Strategies for Phasing and Imputation in a Population Isolate

Supplementary Material

Anthony Francis Herzig (1,2)

Teresa Nutile (3)

Marie-Claude Babron (1,2)

Marina Ciullo (3,4)

Céline Bellenguez (5,6,7,8)

Anne-Louise Leutenegger (1,2,8)

(1) Université Paris-Diderot, Sorbonne Paris Cité, U946, F-75010 Paris, France

(2) Inserm, U946, Genetic variation and Human diseases, F-75010 Paris, France

(3) Institute of Genetics and Biophysics A. Buzzati-Traverso - CNR, Naples, Italy

(4) IRCCS Neuromed, Pozzilli, Isernia, Italy

(5) Inserm, U1167, RID-AGE - Risk factors and molecular determinants of aging-related diseases, F-59000 Lille, France

(6) Institut Pasteur de Lille, F-59000 Lille, France

(7) Université de Lille, U1167 - Excellence Laboratory LabEx DISTALZ, F-59000 Lille, France

(8) These authors contributed equally to this study

Correspondence:

Anthony Francis Herzig

INSERM UMR 946

27 Rue Juliette Dodu

75010, PARIS, FRANCE.

anthony.herzig@inserm.fr

+33172639313

Supplementary Materials

Data Simulation

Founder haplotypes were created with HapGen using a theoretical population size of 3,000 and the default mutation rate. This choice led to simulated data with similar kinship to the observed genotypes in Campora (Supplementary Figure 3). The percentage of sites phased by SLRP serves as a good proxy for the proportion of IBD that can be found within the sample. When phasing the true data from Campora, 99% of heterozygous sites were phased, a similar percentage to those observed on the HapGen+Pedigree simulation, Supplementary Figure 8a.

Error Models.

Here we describe our error model for the WGS data of the 93 SSP individuals. For each simulated genotype, a set of bases was sampled from the two possible alleles of the genotype in order to represent the bases across multiple reads containing the position. Error bases are simulated within this set and can take any value out of A,C,T and G. The depth of the set was randomly selected from the depths observed in the Campora WGS data at the corresponding position. We then used the approach of Kim et al. (2011) to make approximate calculations of genotype likelihoods from which we calculated genotype qualities based on the models implemented by the next-generation sequence calling software GATK (DePristo et al., 2011).

As in to Kim et al. (2011), our error models were not symmetric. Lower genotype quality was observed in Campora on AT and GT SNPs. Hence we simulated higher error rates for error types A → T, T → A, G → T, and T → G as shown in Supplementary Table 1. For all genotypes, the error rate from the true base to the base corresponding to the other possible allele at the position (according to the simulated genotype) was augmented by 1/120. For example, when simulating error bases for a true base of A at a position with alleles A and G present in the simulated data, the error rate A → G was 1/120 + 1/120. However, if the alleles present in the data were A and T, the error rate A → G would remain at 1/120 and the error rate A → T would be augmented by 1/120. Such error models were chosen in order to create similar distributions of genotype quality as had been observed in Campora and overall genotyping error rates high enough to be of interest when analysing their effect on phasing and imputation. Our error models do not attempt to provide a faithful representation of the calling of genotypes from raw sequence reads but simply to create errors and missingness simply, randomly, and in similar patterns to those observed in WGS data in Campora.

Quality Control

ARRAY variants were removed for high missingness (> 5%), low MAF (< 1%), and significant deviation from Hardy-Weinberg equilibrium ($p < 10^{-5}$). WGS variants were removed for high missingness (> 10%), low Minor Allele Count (< 2), and significant deviation from Hardy-Weinberg equilibrium ($p < 10^{-5}$).

Phasing

When using EAGLE2+1000G we set the parameter 'pbwtiter' to 3 which significantly improved the phasing and ensures that phasing inference was made not just from the 1000G but from estimated haplotypes of other individuals in the sample. When phasing with BEAGLE we found that results were generally robust to changes in the 'window' and 'overlap' parameters. For EAGLE2 and BEAGLE we allowed multiple threading (four threads) after observing that restricting to one thread did not significantly change results. However, for SHAPEIT2 it is recommended to not use multiple threading so we used a single thread for each phasing run. Using BEAGLE with the 1000G as an external reference panel proved problematic as many variants were removed from the analysis by the algorithm due to high differences in MAFs between the sample and the reference panel. As population isolate data has been simulated, it is to be expected that MAFs differ from those observed in 1000G. We thus did not present BEAGLE results for this option. We did not test SHAPEIT3 with the 1000G as a reference panel due to the similarity between SHAPEIT2 and SHAPEIT3. Otherwise software were used with default settings.

Note that it is not possible to calculate SER at the exact site of a genotype error or missing genotype as there is no true phase from which to make a comparison with. Hence all calculations are made irrespective of error sites in each replication. An error can still cause a SER in a direct way but this would be measured at the preceding and following heterozygous sites on the chromosome (Supplementary Figure 10).

IBD-Sharing

To create Supplementary Figure 11 for SHAPEIT2+duohmm+1000G and EAGLE2 we randomly selected 200 heterozygous sites with switch errors and 200 heterozygous sites which were phased correctly. For each of the 400 sites we then counted the number of haplotypes in the sample IBD to the individual at the site. This analysis was performed on ARRAY data with no genotyping errors or missingness simulated and on the Pedigree simulation where the exact IBD sharing information was accessible. This is because on the Pedigree simulation, we know exactly which founder haplotype has been copied at every site for each individual. On the HapGen+Pedigree simulation, we could not keep track of this information.

Imputation

Following user manual recommendations for IMPUTE2, the region was split into four regions each of width 5Mb and using a buffer region of 0.25Mb. Identical settings were used for IMPUTE4. Outputs from the four runs were then concatenated after imputation. We experimented with the 'k-haps' parameter in IMPUTE2 and were unable to observe significant changes in accuracy and so the default parameter was used. MINIMAC3 was run with default parameters as was BEAGLE as we found that results were not sensitive to changes in the 'window', 'overlap' and 'Ne' (effective population size) parameters. A detailed investigation on the effects of model parameters on IMPUTE2, MINIMAC3 and BEAGLE is to be found in Browning and Browning (2016). As population isolates are the subject of this investigation it might be suggested that a lower value of 'Ne' would theoretically be suitable. In the context of imputation, the 'Ne' parameter controls the expected rate of recombination. Whilst our simulated individuals were constructed as mosaics of founder haplotypes with relatively few recombinations, a high recombination rate is still required in order to model each individual as an imperfect mosaic of external reference haplotypes. For IMPUTE2 we took advantage of the 'merge-ref-panel' option to perform cross imputation between the 1000G and our WGS SSP.

Difference in MAF between sample and reference panel

We compared the absolute difference in MAF between the simulated data in each replicate and either the 1000G Europeans populations or the complete 1000G. We averaged these differences over all variants used to estimate imputation accuracy and compared them to the baseline mean difference in MAF between the UK10K (our source of founding haplotypes) and the 1000G (Supplementary Figure 15). Compared to the Pedigree simulation, the HapGen+Pedigree simulation strategy produced simulated data with greater disparity in MAF compared to the 1000G. It was possible to observe a pattern between this disparity in MAF and lower imputation accuracy (Supplementary Figure 16a). To illustrate the importance of this difference in MAF we selected variants with a high MAF in our simulated data set (>0.3) and a large difference in MAF compared to the 1000G (top 10% of all MAF differences). We also excluded variants with imputation quality score ('info') below 0.7 coming from imputation using IMPUTE2+1000G. In the Pedigree simulation an average of 2,340 variants fulfilled these criteria and the average 90th percentile of absolute MAF differences was 0.17. In the HapGen+Pedigree simulation there were an average of 2,166 variants and the average 90th percentile of absolute MAF differences was 0.20. We compared the mean imputation accuracy from imputation using IMPUTE2+1000G and IMPUTE2+SSP over this selection of variants to the mean imputation accuracy over a random selection of variants with similar MAFs (Supplementary Figure 16a). In both simulation strategies

variants with a large difference in MAF to the 1000G were harder to impute under IMPUTE2+1000G but were imputed with similar accuracy under IMPUTE2+SSP.

To investigate further, we also selected variants with either a MAF significantly higher in the 1000G reference panel than in the sample or vice-versa. We then calculated the percentage increase in imputation accuracy by changing reference panel from the 1000G reference panel to the SSP (Supplementary Figure 16b). To put these increases into context, we again selected random selections of variants with a similar MAF to the chosen variants but without the large differences in MAF between the sample and the 1000G. Variants with significantly higher MAF in the sample (compared to the 1000G reference panel) experienced the most benefit from the change of reference panel for imputation.

A final analysis was made on variants which were monomorphic in the sample. Such variants may represent the greatest difference in MAF between the sample and an external reference panel. We have compared imputation accuracy on the telomeric region of the short arm of chr10 (20Mb in length). In this region 102,100 variants (found in the UK10K) were simulated and 22% and 31% of these variants became monomorphic in the Pedigree and HapGen+Pedigree simulation strategies respectively due to the founder effects that we simulated. From each replicate of each strategy, we selected 100 monomorphic variants at random. From this selection, an average of 18% and 19% of the variants passed a 0.4 threshold on the IMPUTE2 imputation quality score 'info'. For each variant that passed the threshold, we called imputed genotypes from imputed dosages by assigning the genotype to the highest genotype likelihood if and only if one genotype likelihood exceeded 0.8. In Supplementary Figure 16c we present the number of individuals with an incorrect called genotype for the assembly of all variants considered across replicates. A few variants present extreme results, these were noted to be variants with extremely different MAF between the UK10K and the 1000G. For example, the highest point on the left panel of Supplementary Figure 16c has a MAF of 0.47 in the 1000G and 0.0026 in the UK10K. Supplementary Figure 16d shows a zoom-in on Supplementary Figure 16c.

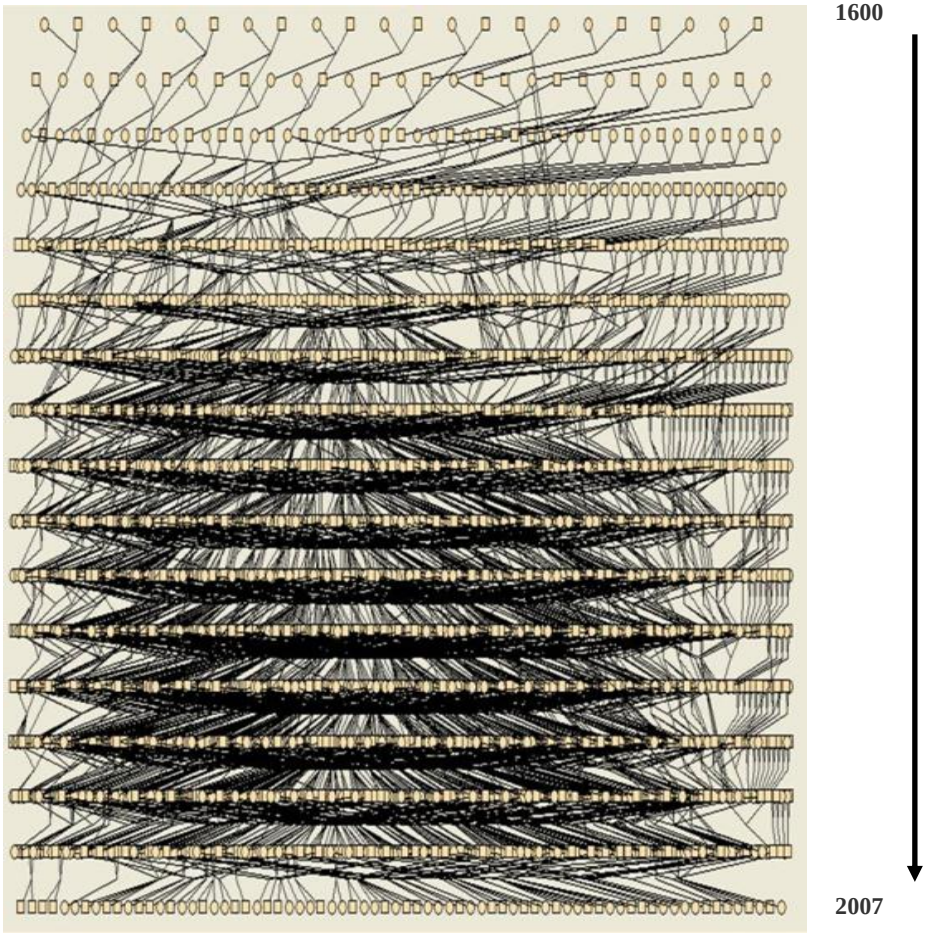
Genetic kinship coefficients between all 477 simulated individuals and the 1000G Europeans were computed using WGS positions after LD-pruning. The mean pairwise kinship on the Pedigree simulation was 1.2×10^{-4} and 2.5×10^{-6} on the HapGen+Pedigree simulation. Again this demonstrated greater dissimilarity between the HapGen+Pedigree simulated data and 1000G than between the Pedigree simulated data and 1000G.

Imputation Quality scores: 'info' and 'RSQ'

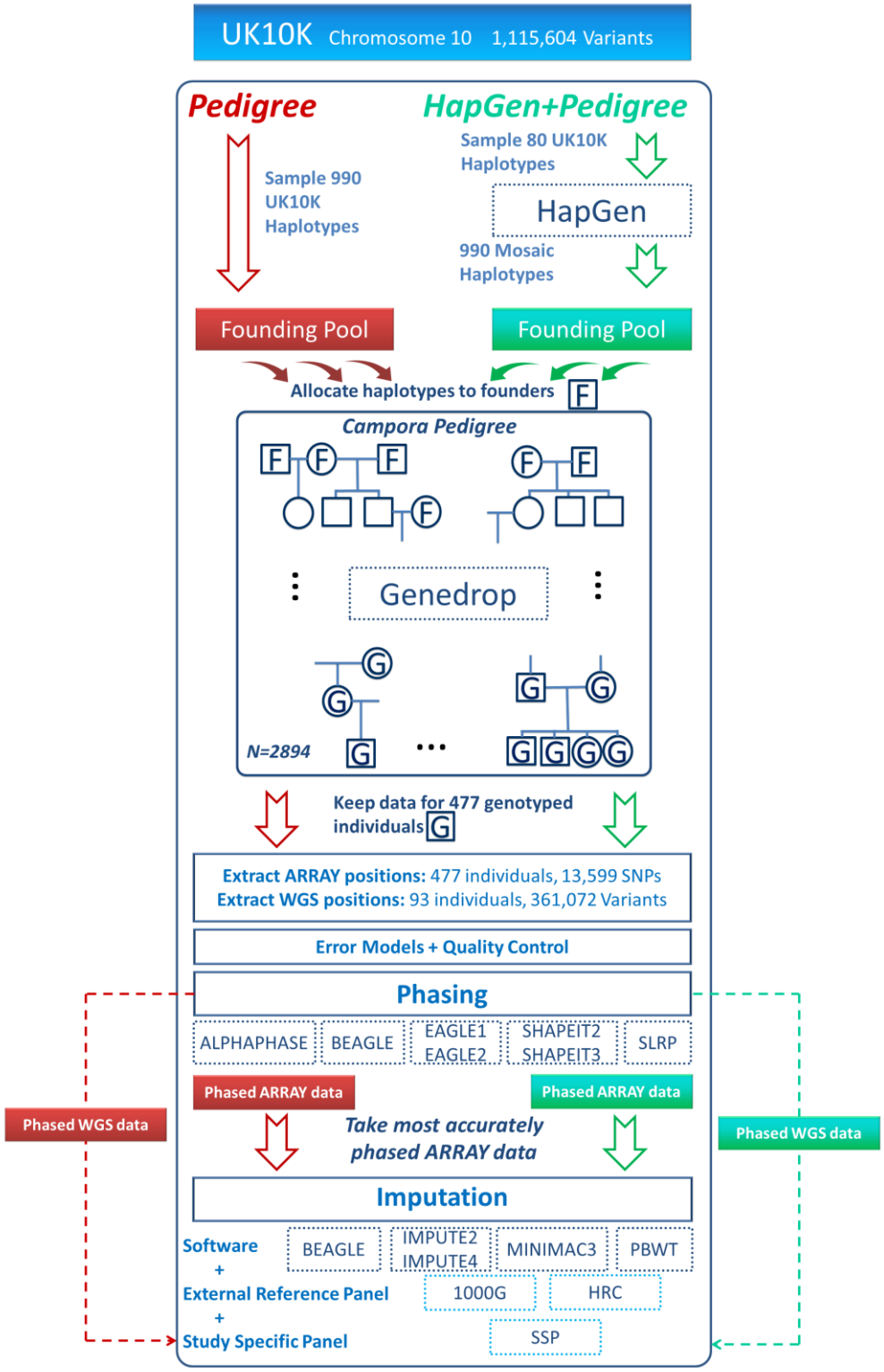
First, we applied the standard thresholds for common variants of 0.4 for 'info' and 0.3 for 'RSQ' (Li, Willer, Ding, Scheet, & Abecasis, 2010; Pistis et al., 2015) (Supplementary Figure 18a).

We then specified different thresholds for 'info' and 'RSQ' and calculated the resulting mean imputation accuracy in the remaining variants (Supplementary Figure 18b) under MINIMAC3+1000G and IMPUTE2+1000G on the HapGen+Pedigree simulation strategy. We observed that increasing the thresholds continued to give gains in mean imputation accuracy at a price of removing large numbers of variants. Particularly for low MAF variants, greater increases in imputation accuracy were observed by placing thresholds on the 'RSQ' measure than 'info'. Furthermore, the mean imputation accuracy of remaining variants became almost equal across all MAF bins when using the 'RSQ' measure while greater differences remain between MAF bins when using the 'info' score.

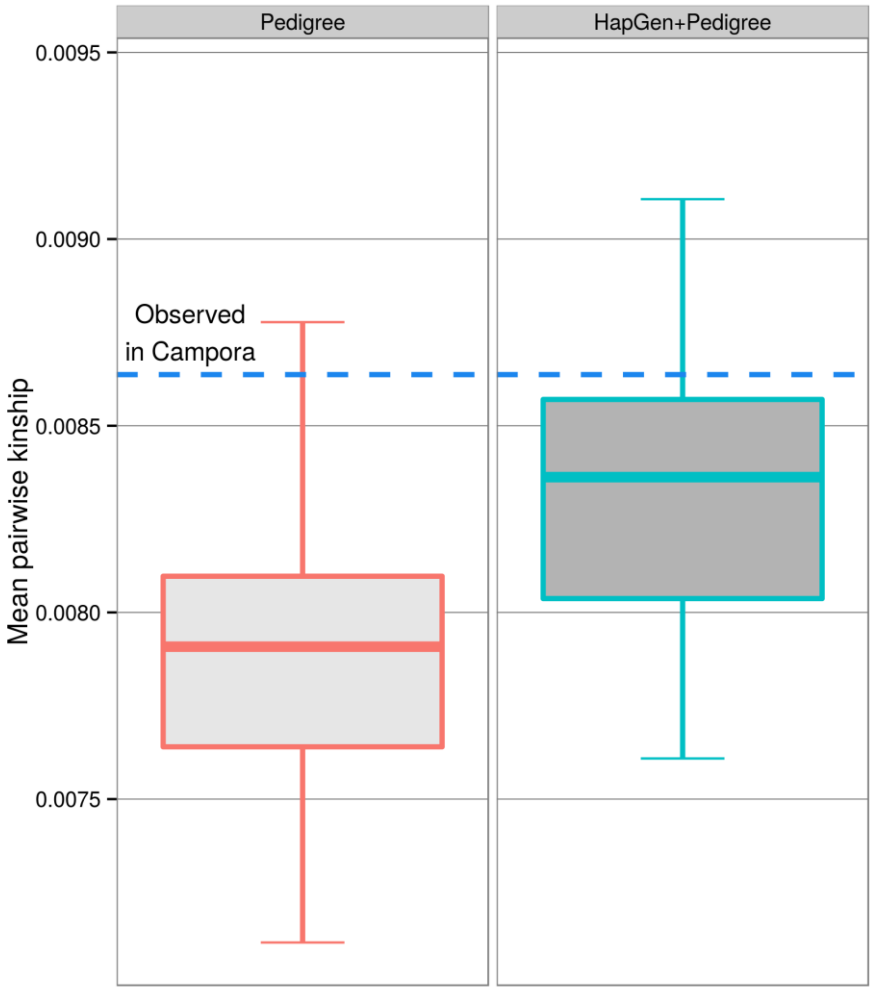
By defining sets of well and poorly imputed variants (imputation accuracy above 0.5 or below 0.2) we observed that the standard thresholds of 0.4 for 'info' and 0.3 for 'RSQ' fail to remove many poorly imputed variants (Supplementary Table 4). Furthermore for rare variants, the quality scores have less ability to separate well imputed variants from poorly imputed variants as reported by others (Liu et al., 2012; Pistis et al., 2015). To ensure that the majority of poorly imputed low MAF variants will be removed, higher thresholds than the standard ones are required. For common variants we observed that similarly high thresholds could be used and only a small number of well imputed variants would be lost and more poorly imputed variants would be removed. The choice of threshold represents a compromise between attempting to remove all badly imputed variants while hoping to not discard too many well imputed variants that could be highly valuable to subsequent analyses. Poorly imputed variants could give false positive results. However, if the motivation for imputation was envisaged single point analyses, the damage would be minimal as the researcher could still access the 'info' or 'RSQ' scores in order to see whether significantly associated variants had a very high imputation quality score or one just above the threshold. If multipoint analyses (gene-based or haplotype based) were envisaged, then poorly imputed variants have the potential to cause false negative results which would be harder to rectify; suggesting that in this scenario higher thresholds should be taken.



Supplementary Figure 1. The pedigree of Campora as recorded from parish records.

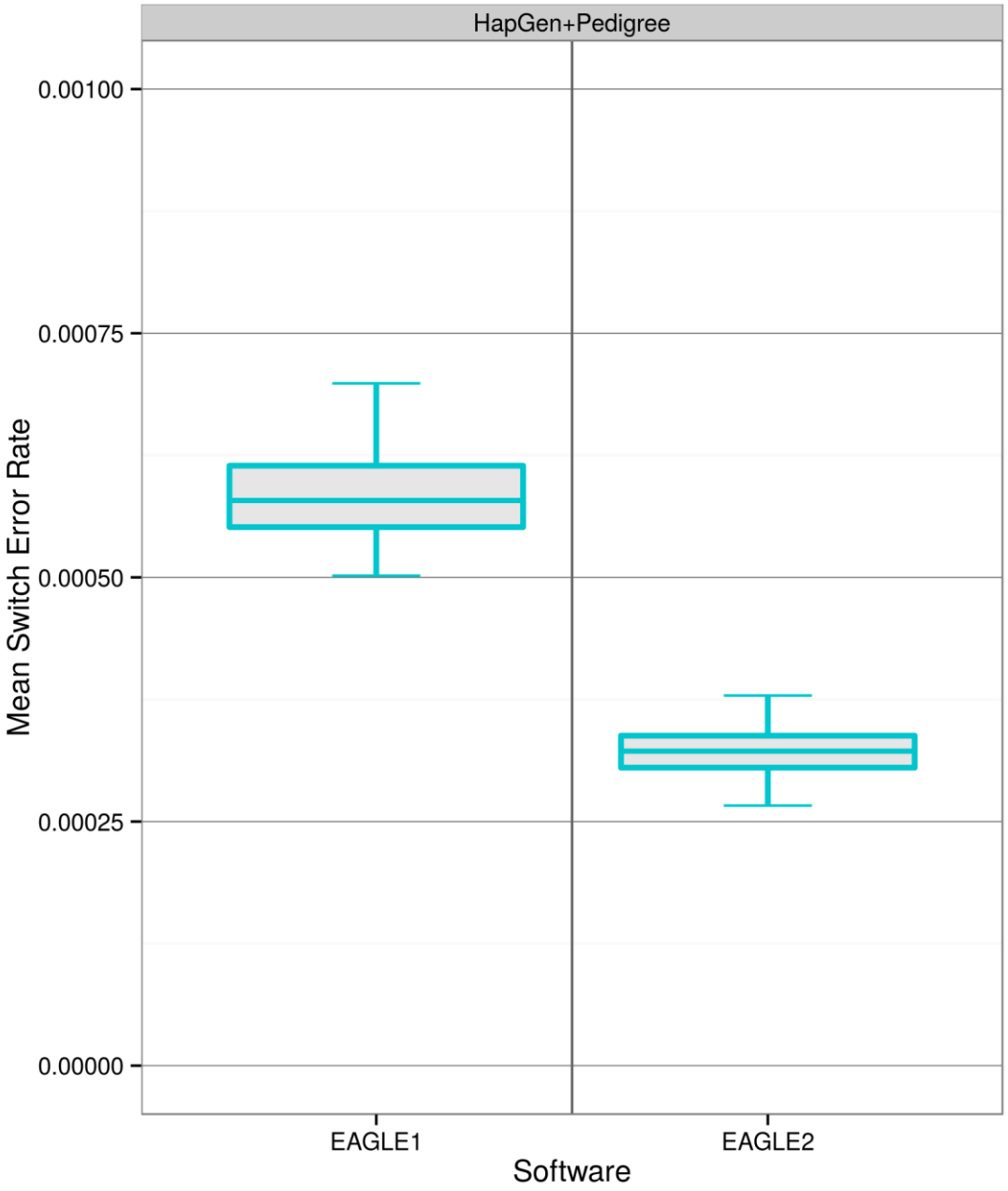


Supplementary Figure 2. Schematic of the two simulation strategies.

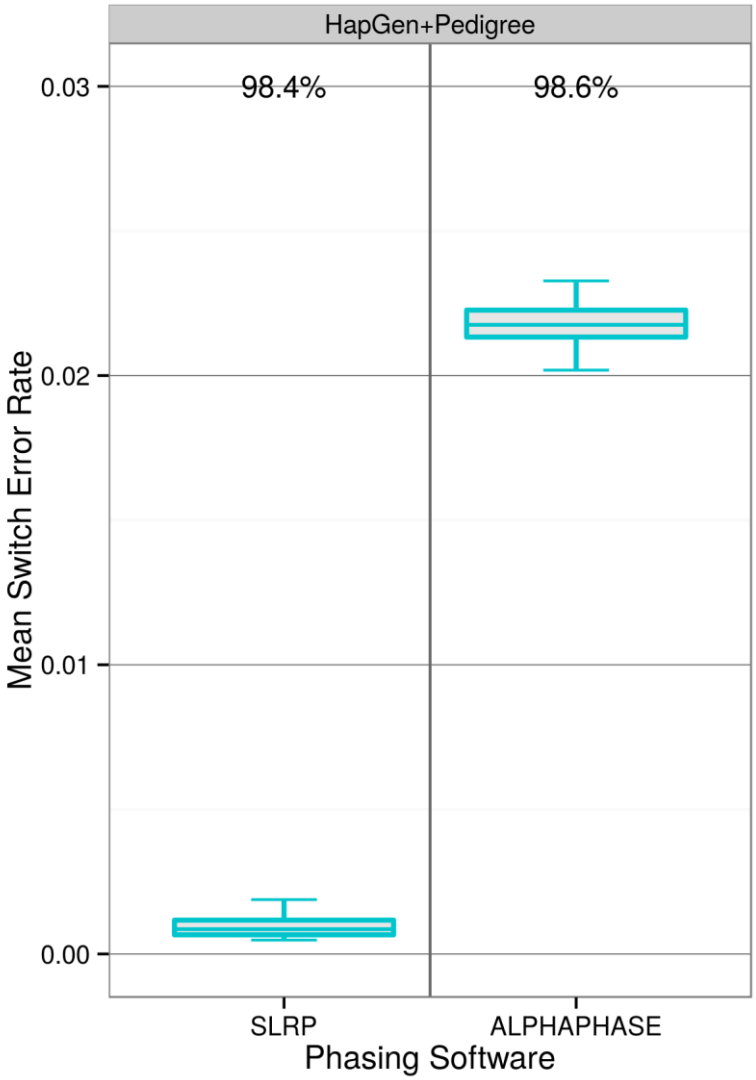


147

148 **Supplementary Figure 3.** Comparison of mean pairwise genetic kinship coefficients estimated on simulated
149 ARRAY data for 477 individuals for both simulation strategies. The HapGen+Pedigree simulation created data
150 with closer mean pairwise genetic kinship to the mean pairwise genetic kinship calculated on the observed
151 genotypes in Campora for the same individuals (dashed line). As every pedigree founder haplotype is first
152 generated from 80 UK10K haplotypes in the HapGen+Pedigree simulation, the pedigree founders are no longer
153 independent and share regions of IBD. Proportions of IBD are consequentially elevated throughout the sample
154 and surpass those predicted solely by the pedigree information.



Supplementary Figure 4. Mean Switch Error Rates for EAGLE1 and EAGLE2 on the HapGen+Pedigree simulation strategy.



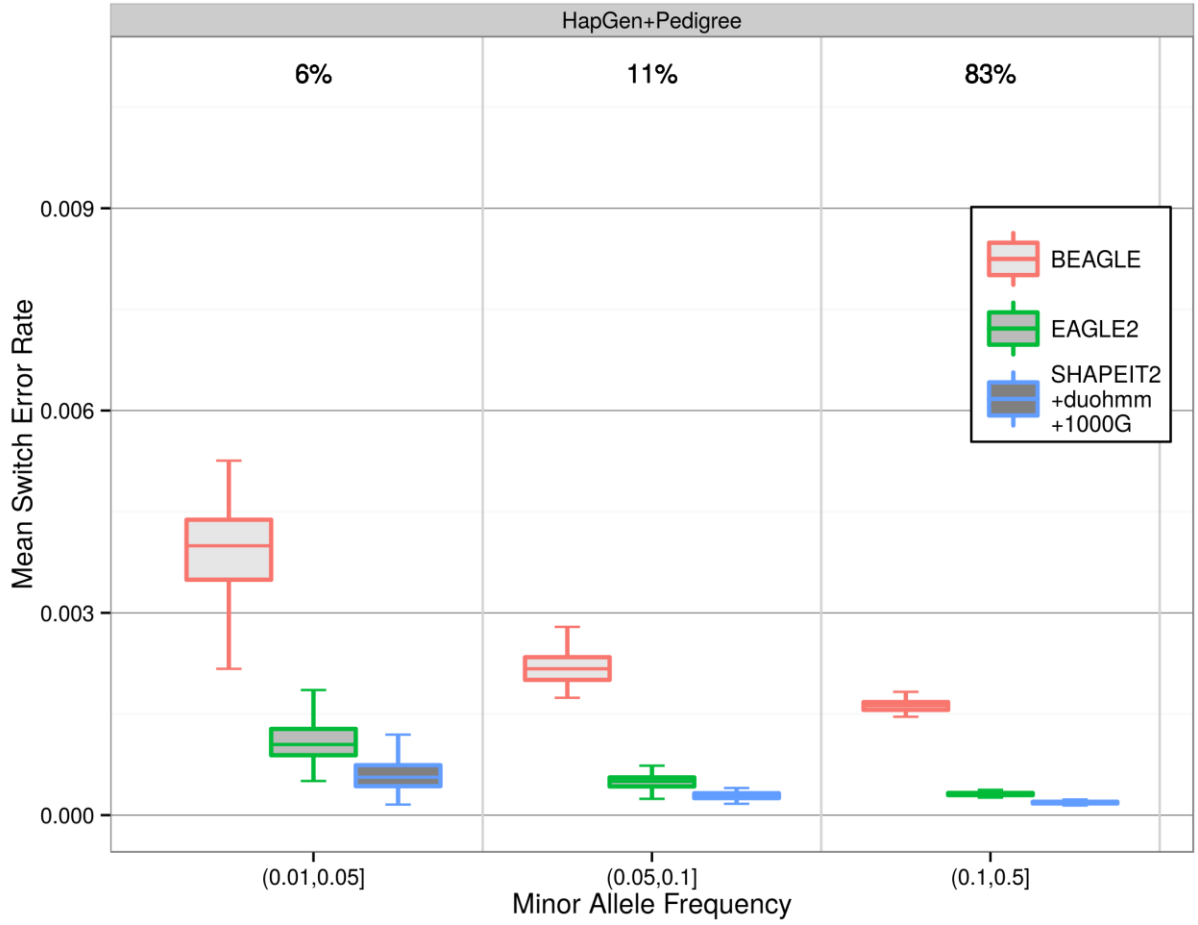
158

159 **Supplementary Figure 5.** Comparison of Long Range Phasing Software SLRP and ALPHAPHASE on the
160 HapGen+Pedigree simulation strategy. The percentages of heterozygous sites phased are displayed atop the
161 figure.

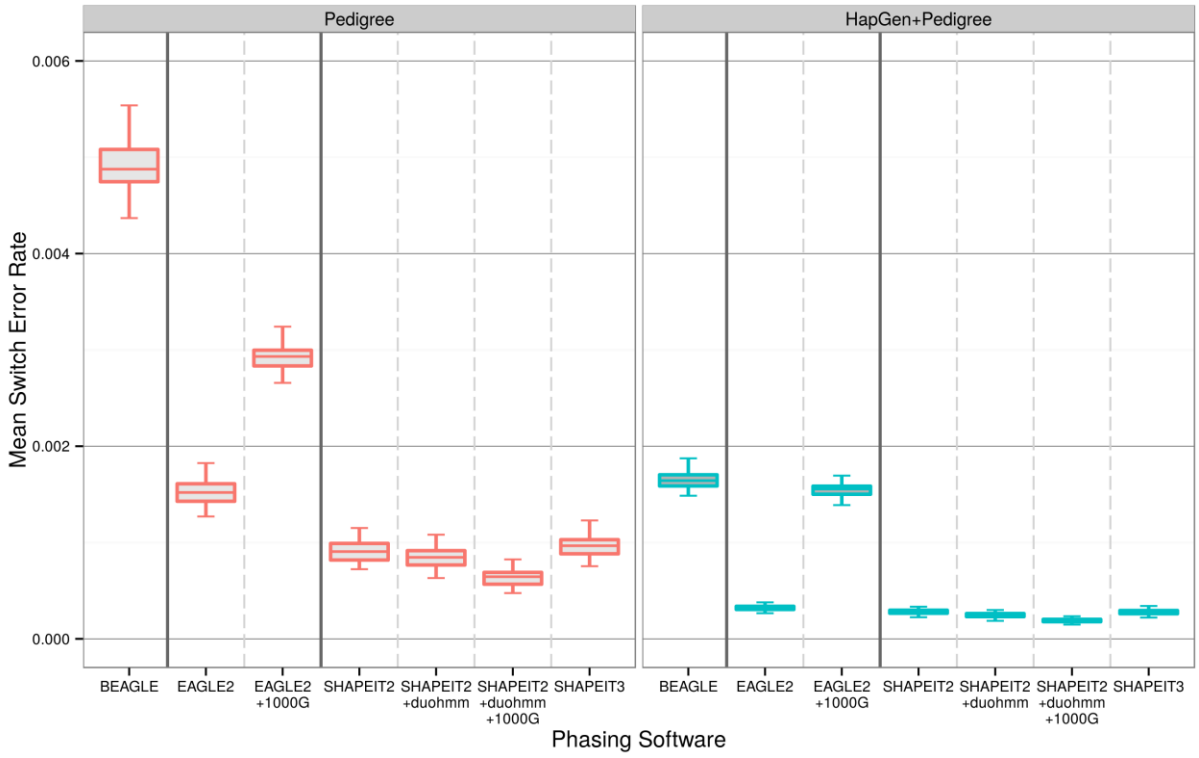
162

163

164



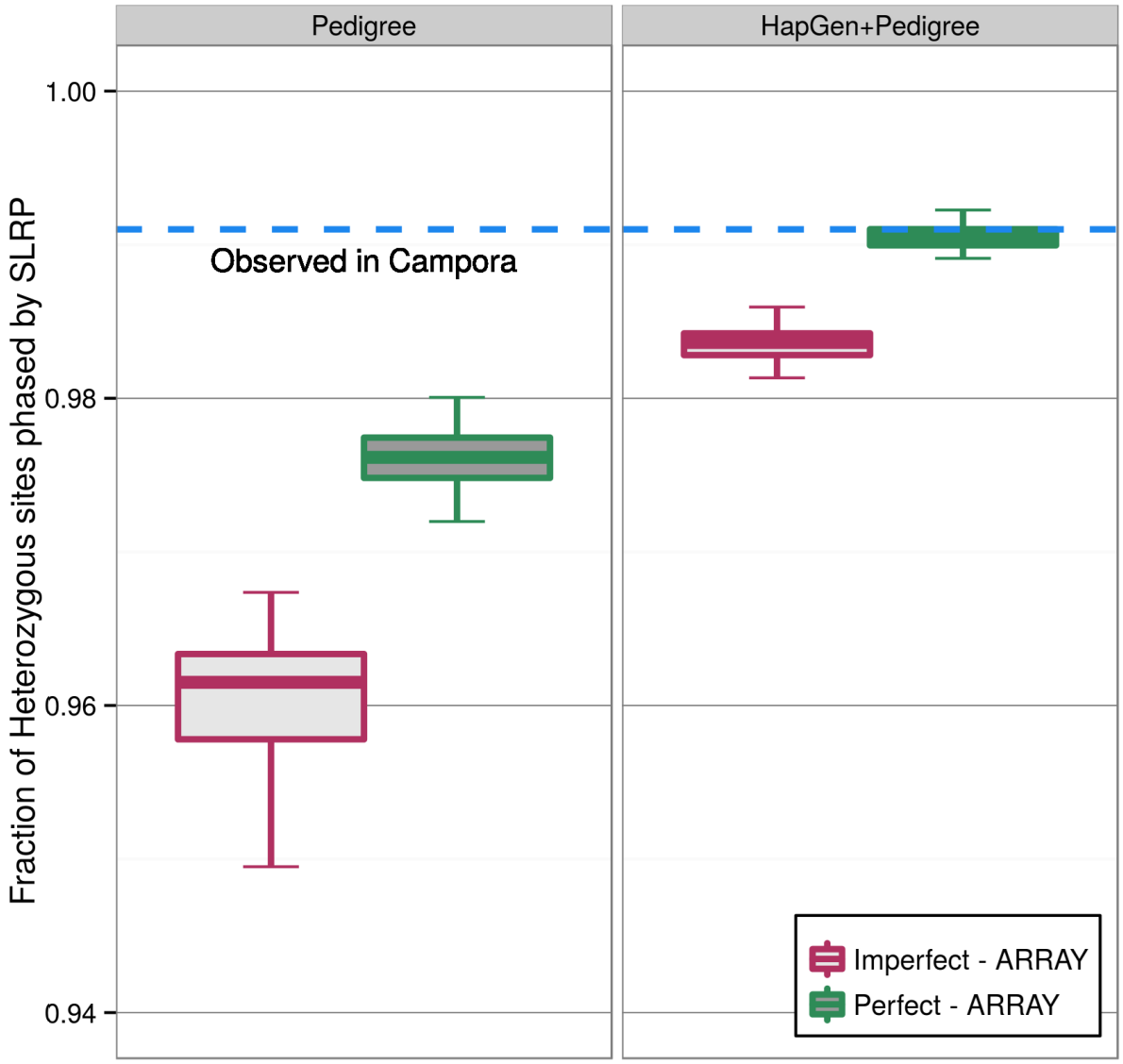
Supplementary Figure 6. Comparison of SERs according to MAF for BEAGLE, EAGLE2 and SHAPEIT2+duohmm+1000G on the HapGen+Pedigree simulation strategy. In each MAF bin, the mean SER over all variants is displayed. The percentages of variants in each MAF bin are displayed atop the figure.



171

172

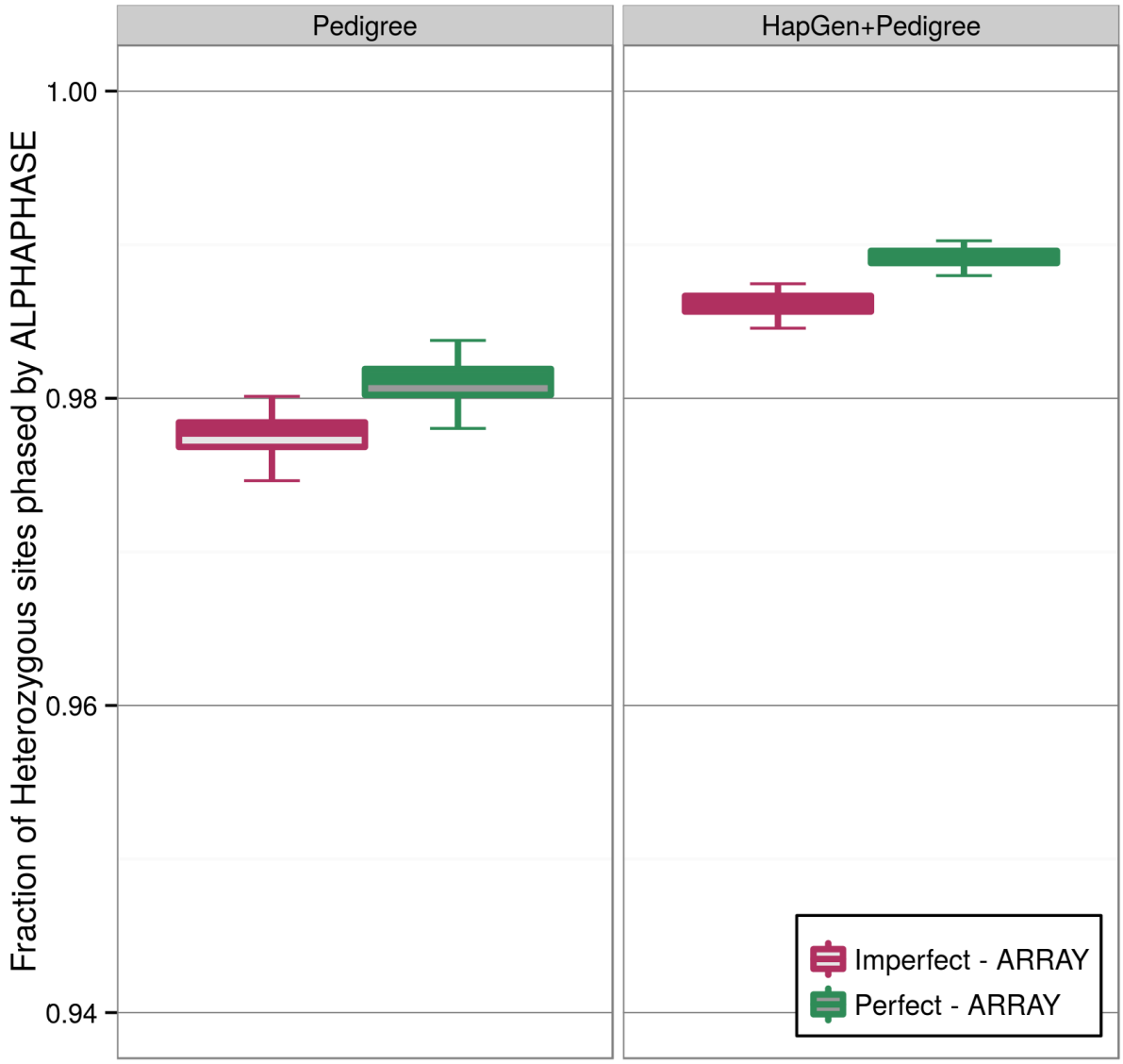
173 **Supplementary Figure 7.** Global SERs from both simulation strategies for all LD-based software.



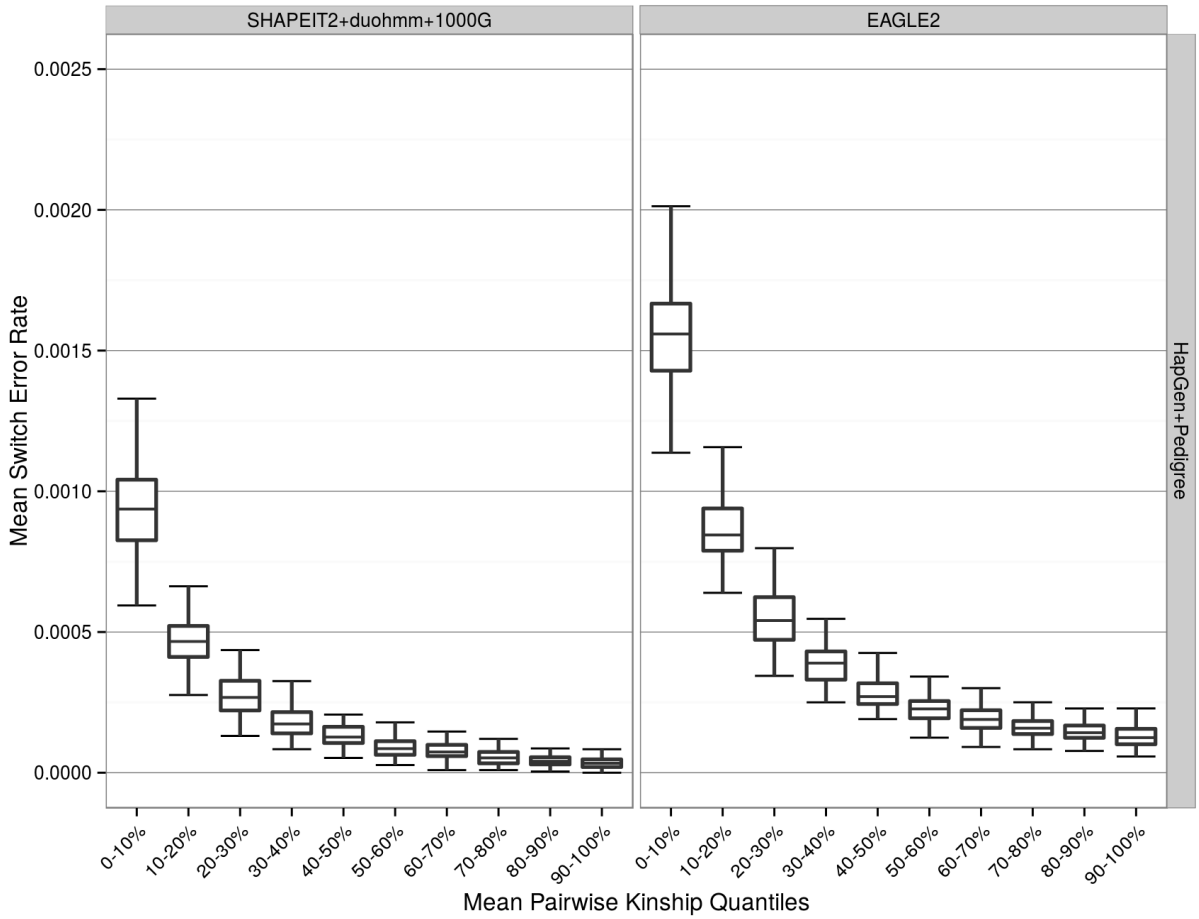
174

175 **Supplementary Figure 8a.** Comparison of the fraction of heterozygous sites phased by SLRP for both
176 simulation strategies. SLRP phased a higher proportion of sites when applied to the HapGen+Pedigree
177 simulation, similar to the proportion of sites as when applied to the observed ARRAY genotypes in Campora by
178 SLRP (blue line). Genotype errors and missingness led to a reduction in the number of sites that SLRP was able
179 to phase in both simulation strategies.

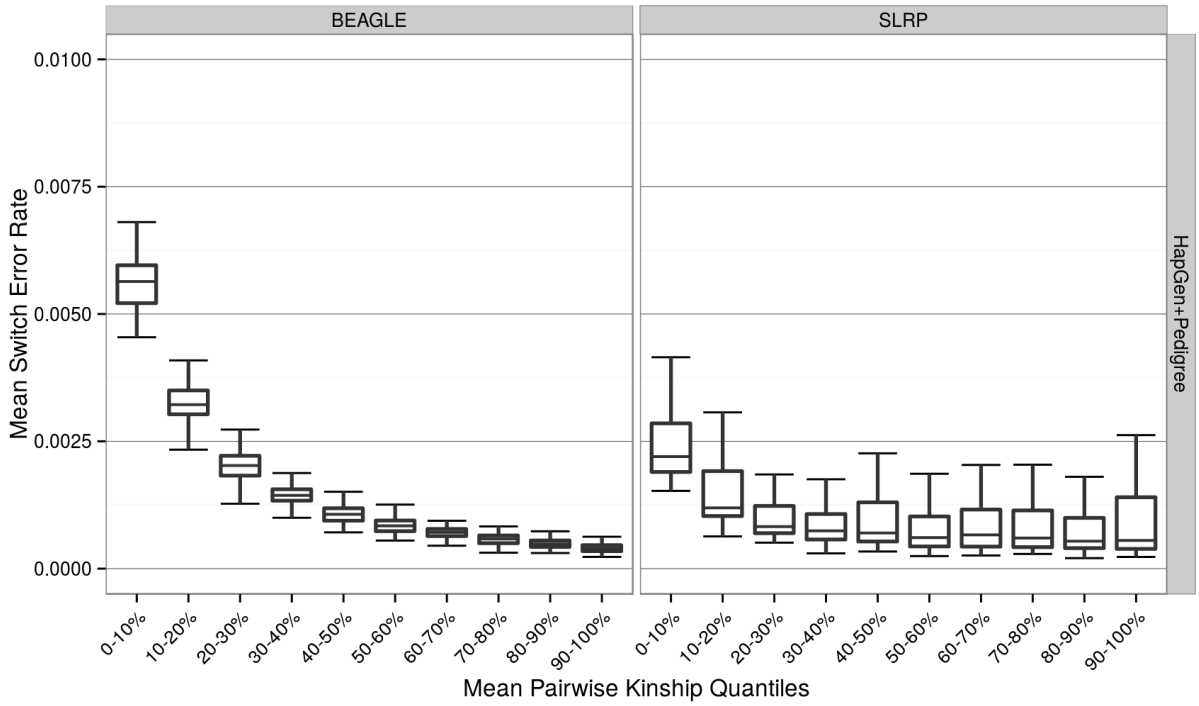
180



Supplementary Figure 8b. Comparison of the fraction of heterozygous sites phased by ALPHAPHASE for both simulation strategies. ALPHAPHASE phased a higher proportion of sites when applied to the HapGen+Pedigree simulation. Genotype errors and missingness led to a reduction in the number of sites that ALPHAPHASE was able to phase in both simulation strategies.



Supplementary Figure 9a. Relationship between mean pairwise genetic kinship and individual SER for SHAPEIT2+duohmm+1000G and EAGLE2. In each replicate, ARRAY genotypes were used to calculate the mean pairwise genetic kinship coefficient of each individual to all others. We considered 10 equally sized bins of mean pairwise genetic kinship based on the quantiles of the distribution of mean pairwise genetic kinship. In each group we then calculated the mean SER for all individuals in the group.

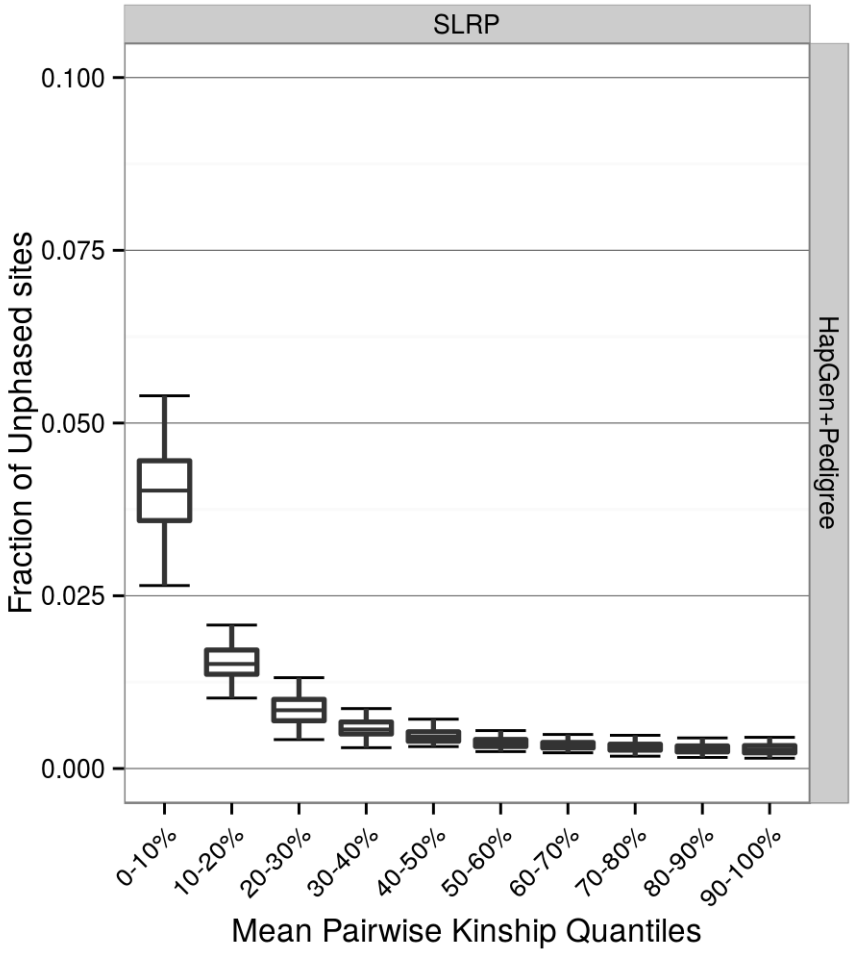


193

194 **Supplementary Figure 9b.** As Supplementary Figure 9a, but for BEAGLE and SLRP. Note the different scale

195 on the y-axis compared to Supplementary Figure 9a.

196



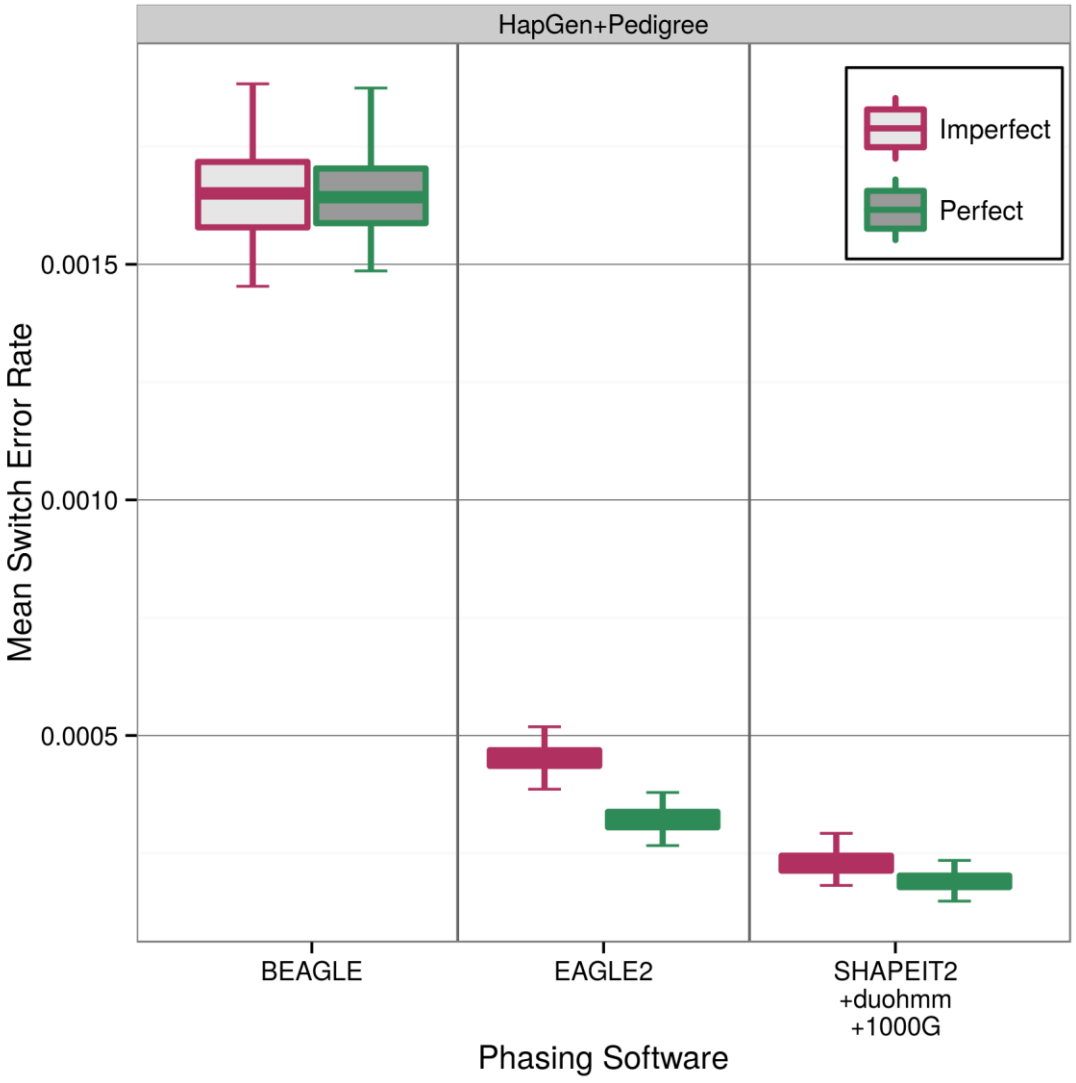
197

198 **Supplementary Figure 9c.** Fraction of all heterozygous sites left unphased by SLRP according to mean

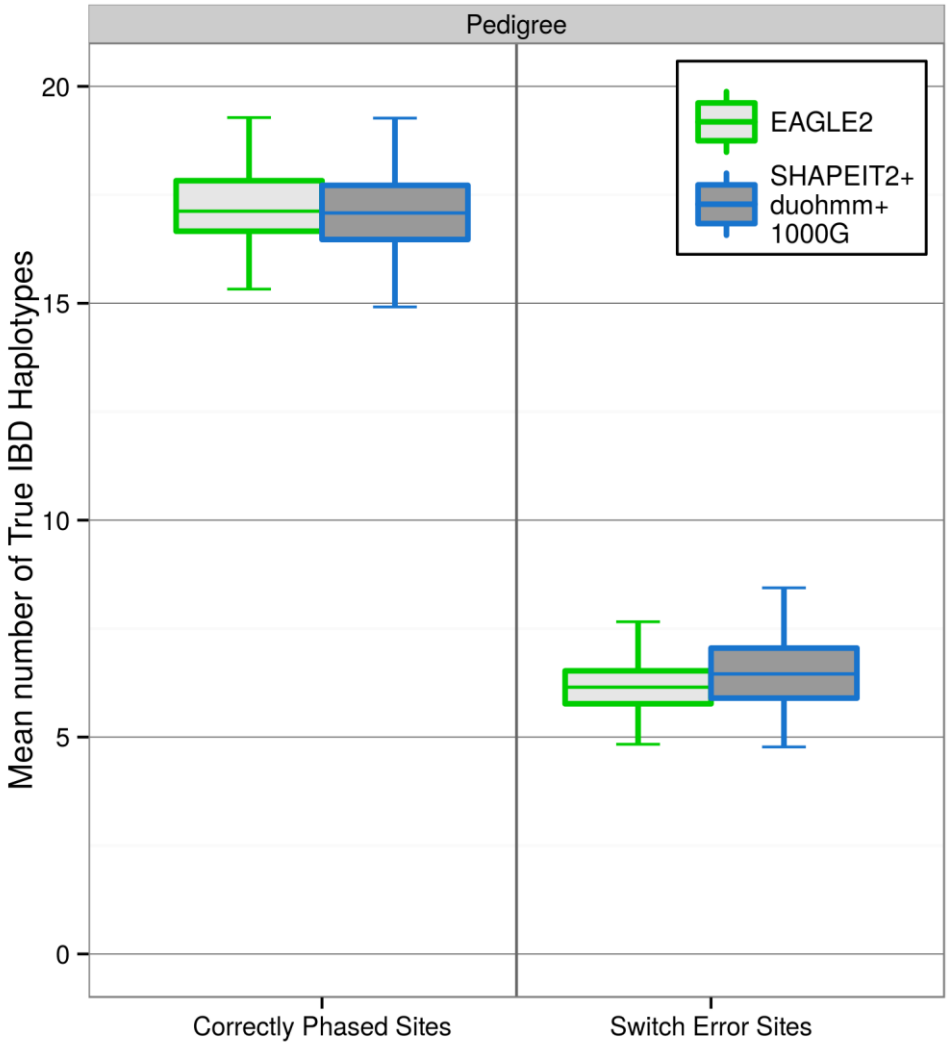
199 pairwise genetic kinship.

200

201



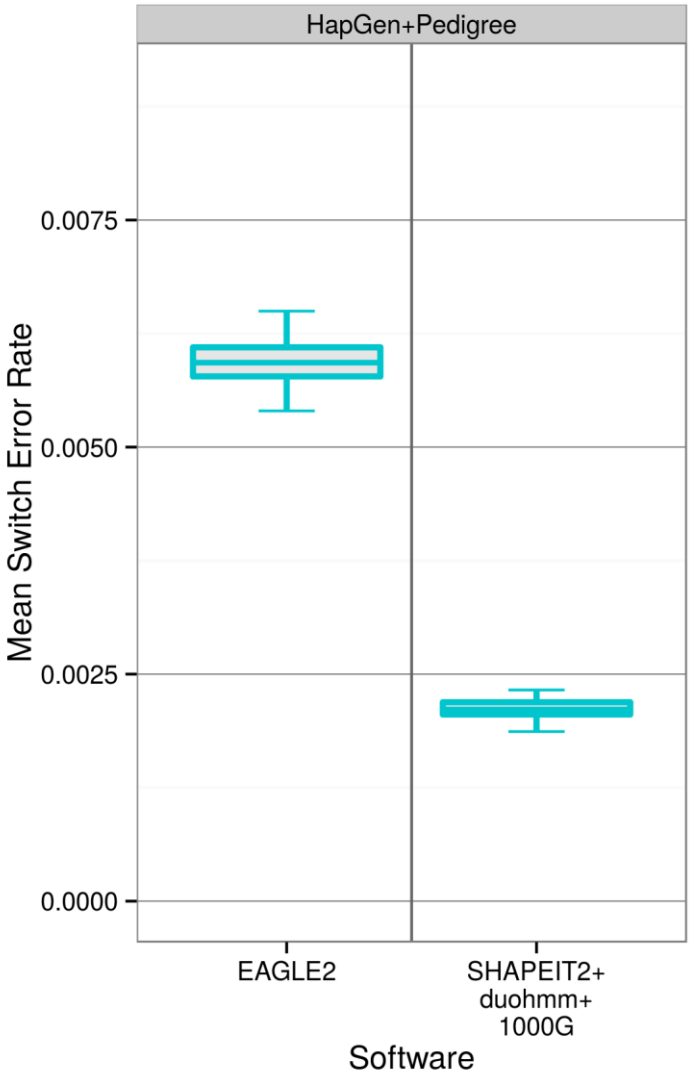
Supplementary Figure 10. Effect of genotype errors and missingness (Imperfect) on mean Switch Error Rate according to software and on the HapGen+Pedigree simulation strategy.



207

208 **Supplementary Figure 11.** Comparison of mean number of true IBD haplotypes at either correctly phased sites
209 or switch error sites for SHAPEIT2+duohmm+1000G or EAGLE2. This analysis was only possible on the
210 Pedigree simulation where the exact locations of simulated IBD-sharing were known.

211



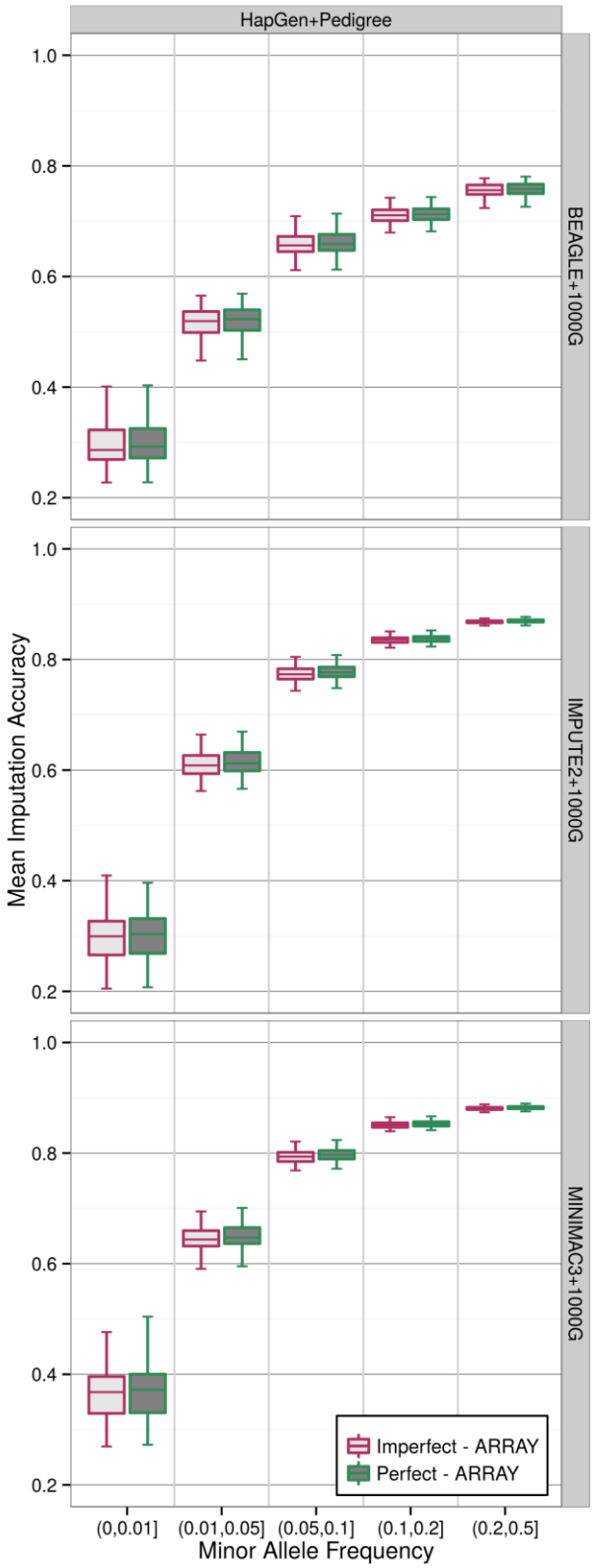
212

213 **Supplementary Figure 12.** Mean SER for the phasing of the 93 WGS SSP individuals with
214 SHAPEIT2+duohmm+1000G and EAGLE2. There are two likely causes for the higher SERs as compared to the
215 phasing of ARRAY data: firstly, a smaller number of individuals are involved, and secondly the WGS data
216 contained a higher proportion of variants with MAF below 0.05.

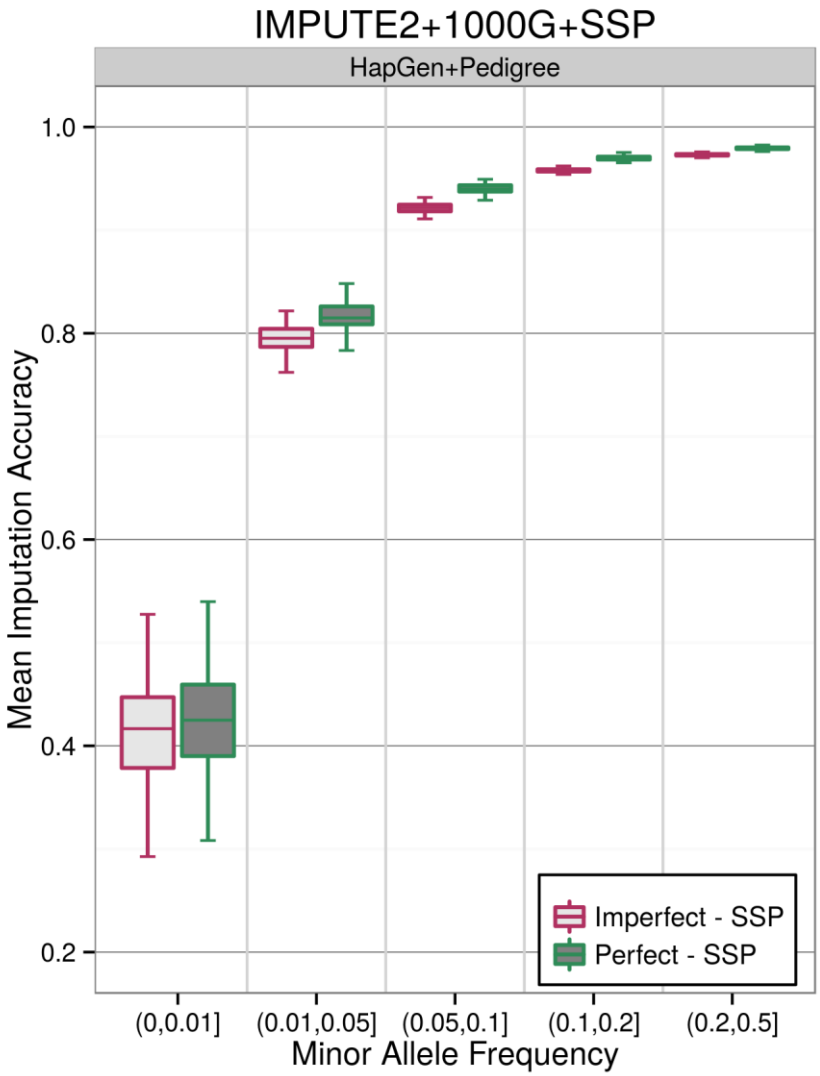
217

218

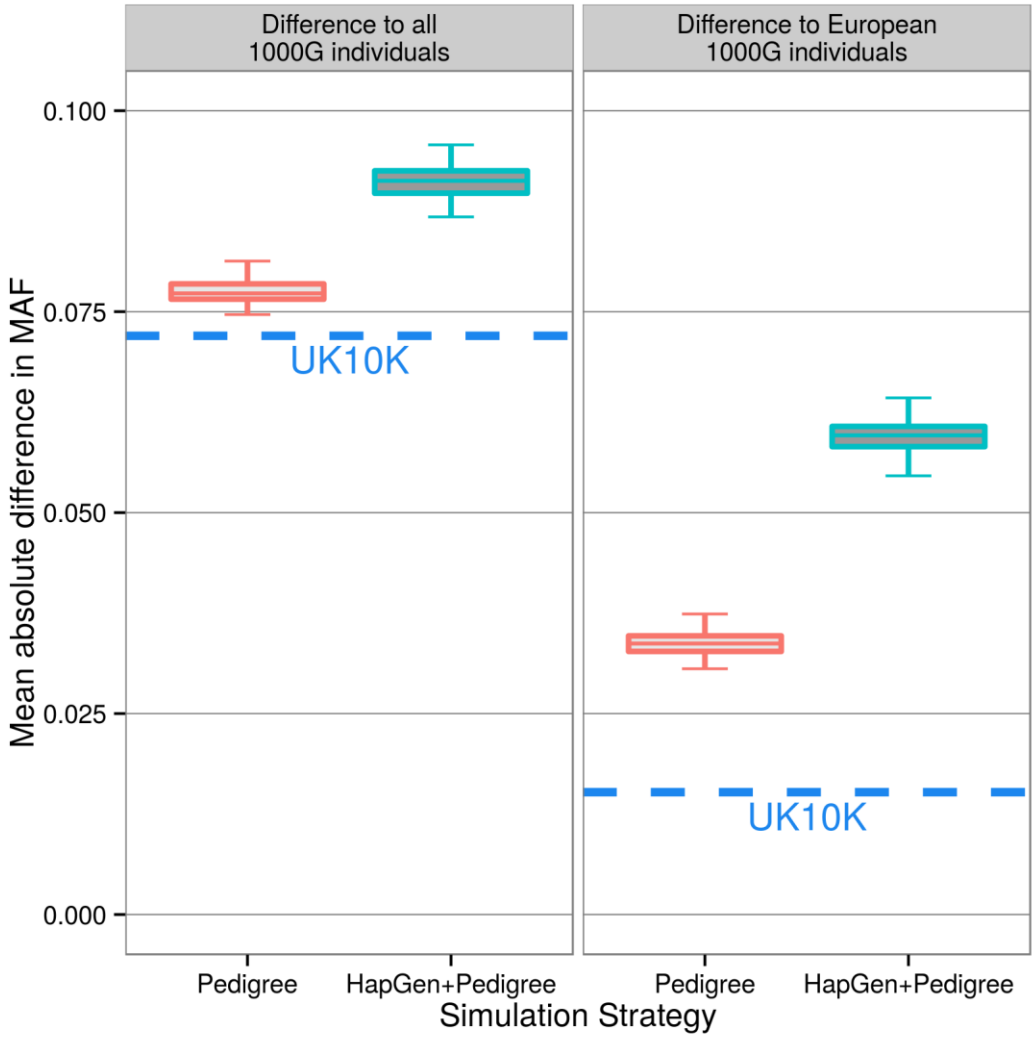
219



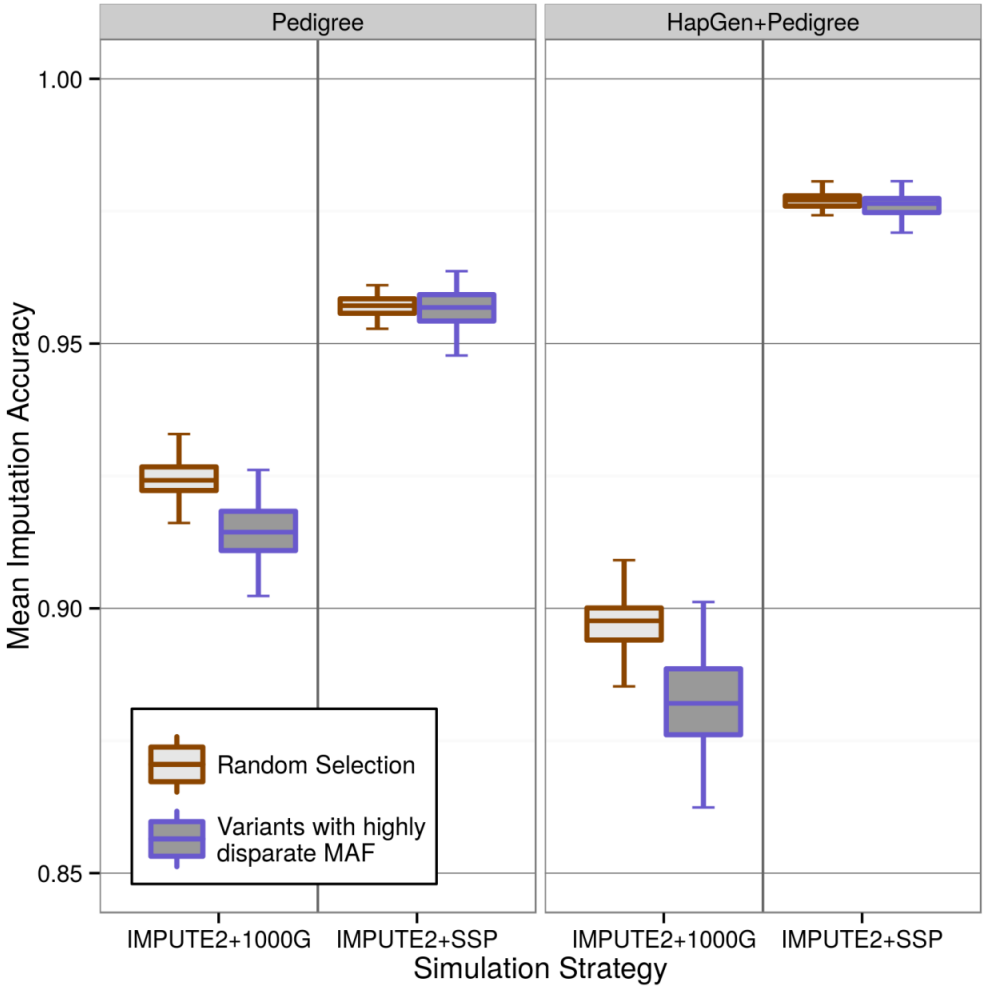
Supplementary Figure 13. Effect of genotype errors and missingness on the ARRAY data on imputation accuracy.



Supplementary Figure 14. Effect of genotype errors and missingness on the SSP on imputation accuracy. Imputation accuracy calculated from imputation strategy IMPUTE2+1000G+SSP.

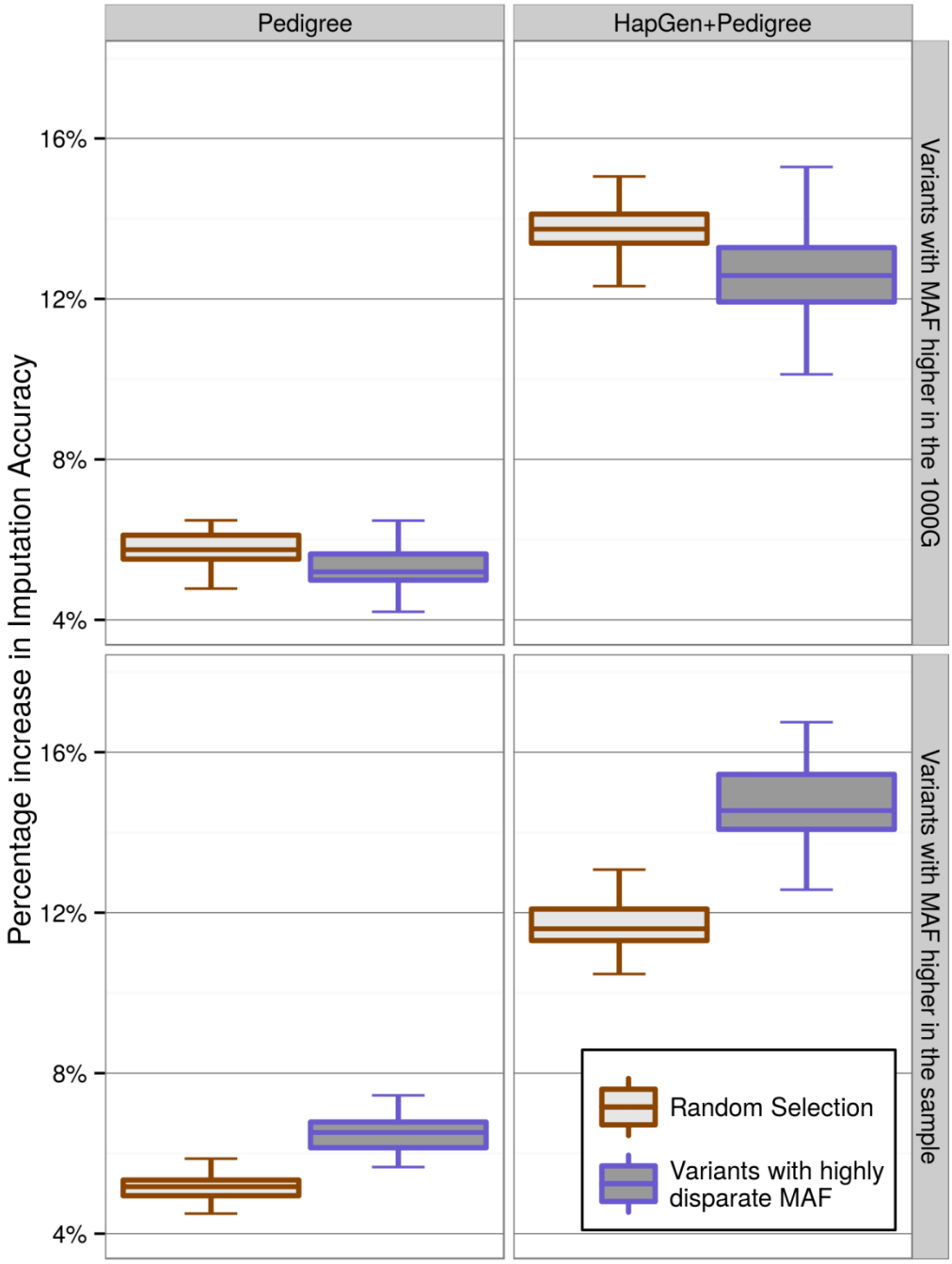


Supplementary Figure 15. Comparison of absolute difference in MAF between simulated data and the 1000G panel for both simulation strategies. Dashed lines represent the mean difference in MAF between the UK10K (founding pool used for the simulation) and the 1000G.



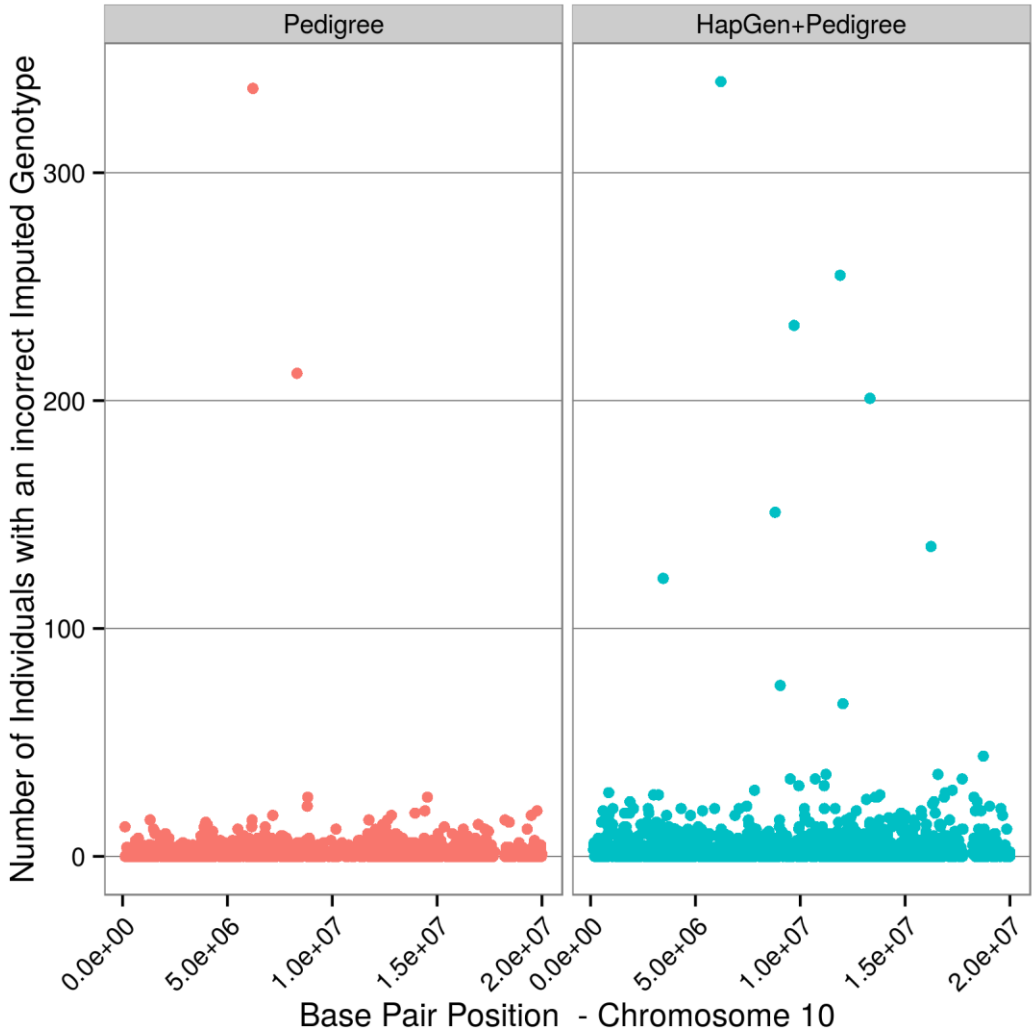
234

235 **Supplementary Figure 16a.** Comparison of imputation accuracy for sets of variants with particularly high
236 differences in MAF compared to the 1000G panel against random selections of similar variants without such
237 elevated disparities. Imputation accuracy was calculated from the imputation strategies IMPUTE2+1000G and
238 IMPUTE2+SSP.

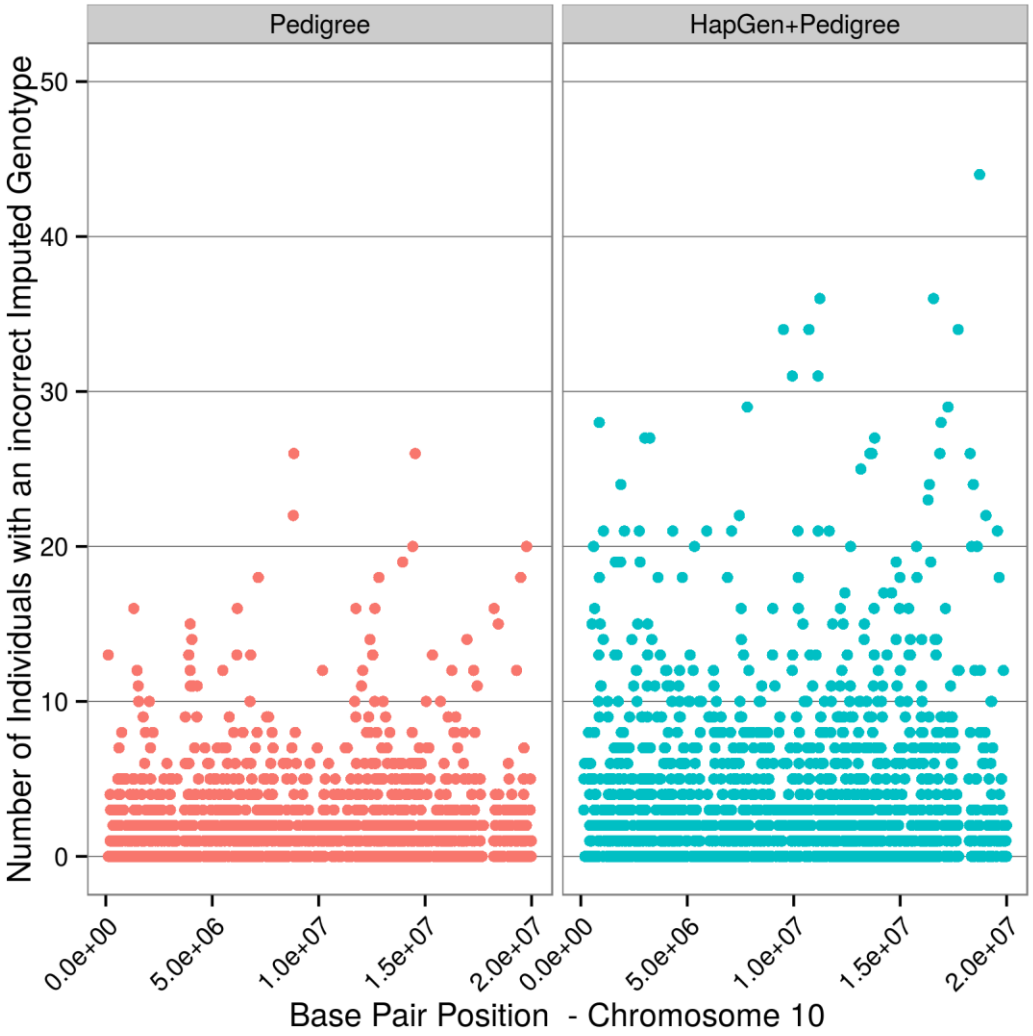


239

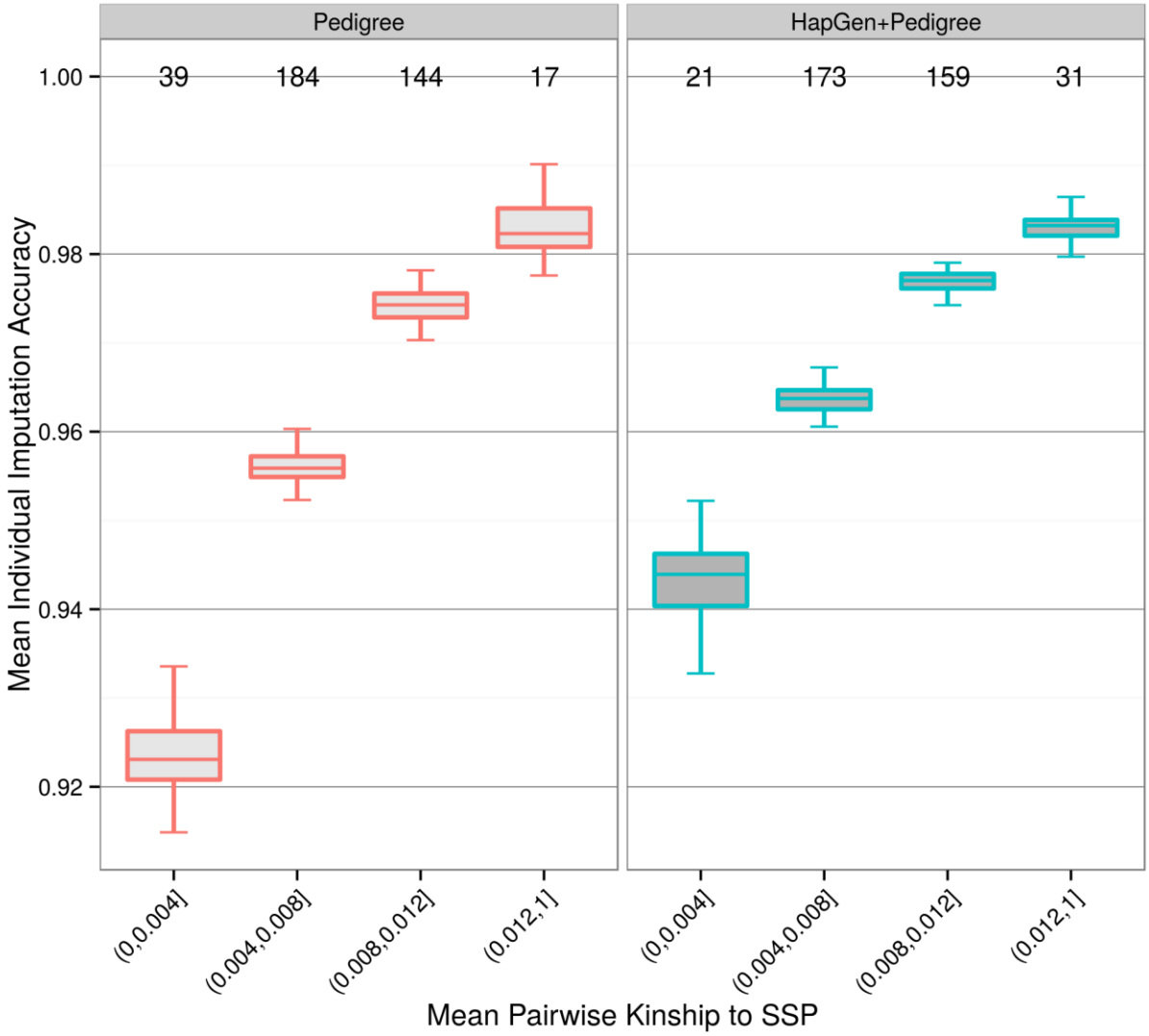
240 **Supplementary Figure 16b.** Increase in imputation accuracy by changing from IMPUTE2+1000G to
241 IMPUTE2+SSP for sets of variants with either MAF greater in the 1000G reference panel compared to the
242 sample or vice-versa. Once again, for each set of chosen variants for comparison, we selected a random
243 selection of control variants with similar MAF in the sample to the chosen set.



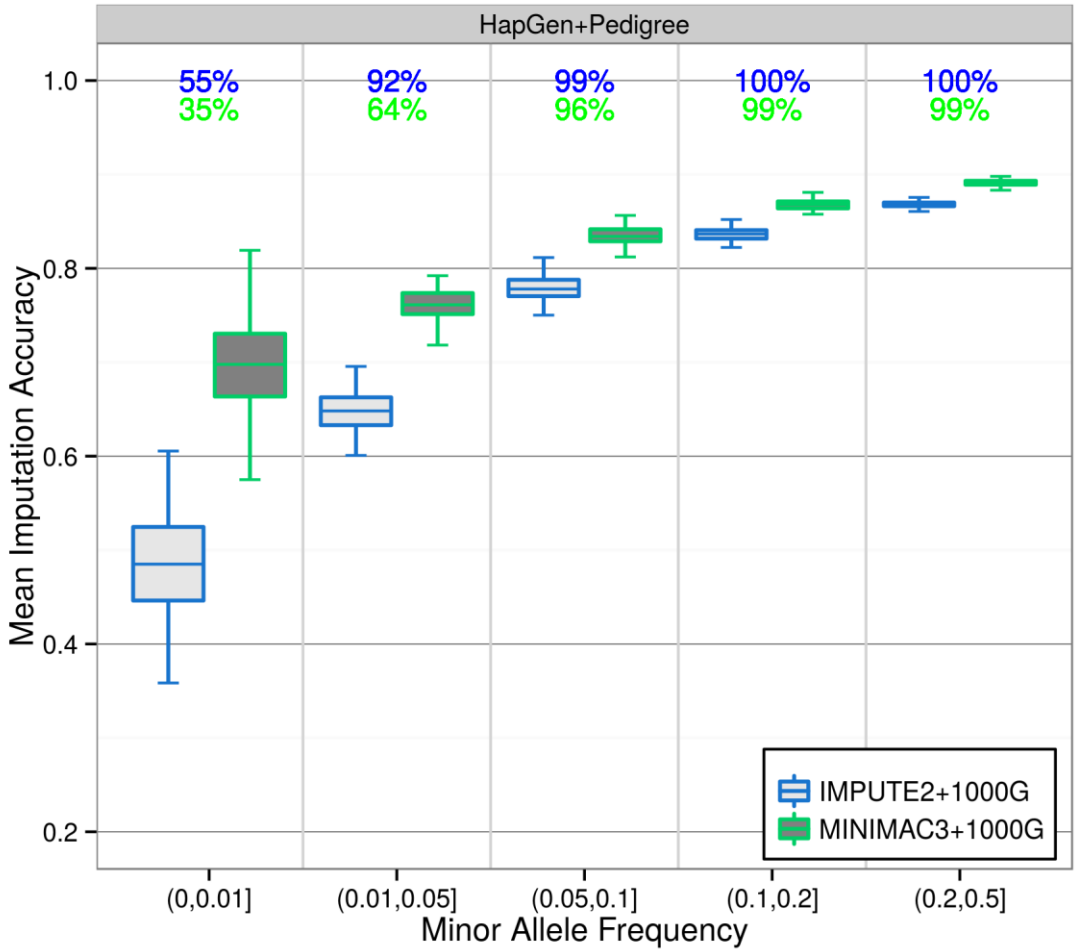
Supplementary Figure 16c. Imputation of monomorphic variants in the sample under IMPUTE2+1000G. The number of individuals with an incorrectly imputed genotype (after taking a hard call) against base pair position on chromosome 10.



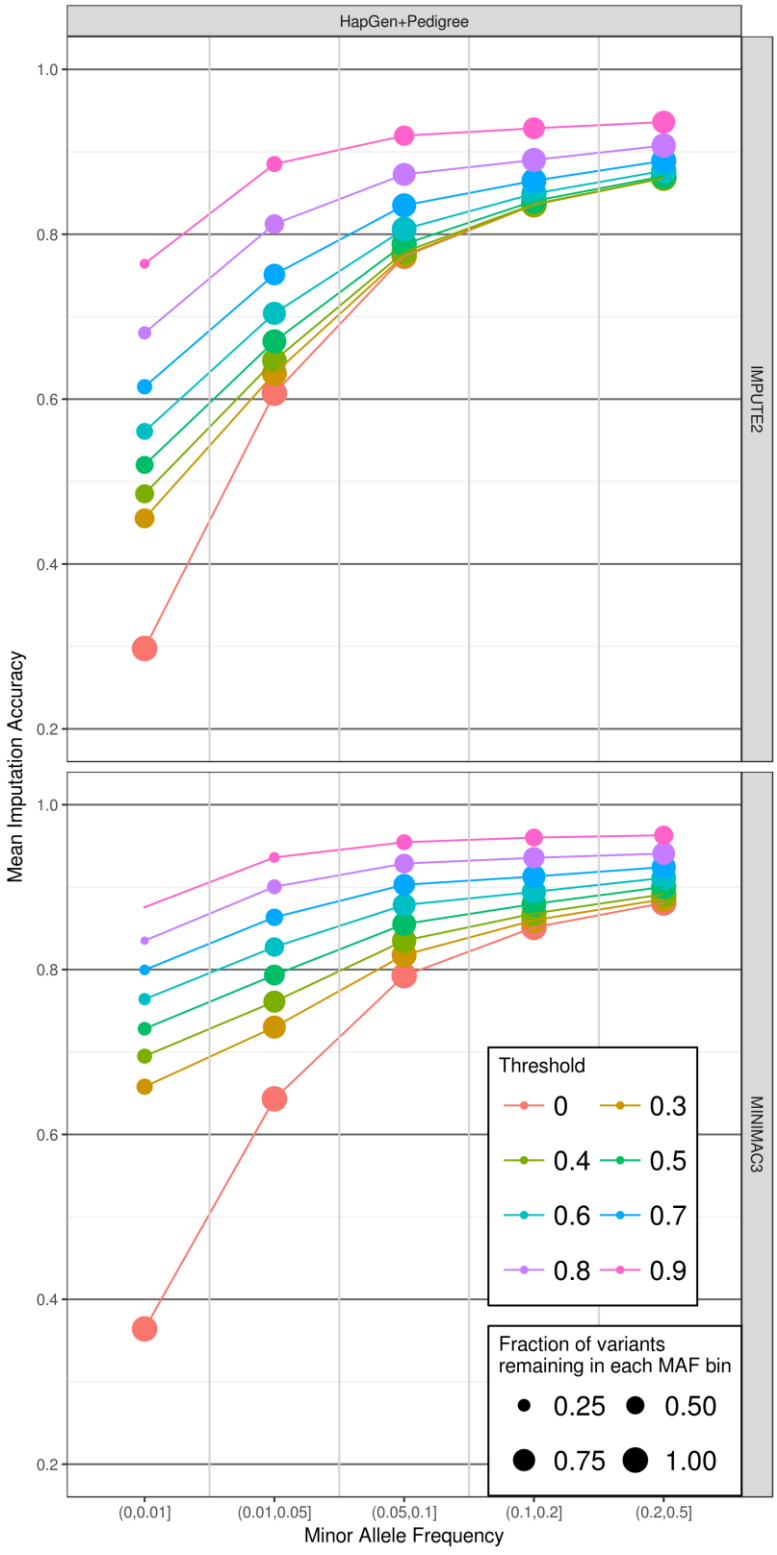
Supplementary Figure 16d. Zoom-in onto Supplementary Figure 16c showing the distribution of points with y-axis values less than 50.



Supplementary Figure 17. Comparison of individual imputation accuracy against the mean pairwise genetic kinship between each non-SSP member and all 93 SSP members. Mean numbers of individuals contributing to each bin of individual mean pairwise genetic kinship are displayed atop the figure. The minimum observed imputation accuracy for a single individual was just above 0.85 and just above 0.88 for the Pedigree and HapGen+Pedigree simulation strategies respectively.



Supplementary Figure 18a. Imputation accuracy across all MAFs following post imputation quality control based on a 0.4 threshold on 'info' scores for IMPUTE2 and a 0.3 threshold on 'RSQ' scores for MINIMAC3. These are the often recommended thresholds for 'info' and 'RSQ'. The percentages of variants that remain in each MAF bin after thresholding are displayed atop the figure (blue for IMPUTE2 and green for MINIMAC3).



Supplementary Figure 18b. Imputation accuracy across all MAFs following post imputation quality control based on either 'info' scores for IMPUTE2 or 'RSQ' scores for MINIMAC3. Imputation accuracy and imputation quality scores are derived from IMPUTE2+1000G or MINIMAC3+1000G imputation.

True Base	Error Base rates				
	G	**T**	**C**	**A**	**Total**
G	-	1/60	1/120	1/120	1/30
T	1/60	-	1/120	1/60	1/24
C	1/120	1/120	-	1/120	1/40
A	1/120	1/60	1/120	-	1/30

Supplementary Table 1. Error rates between specific bases for the simulation of WGS data.

Phasing Software + Options	Mean Switch Error Rate	
	Pedigree	**HapGen+Pedigree**
ALPHAPHASE †	0.0235	0.0218
BEAGLE	0.00490	0.00165
EAGLE1	0.00267	0.000589
EAGLE2	0.00152	0.000321
EAGLE2+1000G	0.00293	0.00155
SHAPEIT2	0.000910	0.000283
SHAPEIT2+duohmm	0.000845	0.000247
SHAPEIT2+duohmm+1000G	0.000638	0.000191
SHAPEIT3	0.000957	0.000279
SLRP †	0.00117	0.000950

† Not all variants were phased.

Supplementary Table 2. Mean global SER across simulation replicates for all phasing strategies considered.

Imputation Software + Reference Panel	Mean Imputation Accuracy									
	Pedigree					HapGen+Pedigree				
MAF	**(0,0.01]**	**(0.01,0.05]**	**(0.05,0.10]**	**(0.10,0.20]**	**(0.20,0.50]**	**(0,0.01]**	**(0.01,0.05]**	**(0.05,0.10]**	**(0.10,0.20]**	**(0.20,0.50]**
BEAGLE+1000G †	0.423	0.605	0.744	0.792	0.832	0.296	0.518	0.658	0.710	0.757
IMPUTE2+1000G †	0.472	0.670	0.833	0.882	0.904	0.299	0.608	0.774	0.836	0.868
IMPUTE4+1000G †	0.524	0.714	0.845	0.890	0.912	0.351	0.633	0.786	0.845	0.877
MINIMAC3+1000G †	0.530	0.722	0.852	0.900	0.916	0.366	0.644	0.793	0.851	0.881
PBWT+1000G (20 replicates) †	0.426	0.640	0.791	0.851	0.883	0.290	0.555	0.724	0.798	0.840
MAF	**(0,0.05]**		**(0.05,0.10]**	**(0.10,0.20]**	**(0.20,0.50]**	**(0,0.05]**		**(0.05,0.10]**	**(0.10,0.20]**	**(0.20,0.50]**
IMPUTE2+1000G ‡	0.749		0.863	0.883	0.907	0.670		0.811	0.834	0.871
IMPUTE2+SSP ‡	0.845		0.921	0.938	0.954	0.916		0.951	0.963	0.974
IMPUTE2+1000G+SSP ‡	0.872		0.933	0.946	0.960	0.914		0.950	0.961	0.973
MINIMAC3+1000G ‡	0.779		0.882	0.900	0.920	0.703		0.831	0.855	0.884
MINIMAC3+HRC ‡	0.844		0.918	0.930	0.942	0.752		0.860	0.879	0.903
MINIMAC3+SSP ‡	0.840		0.917	0.935	0.953	0.909		0.946	0.958	0.971
MINIMAC3+HRC+SSP ‡	0.905		0.951	0.961	0.971	0.918		0.953	0.964	0.974

† Corresponds to a comparison on 40,989 and 40,407 variants on the Pedigree and HapGen+Pedigree simulation strategies respectively.

‡ Corresponds to a comparison on 35,058 and 34,605 variants present in the SSP on the Pedigree and HapGen+Pedigree simulation strategies respectively.

Supplementary Table 3. Mean Imputation accuracy across simulation replicates split by MAF, these results correspond to Figures 3, 4, and 5 in the main text.

MAF	(0,0.01]		(0.01,0.05]		(0.05,0.10]		(0.10,0.20]		(0.20,0.50]	
	Good	Bad	Good	Bad	Good	Bad	Good	Bad	Good	Bad
N	313	595	2232	670	3568	272	7621	210	21682	359
	Variants remaining (%) after threshold was applied									
info										
0.3	96	37	100	80	100	97	100	99	100	100
0.4	94	30	100	69	100	90	100	96	100	99
0.5	91	23	100	54	100	73	100	79	100	82
0.6	89	17	99	37	100	48	100	47	100	43
0.7	80	11	98	20	98	23	99	20	100	13
0.8	68	6	93	8	92	9	95	7	97	3
0.9	46	2	77	2	74	2	80	3	85	1
RSQ										
0.3	85	13	96	37	100	55	100	60	100	65
0.4	79	9	93	23	99	34	100	35	100	33
0.5	71	6	88	14	97	18	99	18	100	10
0.6	62	4	80	7	92	9	96	8	98	3
0.7	51	2	67	4	84	4	90	4	94	1
0.8	37	1	51	2	70	2	77	2	84	1
0.9	20	0	29	0	47	1	53	0	60	0

Supplementary Table 4. Mean number of variants (N) in each MAF bin that were well imputed (Good) or poorly imputed (Bad) as defined by whether Imputation accuracy exceeded 0.5 or fell below 0.2 respectively. The body of the table displays the mean percentage of variants remaining after 'info' or 'RSQ' thresholds have been applied. 'Info' and 'RSQ' scores pertain to IMPUTE2+1000G and MINIMAC3+1000G imputation respectively.

Phasing	Real Time	Computational Time
BEAGLE	0:05:53	0:36:58
SLRP	3:39:20	3:34:38
ALPHAPHASE	0:04:05	0:03:59
EAGLE1	0:04:33	0:16:47
EAGLE2	0:04:11	0:15:27
EAGLE2+1000G	0:15:46	0:44:32
SHAPEIT2	1:00:46	1:00:40
SHAPEIT2+duohmm	1:01:58	1:01:53
SHAPEIT2+duohmm+1000G	1:14:23	1:09:08
SHAPEIT3	0:46:46	0:46:45
Imputation		
BEAGLE+1000G	0:17:54	3:35:49
IMPUTE2+1000G	2:03:24	1:57:39
IMPUTE4+1000G	0:13:55	0:12:56
IMPUTE2+SSP	0:02:49	0:02:42
IMPUTE2+1000G+SSP	5:00:29	4:53:01
MINIMAC3+1000G †	1:07:13	1:05:05
MINIMAC3+SSP †	0:03:30	0:03:20
MINIMAC3+HRC †	7:39:29	7:35:56

† Part of the duration taken by MINIMAC3 was attributed to reformatting the reference panel into a specialised MINIMAC3 format.

Hours:Minutes:Seconds

Supplementary Table 5. Time requirements for phasing ARRAY data on the whole of chromosome 10 for and imputing 20Mb of chromosome 10.

297 **References**

298 *Browning, Brian L., & Browning, Sharon R. (2016). Genotype Imputation with Millions of Reference*
299 *Samples. Am J Hum Genet, 98(1), 116-126. doi: 10.1016/j.ajhg.2015.11.020*
300 *DePristo, M. A., Banks, E., Poplin, R. E., Garimella, K. V., Maguire, J. R., Hartl, C., . . . Daly, M. J.*
301 *(2011). A framework for variation discovery and genotyping using next-generation DNA*
302 *sequencing data. Nat Genet, 43(5), 491-498. doi: 10.1038/ng.806*
303 *Kim, S. Y., Lohmueller, K. E., Albrechtsen, A., Li, Y., Korneliussen, T., Tian, G., . . . Nielsen, R. (2011).*
304 *Estimation of allele frequency and association mapping using next-generation sequencing*
305 *data. BMC Bioinformatics, 12, 231-231. doi: 10.1186/1471-2105-12-231*
306 *Li, Y., Willer, C. J., Ding, J., Scheet, P., & Abecasis, G. R. (2010). MaCH: Using Sequence and Genotype*
307 *Data to Estimate Haplotypes and Unobserved Genotypes. Genet Epidemiol, 34(8), 816-834.*
308 *doi: 10.1002/gepi.20533*
309 *Liu, E. Y., Buyske, S., Aragaki, A. K., Peters, U., Boerwinkle, E., Carlson, C., . . . Li, Y. (2012). Genotype*
310 *Imputation of Metabochip SNPs Using a Study-Specific Reference Panel of ~4,000 Haplotypes*
311 *in African Americans From the Women's Health Initiative. Genet Epidemiol, 36(2), 107-117.*
312 *doi: 10.1002/gepi.21603*
313 *Pistis, G., Porcu, E., Vrieze, S. I., Sidore, C., Steri, M., Danjou, F., . . . Sanna, S. (2015). Rare variant*
314 *genotype imputation with thousands of study-specific whole-genome sequences:*
315 *implications for cost-effective study designs. Eur J Hum Genet, 23(7), 975-983. doi:*
316 *10.1038/ejhg.2014.216*

317

318