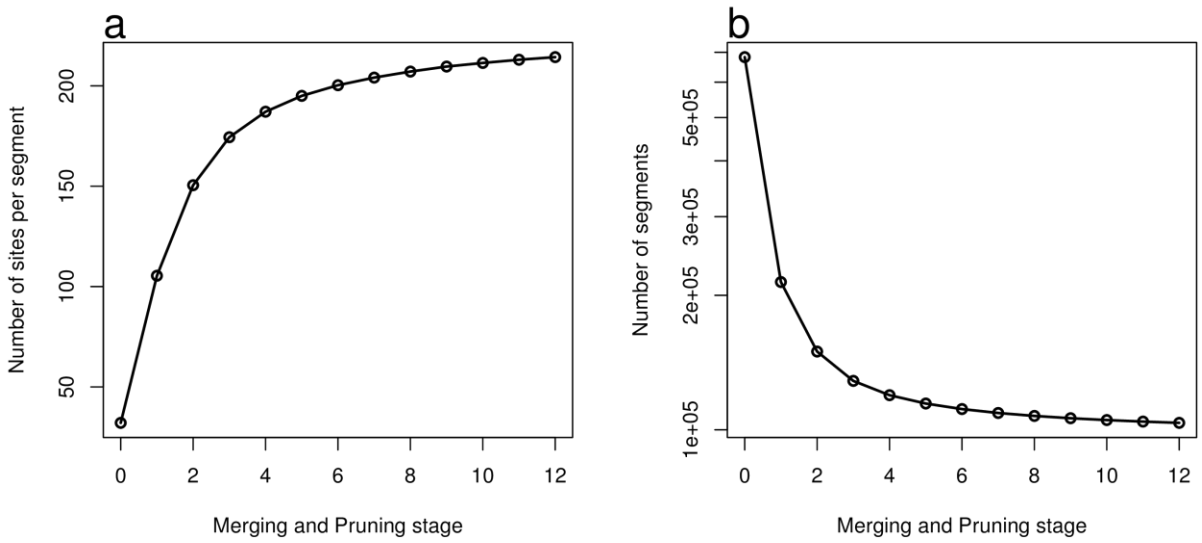
Illustration of the 3 levels of our model. Panel (a) shows an example of possible haplotype graph that can be derived from a vector of genotype likelihoods. The segments contains few SNPs in order to limit the number of possible haplotypes they contain. Panel (b) illustrates what happens to the situation in A when some of the genotypes are initialised using the Beagle posteriors (in red): the segments contain more sites (in blue). Panel (c) illustrates what happens to the situation in B when a haplotype scaffold is used (in red): some of the transitions inconsistent with the scaffold are discarded (leaving just those in blue).

Supplementary Figure 4



Calibration of Beagle posterior probabilities. Plots (a) and (b) give for several threshold values the ALT and ALL discordance rates as a function of the proportion of genotypes discarded by a given threshold. For example, the points highlighted in red show that using a threshold of 0.995 on the Beagle posteriors, we discard 3% of the genotypes and the remaining 97% have ALT and ALL discordance rates of ~0.25% and ~0.07% respectively. Panels (c) and (d) give similar information but this time the proportion of genotypes discarded by a given threshold is plotted against the percentage of the overall discordance accounted for by the genotypes passing the filter. For example, the two red points show that using a threshold of 0.995 on the Beagle posteriors, we discard 3% of the genotypes and the remaining 97% contain only ~16% and ~18% of the ALT and ALL discordances contained in the full call set.

Supplementary Figure 5



Convergence of the SHAPEIT2 MCMC algorithm. Plot (a) shows the average number of sites per segment (y-axis) as the number of merging and pruning stages increases (x-axis). Plot (b) shows the total number of segments as the number of pruning and merging stages increases (x-axis). A value of 0 for the pruning and merging stage gives the initial number before any iteration.

Supplementary Table 1

Population	OMNI	1000GP	1000GP phased as Unrelated	1000GP phased as Duo	1000GP phased as Trio
ACB	102	0	0	0	0
ASW	97	61	17	22	22
CDX	100	0	0	0	0
CEU	104	85	83	0	2
CHB	100	97	97	0	0
CHD	1	0	0	0	0
CHS	150	100	0	0	100
CLM	107	60	1	1	58
FIN	100	93	93	0	0
GBR	101	89	88	1	0
GIH	100	0	0	0	0
IBS	150	14	0	0	14
JPT	100	89	89	0	0
KHV	121	0	0	0	0
LWK	100	97	97	0	0
MKK	31	0	0	0	0
MXL	100	66	10	0	56
PEL	105	0	0	0	0
PUR	111	55	0	0	55
TSI	100	98	98	0	0
YRI	161	88	3	12	73
Total	2141	1092	676	36	380

Number of 1000GP and OMNI samples per population. The first column gives the population acronym. The second and third columns give the numbers of OMNI and 1000GP samples respectively per population. The fourth, fifth and sixth columns give the number of 1000GP samples that are phased as unrelated or as part of a duo or a trio when phased in OMNI.

Supplementary Table 2

Population	Group	TGP1	CGI1 – TGP1	CGI1 ∩ TGP1	CGI2 – TGP1	CGI2 ∩ TGP1
ASW	AFR	61	0	5	0	0
LWK	AFR	97	3	1	0	8
MKK	AFR	0	4	0	0	0
YRI	AFR	98	2	7	2	19
CLM	AMR	60	0	0	0	0
MXL	AMR	66	3	2	0	0
PEL	AMR	0	0	0	37	0
PUR	AMR	55	0	2	0	0
CHB	ASN	97	0	4	0	0
CHS	ASN	100	0	0	0	47
GIH	ASN	0	4	0	0	0
JPT	ASN	89	0	4	0	0
PJL	ASN	0	0	0	2	0
CEU	EUR	85	3	6	10	51
FIN	EUR	93	0	0	0	0
GBR	EUR	89	0	0	0	0
IBS	EUR	14	0	0	0	0
TSI	EUR	98	1	3	0	0

Number of CG1, CG2 and 1000GP samples per population and continental group. The first and second columns give the population and the continental groups. The third column gives the number of 1000GP samples per population. The fourth and sixth columns give the numbers per population of CG1 and CG2 samples not in 1000GP. These individuals are used to carry out imputation experiments using the call sets. The fifth and seventh columns give the numbers per population of CG1 and CG2 samples also included in 1000GP. These individuals are used to carry out discordance analysis of the call sets.

Supplementary Table 3

Set	Type	Chr 10	Chr 20	Chr 1-22
TGP1	SNP	1,814,147	824,953	36,820,992
TGP1	Indel	68,118	30,010	1,389,601
CGI1 ∩ TGP1	SNP	747,042	341,674	15,060,295
CGI2 ∩ TGP1	SNP	867,976	399,477	17,399,956
CGI2 ∩ TGP1	Indel	27,511	12,751	554,886

Number of SNPs and Indels in common between 1000GP, CG1 and CG2.