Fast and accurate genotype imputation in genome-wide association studies through pre-phasing

Supplementary information

Bryan Howie[1,6], Christian Fuchsberger[2,6], Matthew Stephens[1,3], Jonathan Marchini[4,5], and Gonçalo R. Abecasis[2]

[1]Department of Human Genetics, University of Chicago, Chicago, US
[2]Department of Biostatistics, University of Michigan, Ann Arbor, US
[3]Department of Statistics, University of Chicago, Chicago, US
[4]Department of Statistics, University of Oxford, Oxford, UK
[5]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK
[6]These authors contributed equally to this work

Nature Genetics: doi:10.1038/ng.2354

Supplementary Note

Multiple Imputation as Alternative to Most Likely Haplotype Solution

Both MaCH and IMPUTE2 estimate haplotypes via Markov chain Monte Carlo (MCMC) algorithms that probabilistically sample phased haplotypes for each study individual, conditional on the current haplotype guesses for other individuals in the dataset. While the main text focuses on collapsing the haplotypes sampled by these algorithms into best-guess estimates, it also mentions another form of pre-phasing in which several sampled haplotype configurations are stored for later use. Given these, we can impute from any reference panel into each of an individual's sampled haplotype pairs and average the results to produce probability distributions on missing genotypes. In statistics terminology, this is a form of "multiple imputation." This approach takes longer than single imputation into best-guess haplotypes, but we explore this strategy here to inform investigators who want to maximize accuracy at the cost of extra computing power.

We used the WTCCC2 and GAIN datasets to assess the properties of imputation based on sampled haplotypes. We considered two variables that drive the phasing algorithm: the number of MCMC iterations (parameterized by *-iter* in IMPUTE2 and *-r* in MaCH; each iteration involves sampling a new haplotype pair for every study individual) and the number of non-self haplotypes that are copied when sampling a new pair of haplotypes (parameterized by *–k* in IMPUTE2 and *–s* in MaCH). The cost of phasing grows linearly with the number of iterations and quadratically with the number of copied haplotypes. The cost of imputation grows linearly with the number of imputation steps N, where $N=1$ for imputation into best-guess haplotypes and $N=M$ for imputation into M sets of sampled haplotypes. We

Nature Genetics: doi:10.1038/ng.2354

analyzed the WTCCC2 data with IMPUTE2 runs of 4, 10, 20, or 500 iterations and –*k* values of 40, 80, or 120. (Each run was preceded by 10 burn-in iterations that did not inform the inference.) We analyzed the GAIN data with MaCH runs of 4, 10, 15, or 20 iterations (no burn-in) and *-s* values of 50, 150, or 300. The two methods were assigned different settings because their algorithms and approximations are slightly different.

The results of our experiments are shown in Supplementary Fig. 2 (WTCCC2/IMPUTE2) and Supplementary Fig. 3 (GAIN/minimac), where –*k* and *-s* are plotted on the *x*-axis (respectively) and imputation accuracy is plotted on the *y*-axis. Supplementary Fig. 2A and 3A show the results for SNPs with MAF > 5%, while Supplementary Fig. 2B and 3B show the results for SNPs with MAF < 3%. Each color represents a phasing run that sampled a different number of haplotype configurations per individual. The results from collapsing the haplotypes into best-guess estimates for imputation are shown as solid lines, while the results from imputing into multiple sampled haplotypes and averaging are shown as dashed lines. We also provide benchmark results from IMPUTE1 (Supplementary Fig. 2) and MaCH (Supplementary Fig. 3) as dotted black lines.

The basic trends in Supplementary Fig. 2 and 3 are as expected: imputation accuracy always increases with larger numbers of copied states (–*k* and –*s*) and with larger numbers of sampled haplotypes, regardless of whether those haplotypes were used for best-guess or sample-averaged imputation. In each case, it is clear that using 4-10 sampled haplotype configurations (at the cost of a 4-10x increase in computation time for the imputation step) can outperform analysis using the most likely haplotype pair for each individual. It is also clear that further increases in the number of sampled haplotype configurations (e.g., including up to 500 such configurations in

Nature Genetics: doi:10.1038/ng.2354

Supplementary Fig. 2, which would require 500x more computation time than using the most likely haplotype configuration) provide only modest increases in accuracy.

The mixing of the Markov chains does not appear to be an issue since we obtained the same accuracy from the first *M* sampled haplotypes (*M* = 4,10,20) as from *M* haplotype sets drawn at uniform intervals during 500 sampling iterations (results not shown). The imputation accuracy was consistent across independent runs started from different parts of phasing space (results not shown), which further argues that the algorithms are efficiently exploring the space of GWAS haplotypes.

In general, we anticipate that averaging imputation results across sampled haplotypes will be more valuable in situations where the haplotypes are harder to estimate, such as populations with high genetic diversity. One pragmatic approach might be to use best-guess haplotypes for fast genome-wide imputation, then repeat the imputation with the more intensive sample-averaging approach to refine signals in regions with putative associations. Both MaCH/minimac and IMPUTE2 include functionality to store sampled haplotypes at the pre-phasing step and then average the imputation over these configurations for a given reference panel.

Choice of Methods for Pre-Phasing and Subsequent Imputation

We conducted each pre-phasing analysis in this study within a given software suite: for some datasets we pre-phased with MaCH and then imputed with minimac, while for other datasets we pre-phased and imputed with IMPUTE2. This approach was convenient for our situation, but in principle the pre-phasing and imputation steps could be performed with any

Nature Genetics: doi:10.1038/ng.2354

combination of software packages that can handle the required input and output information.

In general, we recommend using the most accurate method that is computationally feasible for each step. This could be a function of study design (unrelateds, small nuclear families, etc.), study population, and computational resources. The GWAS haplotypes estimated in the pre-phasing step affect all downstream imputation steps, so it is worthwhile to devote substantial computing power to pre-phasing if this will improve accuracy. Recent comparisons suggest that MaCH and IMPUTE2 can produce some of the best phasing results among existing methods for unrelated individuals[1,2], although newer methods like SHAPEIT[2] may achieve better accuracy with less computation.

To illustrate the benefits of using our approach together with alternative phasing methods, we phased the WTCCC2 scaffold genotypes (Affymetrix 500k SNPs) with SHAPEIT v1.r416 on the following settings: --states-phase=600, --burn=10, --prune=10, --main=50, --effective-size=11500. We then passed the estimated haplotypes to IMPUTE2 for imputation from the 1,000 Genomes EUR (Nov 2010) reference panel. Relative to the WTCCC2 results in Table 2 (which are based on using IMPUTE2 for both the pre-phasing and imputation), this analysis yielded a small improvement at SNPs with MAF>5% (~0.005 on the R^2 scale) and a larger improvement at SNPs with MAF<3% (~0.014 on the R^2 scale). These results highlight the importance of accuracy in the pre-phasing step and the potential to improve imputation accuracy through continuing development in phasing algorithms.

Nature Genetics: doi:10.1038/ng.2354

Imputation Results from Different Combinations of Methods and Datasets

In the main text, we illustrated the strengths of pre-phasing through different combinations of imputation methods and cross-validation datasets. We conducted this work collaboratively across multiple institutions, and these combinations were largely determined by which datasets were convenient for each group to analyze. This approach was informed by our understanding that the MaCH and IMPUTE software families share deep mechanistic similarities, such that we do not expect substantial differences in behavior. Nonetheless, we here present results to confirm that both methods produce very similar trends in accuracy and running time when applied to the same dataset (Supplementary Fig. 4; main text Table 1 vs. Supplementary Table 1).

REFERENCES

1. Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* **34**, 816-34 (2010).
2. Delaneau, O., Marchini, J. & Zagury, J.F. A linear complexity phasing method for thousands of genomes. *Nature methods* **9**, 179-81 (2012).

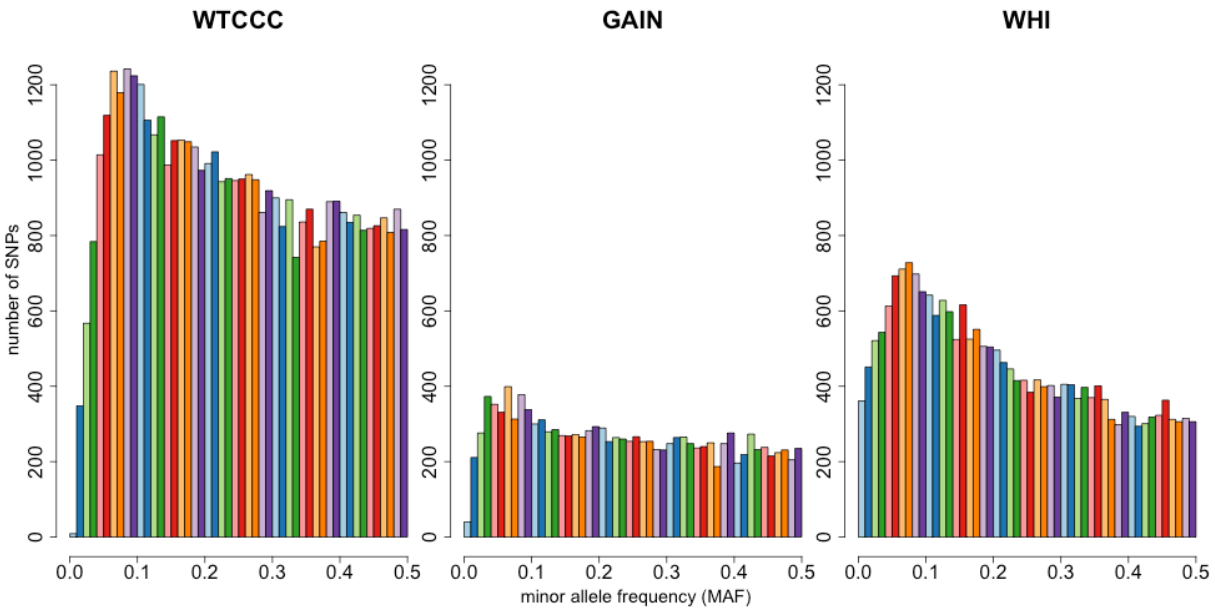
Nature Genetics: doi:10.1038/ng.2354

Supplementary Table 1: Running times in GAIN data (in CPU minutes needed to impute one individual genome-wide) for different imputation methods and reference panels.

	Imputation method	
Reference Panel[a]	**MaCH**	**minimac (pre-phasing[b])**
HapMap 2 CEU (60 indiv, 2.5M SNPs)	12	<1
1000G CEU (60 indiv, 7.3M SNPs)	45	1
1000G EUR (283 indiv, 11.6M SNPs)	1,261	7
1000G EUR (381 indiv, 37.4M SNPs)	7,369[c]	23

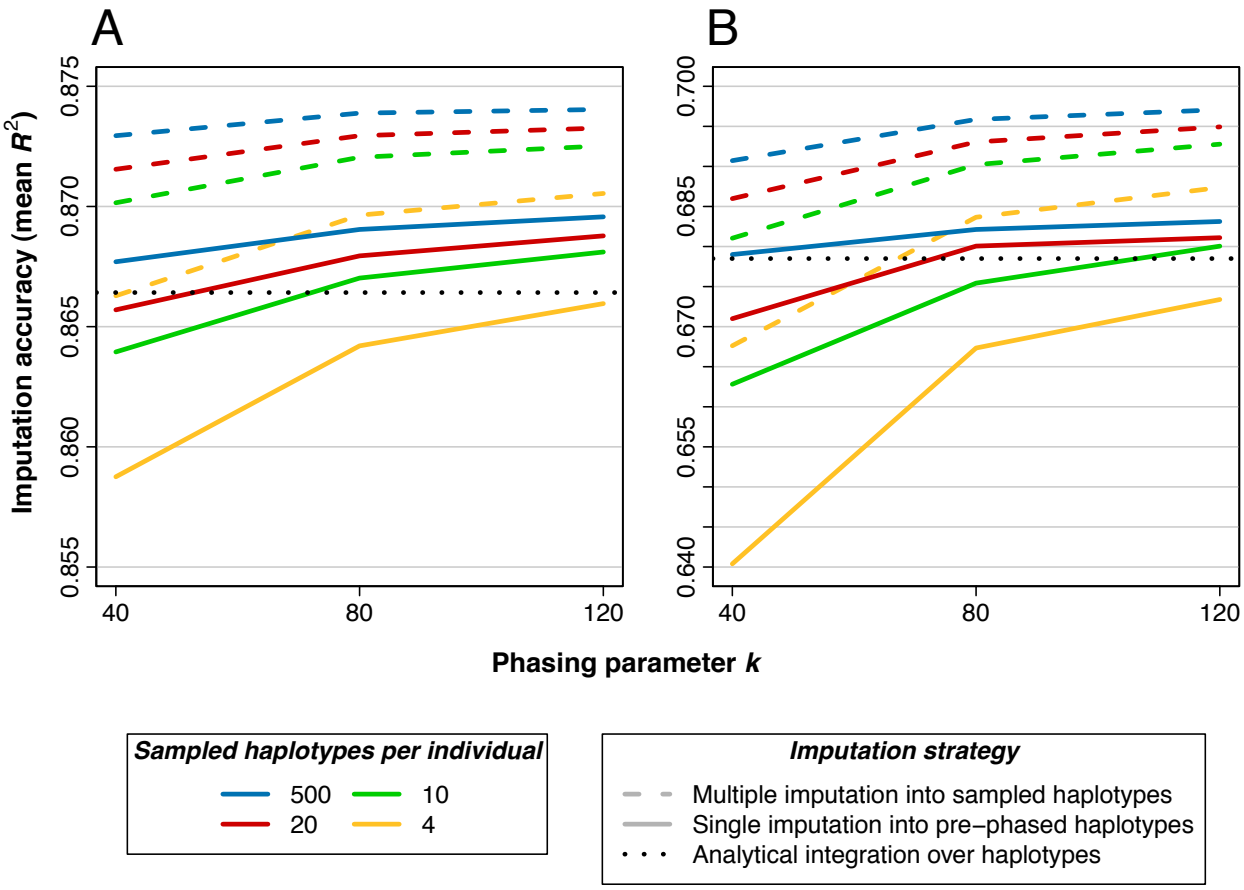
[a]Reference panels, in order: HapMap 2 release #22; 1000 Genomes low-coverage pilot (June 2010); 1000 Genomes interim release (Aug 2010); 1000 Genomes interim Phase I release (Nov 2010).
[b]Running times do not include the initial investment required to phase the GWAS genotypes, which took 25 minutes per individual.
[c]Projected running time extrapolated from existing benchmarks.

Nature Genetics: doi:10.1038/ng.2354



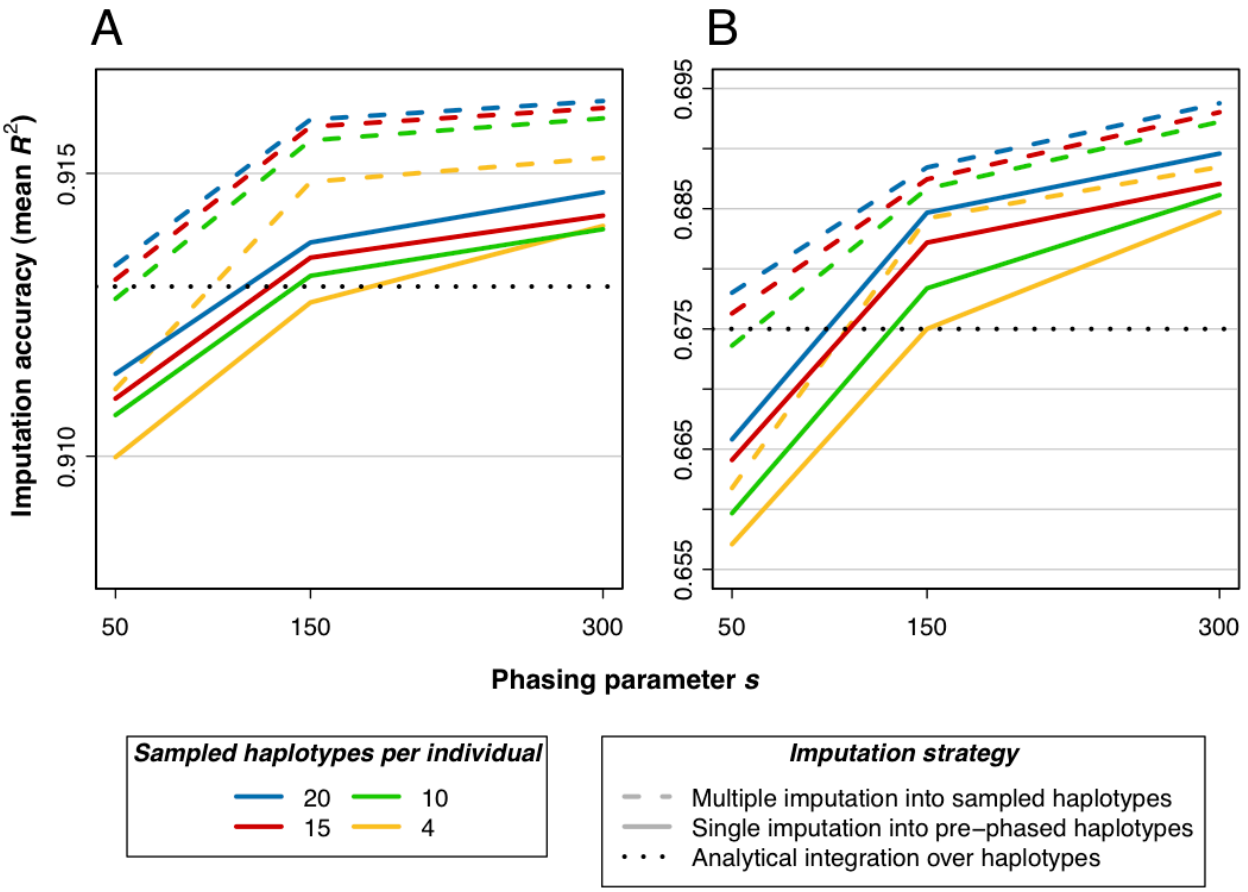
Supplementary Figure 1: Histograms of minor allele frequency on cross-validation SNP chips. The *x*-axis shows 1% frequency bins: (0,0.01], (0.01,0.02],..., (0.49,0.50]. The *y*-axis shows the number of SNPs in each bin that were used to estimate cross-validation accuracy. The WTCCC comparison included SNPs from the Affymetrix 6.0 and Illumina 1M arrays on chromosome 10, minus a set of scaffold SNPs from the Affymetrix 500k array. The GAIN comparison included SNPs from the Affymetrix 6.0 array on chromosome 20, minus a set of scaffold SNPs from a custom Perlegen array. The WHI comparison included SNPs from the Affymetrix 6.0 array on chromosome 20; all SNPs were imputed through a mask-every-10th-SNP sliding window analysis.

Nature Genetics: doi:10.1038/ng.2354



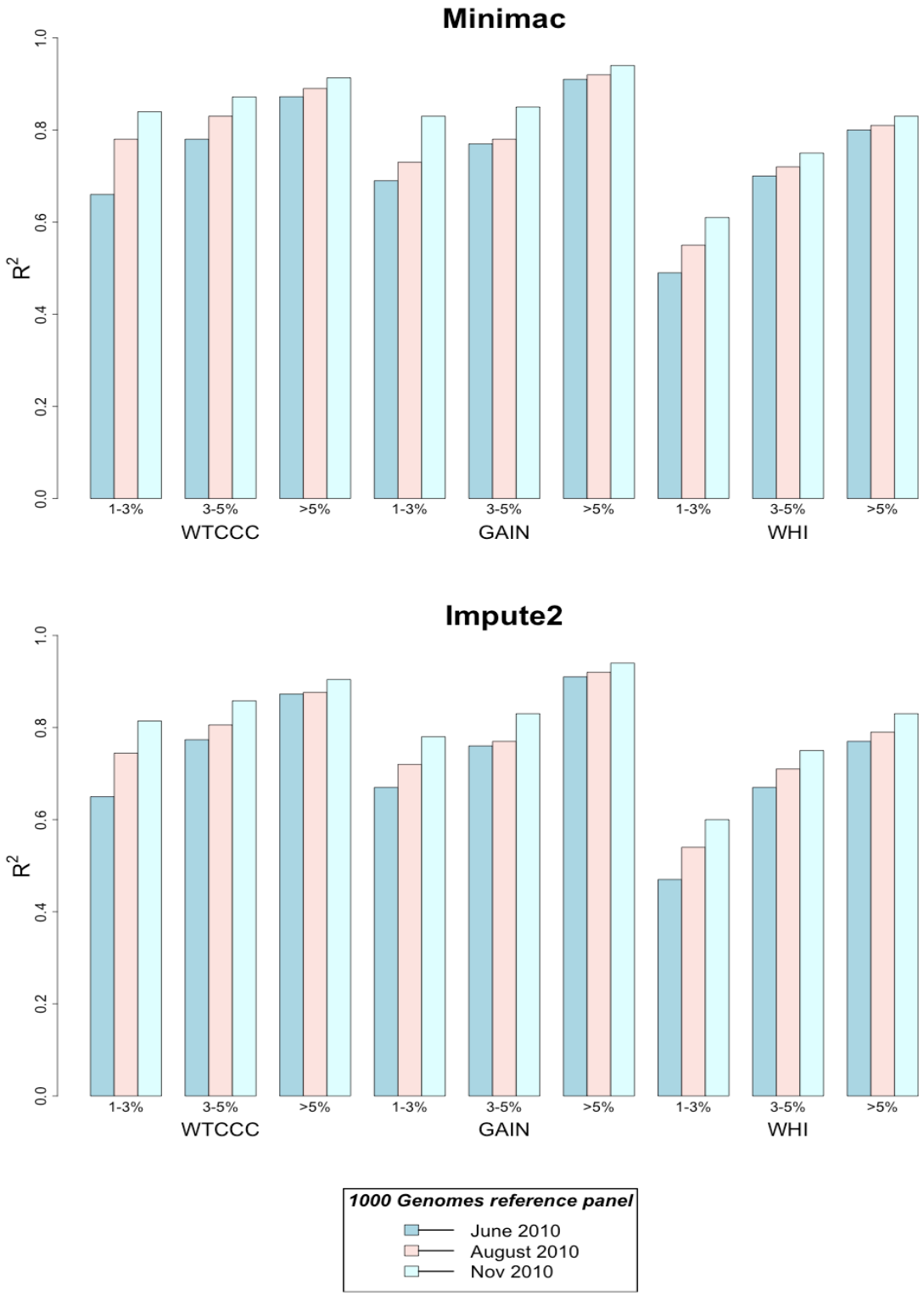
Supplementary Figure 2: Accuracy of imputation from 1000 Genomes CEU panel into WTCCC2 data as a function of phasing parameter *-k*, imputation strategy, and number of haplotypes sampled per individual by IMPUTE2. The analysis used a pseudo-GWAS scaffold of Affymetrix 500k SNPs; all other SNPs in the WTCCC2 dataset were masked and imputed (this analysis was restricted to the first half of chromosome 10 to facilitate parameter exploration). The dashed and solid curves show results based on IMPUTE2's iterative phasing algorithm, which samples a new pair of scaffold haplotypes for every GWAS individual at each iteration. The phasing algorithm was run for 4, 10, 20, or 500 iterations (orange, green, red, and blue lines, respectively). The sampled haplotypes were then used for multiple imputation (dashed curves) or collapsed into best-guess estimates and used for single imputation (solid curves). As a benchmark, we also ran a method that analytically integrates over GWAS haplotype configurations (IMPUTE, dotted black line; results do not depend on *k*). *(A)* Imputed SNPs with MAF > 5% in WTCCC2 data. *(B)* Imputed SNPs with MAF < 3% in WTCCC2 data.

Nature Genetics: doi:10.1038/ng.2354



Supplementary Figure 3: Accuracy of imputation from 1000 Genomes CEU panel into GAIN data as a function of phasing parameter *-s*, imputation strategy, and number of haplotypes sampled per individual by MaCH. We evaluated imputation accuracy by examining the correlation between imputed dosages and array genotypes for markers that were present on the Affymetrix 6.0 arrays but not on the Perlegen custom array (this analysis was restricted to chromosome 20 to facilitate parameter exploration). The dashed and solid curves show results based on MaCH's iterative phasing algorithm, which samples a new pair of scaffold haplotypes for every GWAS individual at each iteration. The phasing algorithm was run for 4, 10, 15, or 20 iterations (orange, green, red, and blue lines, respectively). The sampled haplotypes were then used for multiple imputation (dashed curves) or collapsed into best-guess estimates and used for single imputation (solid curves). As a benchmark, we also ran a method that analytically integrates over GWAS haplotype configurations (MaCH, dotted black line; results do not depend on *s*). *(A)* Imputed SNPs with MAF > 5% in GAIN data. *(B)* Imputed SNPs with MAF < 3% in GAIN data.

Nature Genetics: doi:10.1038/ng.2354



Supplementary Figure 4: Accuracy of pre-phasing imputation with minimac and IMPUTE2 when used to impute genotypes from three different 1000 Genomes reference panels into GWAS data from WTCCC2, GAIN, and WHI.

Nature Genetics: doi:10.1038/ng.2354