Supplementary Data

Luke Jostins, Katherine I. Morley and Jeffrey C. Barrett

1 HapMap Release

For the various reference sets, we used the HapMap2 and HapMap 3 datasets; these datasets have 2545519 and 1375022 polymorhic SNPs, respectively. The datasets were phased by the HapMap Consortium using IMPUTE v2. We used the CEU population from the phased r2 release of the HapMap2 data, and various mixtures of populations from the phased r21 release of the HapMap3 data. The data can be downloaded from the HapMap FTP

```
ftp://ftp.ncbi.nlm.nih.gov/hapmap/phasing
```

2 Performing Imputation

We used the imputation program Beagle (version 3.0.2) for the majority of our imputation tests. Beagle was designed to handle large reference sets; it performs Hidden Markov Model (HMM)-based imputation using groups of haplotypes (haplotype clusters) as states; these clusters can be dynamically merged and split as you move along the chromosome, allowing a large number of samples to be processed without unduely inflating compute time or memory. This makes Beagle well suited to our purposes, as it should be able to extract the useful information from large mixtures of HapMap3 populations, without suffering an overinflation of resource requirements. We used the default settings ($N_{iterations} = 10$, $N_{samples} = 4$).

In order to investigate how the results generalise across imputation methods, we also performed some imputation using the IMPUTE programs, IMPUTE v1 (version 1.0.0) and IMPUTE v2 (version 2.1.0). IMPUTE v1, like Beagle, uses a HMM method, where each reference haplotype is a state, using a recombination map to estimation transition probabilities between states; genotypes for target samples are integrated over their possible haplotypes to give genotype posteriors. IMPUTE v2 uses a similar model to IMPUTE v1, but uses a Markov Chain Monte Carlo method to sample possible haplotype assignments for target individuals, and uses these haplotypes to infer genotypes; in order to cut down on computation, each target sample constructs its haplotypes only from the k nearest known haplotypes (as defined by the Hamming distance), where k is a parameter of the method. For both versions of IMPUTE we used the suggested value of $N_e = 11418$ for the effective population size, and the HapMap combined genetic map. For the IMPUTE v2 MCMC algorithm we used the default value of $k = 50$ and $k_{hap} = 500$, with 30 iterations and a burn-in of 10 iterations.

We also attempted to use the imputation software MACH, but found that the memory and CPU requirements were too high to realistically carry out the tasks we required; in particular, imputation using reference sizes larger than around 200 individuals took over 12 CPU days to perform.

All imputation was carried out using the default or recommended parameter values. For all imputation programs (Beagle, Impute1, Impute2 and MACH) we split up the target SNPs into segments, 5MB long, with a 1MB buffer region; we tested various sizes segment and the buffer region, and found little change in imputation power.

Imputation was performed on the Wellcome Trust Sanger Institutes compute farm. Most computation was carried out using single cores of 2x3.0 GHz quad core Xeon EMT64 processors;

the exception was for the large memory runs of IMPUTE v1, which were run on large memory machines.

Reference Set	Calibration		Quality r^2	
	Common	Rare	Common	Rare
HM2CEU	0.019	0.038	0.78	0.73
CEU	0.008	0.027	0.88	0.76
CEU+TSI	0.002	0.009	0.92	0.79
CEU+TSI+GIH+MEX	-0.006	-0.019	0.93	0.79
WORLD	-0.010	-0.043	0.91	0.76

Table 1: Calibration data for Genome-Wide imputation using our five reference sets. Quality calibration is defined as the mean difference between the actual and predicted dosage r^2 across unfiltered SNPs; a negative value represents conservative quality scores, and a positive value represents liberal quality scores. The quality r^2 is the correlation between the predicted and actual r^2. The SNPs are split into common (MAF > 0.05) and rare (MAF ≤ 0.05).

Population	Accuracy MAF >5%	Accuracy MAF >5%	Calibration	Quality r^2 (Gb)	Memory	CPU Hours
Beagle						
CEU	0.948	0.791	0.006	0.856	16.1	7.6
CEU+TSI	0.954	0.827	-0.003	0.888	16.5	9.0
CEU+TSI+GIH+MEX	0.957	0.862	-0.014	0.914	16.7	12.1
WORLD	0.954	0.873	-0.012	0.891	16.7	34.4
Impute v1						
CEU	0.954	0.871	0.007	0.920	12.0	24.5
TSI	0.957	0.0.884	-0.010	0.937	45.0	56.7
CEU+TSI+GIH+MEX	0.957	0.886	-0.012	0.952	77.3	154.6
WORLD			NA	NA	1237.1[a]	NA
IMPUTE v2						
CEU	0.948	0.832	-0.066	0.881	2.9	4.0
CEU+TSI	0.950	0.841	-0.066	0.890	3.0	5.1
CEU+TSI+GIH+MEX	0.949	0.840	-0.066	0.887	3.2	5.8
WORLD	0.946	0.833	-0.065	0.880	4.1	12.2

Table 2: Unfiltered mean dosage r^2, quality calibration and r^2 and resource use information for various imputation methods and reference sets, on Chromosome 17. We used version 3.0.2 of Beagle, version 1.0.0 of IMPUTE v1 and version 2.1.0 of IMPUTE v2. All computation was carried out using single cores of 2x3.0 GHz quad core Xeon EMT64 processors, with the exception of the large memory runs of IMPUTE v1, which were run on large memory machines. In general, IMPUTE v2 was fastest and required the least memory, whereas IMPUTE v1 was slower and took a large amount of memory. Beagle was intermediate in both resource uses, though closer to IMPUTE v2. [a] We were unable to use the WORLD set with IMPUTE v1, as the program required over a terabyte of memory (as estimated by the software).

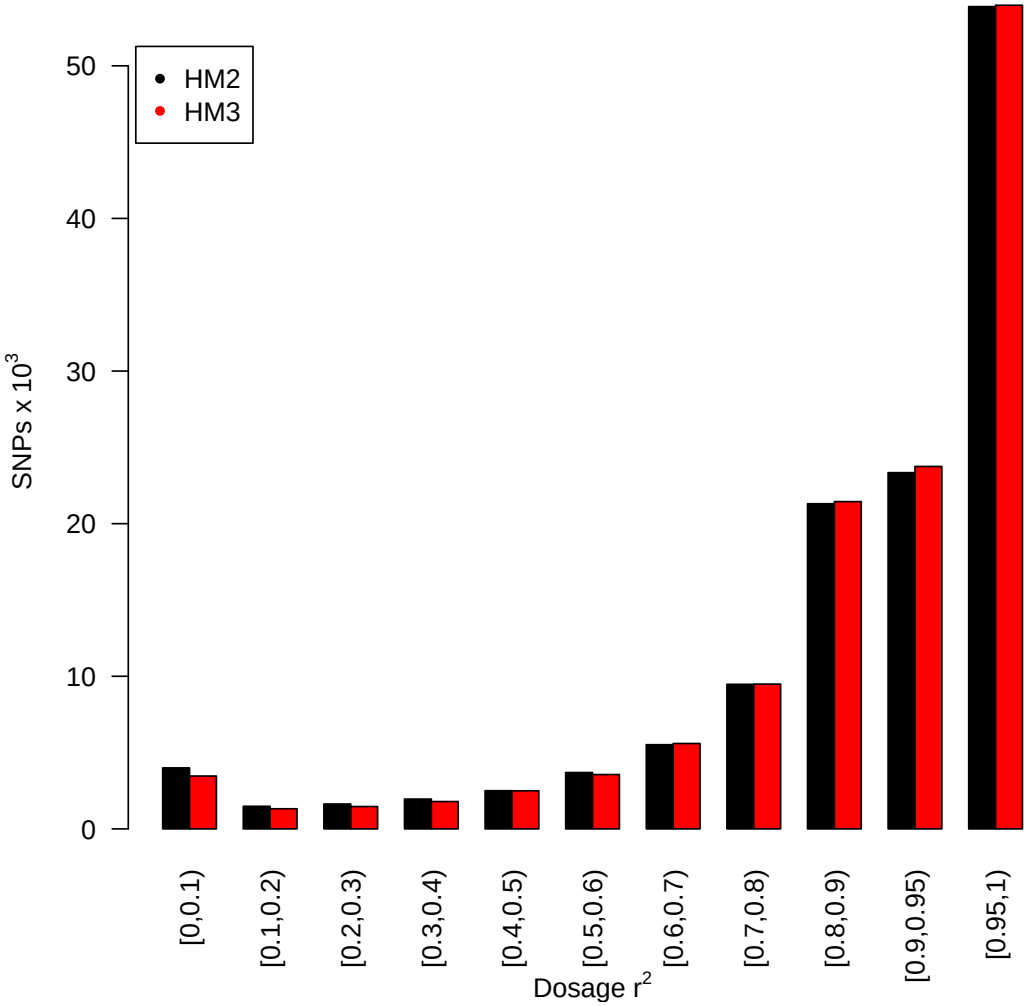
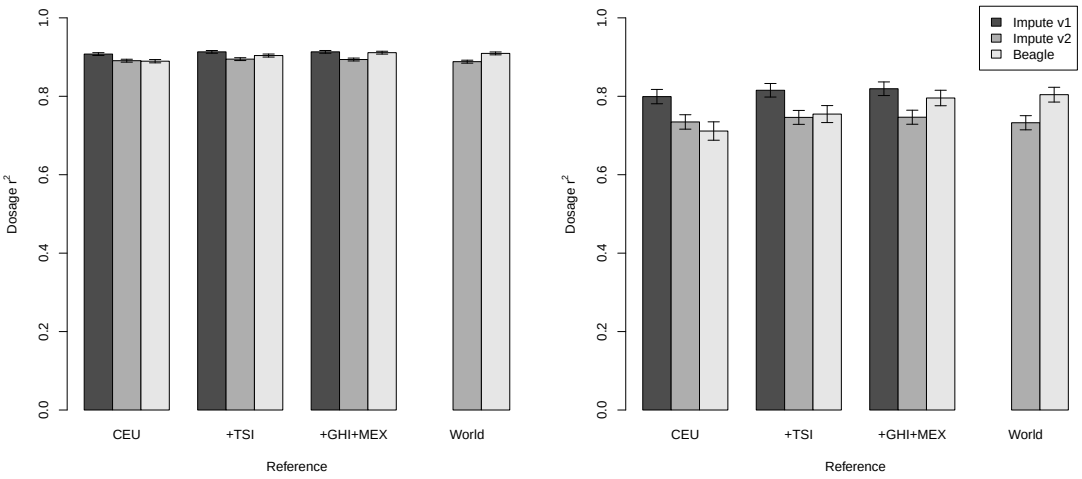


Figure 1: A histogram of dosage r^2 across unfiltered SNPs for a genome-wide imputation using the reduced HapMap2 and HapMap3 sets, which contain only the 1,069,264 SNPs and 56 CEU samples that both HapMap2 and HapMap3 have genotype information for. The means of the distributions are 0.841 and 0.845, and the difference is significant (t = 7.59, df = 256480, p $<10^{-13}$). The dosage r^2 is defined as the square of the Pearson correlation coefficient between the imputed and the actual allele dosage across all imputed samples. The actual dosage is the count of minor alleles for each sample, and the imputed dosage is the expected minor allele count, defined as $2P(aa) + P(Aa)$, where a is the minor allele, and $P(G)$ is the posterior probability of a particular genotype.



(a) Common

(b) Rare

Figure 2: The mean dosage r^2 for the various imputation methods, with different reference sets, for both common (MAF $> 5\%$) and rare (MAF $\leq 5\%$) SNPs. Imputation was performed across chromosome 17. The error bars are 95% confidence intervals. For common variants, all programs performed similarly, with the exception of the WORLD reference set analyses, for which Beagle performs significantly better. For rare variants, IMPUTE v1 and Beagle behave similarly, with an increase in mean dosage r^2 as the reference population increases in size and diversity, though this trend is less pronounced for IMPUTE v1. IMPUTE v2 shows a small improvement for the TSI samples, but the overall imputation accuracy decreases with larger reference sets, especially relative to the performance of Beagle.