The American Journal of Human Genetics

Supplemental Data

Genotype Imputation with Millions of Reference Samples

Brian L. Browning and Sharon R. Browning

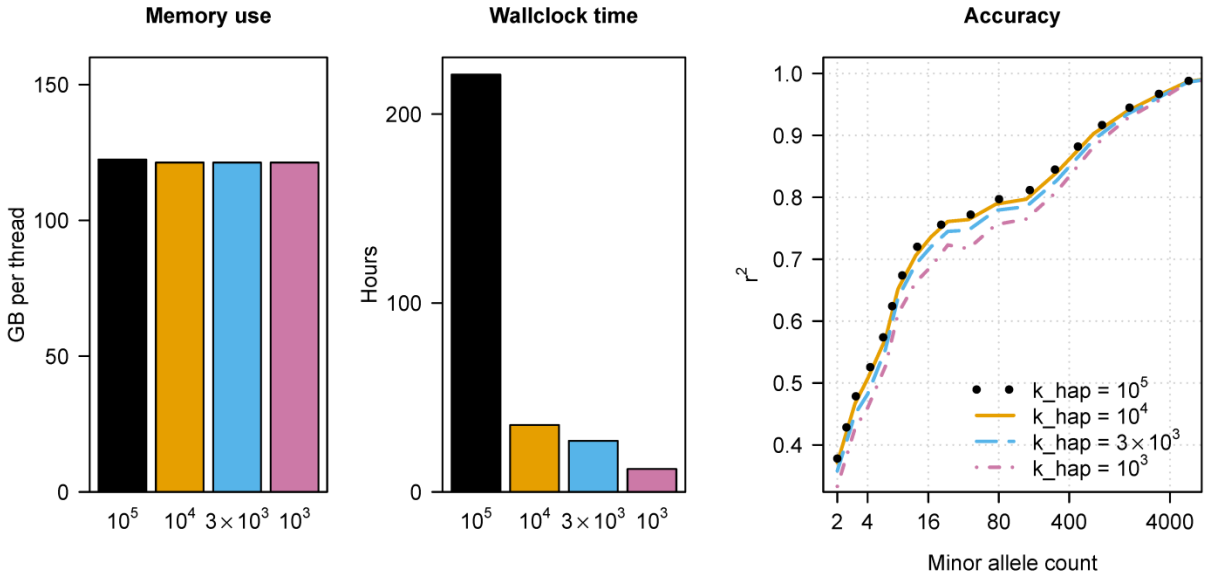

Figure S1. Impute2's performance as the k_hap parameter is varied

Memory use, wallclock time, and genotype imputation accuracy when using Impute2 to impute 10 Mb from a simulated reference panel with 50,000 individuals and 382,425 markers. The simulated imputation target was 1000 individuals genotyped on a 1M SNP array (3,333 markers in the 10 Mb region). The 10 Mb region was broken into six 1.67 Mb regions with a 250 kb buffer appended to each end of each region. Memory use is the maximal memory use for the six runs. Wallclock time is the sum of the wallclock times for each region. Impute2 was run with four different values of the k_hap parameter (10^5, 10^4, 3 x 10^3, and 10^3). Imputed genotypes were binned according to the minor allele count of the marker in the reference panel. The squared correlation between the imputed minor allele dose and the true minor allele dose is reported for each minor allele count bin. The horizontal axis in the genotype imputation accuracy plot is on a log scale.

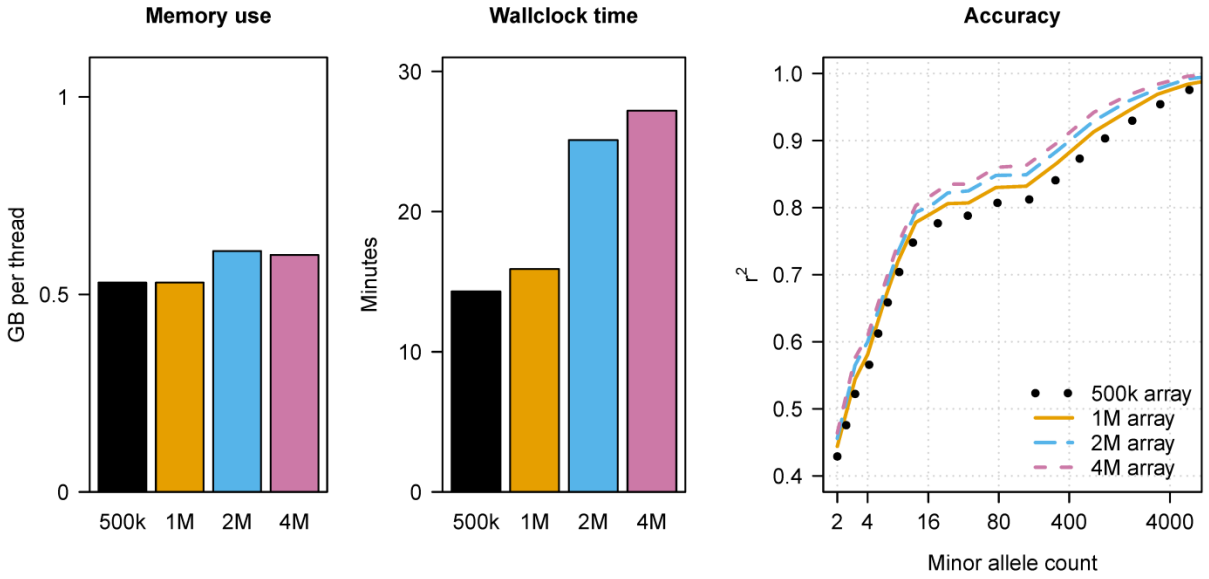

Figure S2. Beagle 4.1's performance as the number of genotyped markers is varied
Memory use, wallclock time, and genotype imputation accuracy when using Beagle 4.1 to impute 10 Mb from a simulated reference panel with 50,000 individuals and 382,425 variants. The simulated imputation target was 1000 individuals. Beagle 4.1 was run with four different genotyped marker densities in the imputation target: 1667, 3333, 6666, and 13,332 markers in the 10 Mb region, which correspond to genome-wide arrays with 500k, 1M, 2M, and 4M SNPs respectively. All Beagle runs used 12 computational threads. Imputed genotypes were binned according to the minor allele count of the marker in the reference panel. The squared correlation between the imputed minor allele dose and the true minor allele dose is reported for each minor allele count bin. The horizontal axis in the genotype imputation accuracy plot is on a log scale.

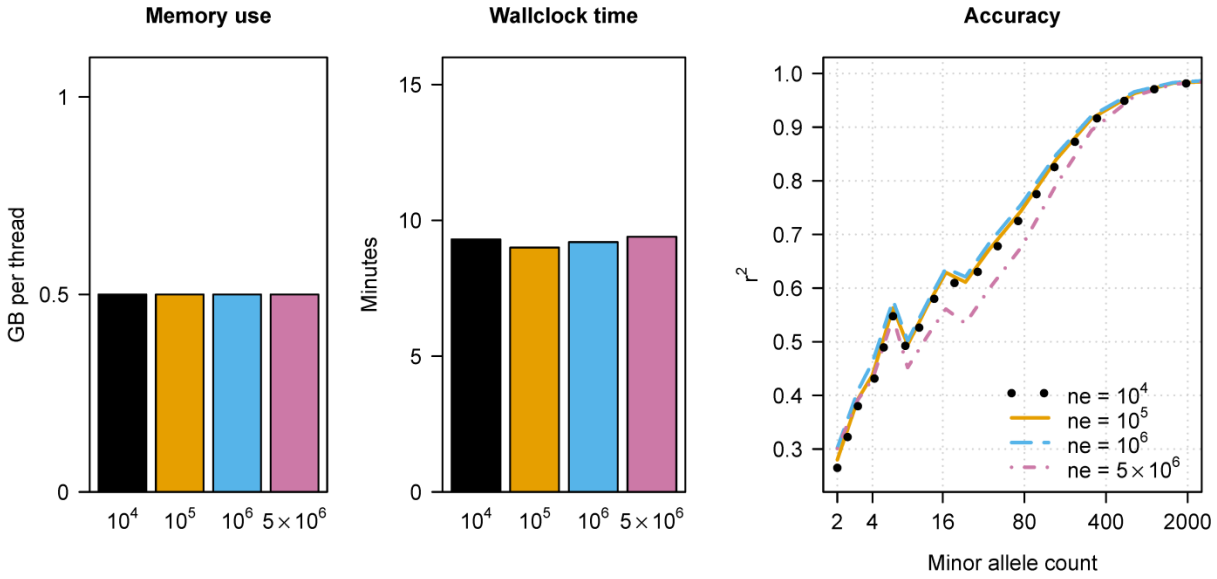

Figure S3. Beagle 4.1's performance as the ne parameter is varied

Memory use, wallclock time, and genotype imputation accuracy when using Beagle 4.1 to impute chromosome 20 variants from a UK10K reference panel (n = 3781). The imputation target was the 1000 Genomes Project European samples. Genotypes that were not on the Illumina Omni2.5 array were masked and imputed in the target samples. Beagle 4.1 was run with four different values of the ne parameter (10^4, 10^5, 10^6, and 5 x 10^6). All Beagle runs used 12 computational threads. Imputed genotypes were binned according to the minor allele count of the marker in the reference panel. The squared correlation between the imputed minor allele dose and the true minor allele dose is reported for each minor allele count bin. The horizontal axis in the genotype imputation accuracy plot is on a log scale.

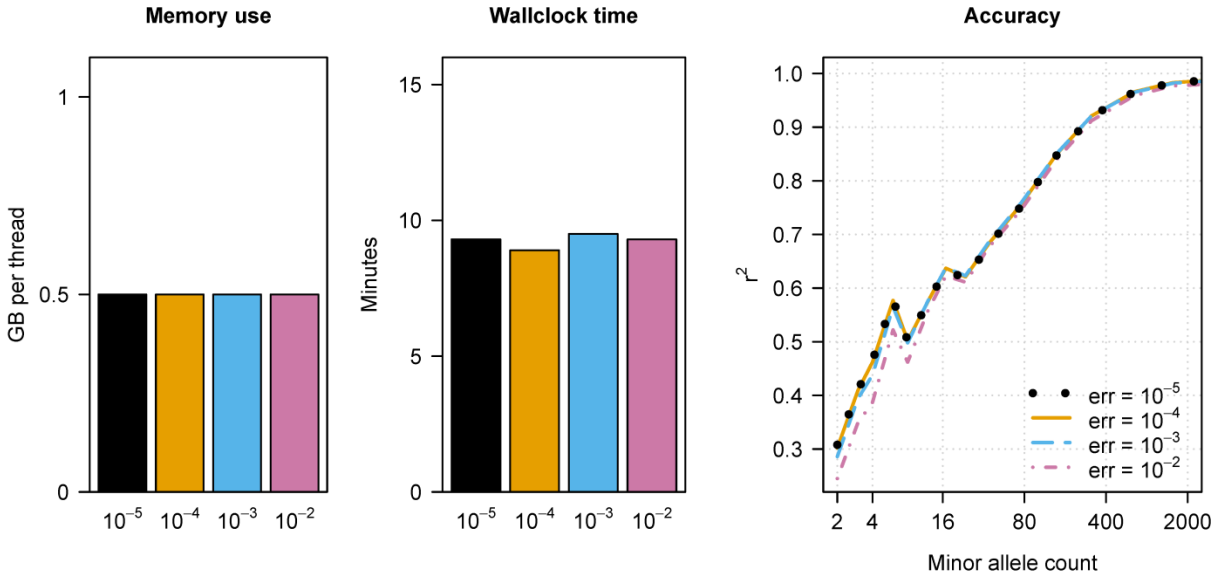

Figure S4. Beagle 4.1's performance as the err parameter is varied

Memory use, wallclock time, and genotype imputation accuracy when using Beagle 4.1 to impute chromosome 20 variants from a UK10K reference panel (n = 3781). The imputation target was the 1000 Genomes Project European samples. Genotypes that were not on the Illumina Omni2.5 array were masked and imputed in the target samples. Beagle 4.1 was run with four different values of the err parameter (0.00001, 0.0001, 0.001, and 0.01). All Beagle runs used 12 computational threads. Imputed genotypes were binned according to the minor allele count of the marker in the reference panel. The squared correlation between the imputed minor allele dose and the true minor allele dose is reported for each minor allele count bin. The horizontal axis in the genotype imputation accuracy plot is on a log scale.

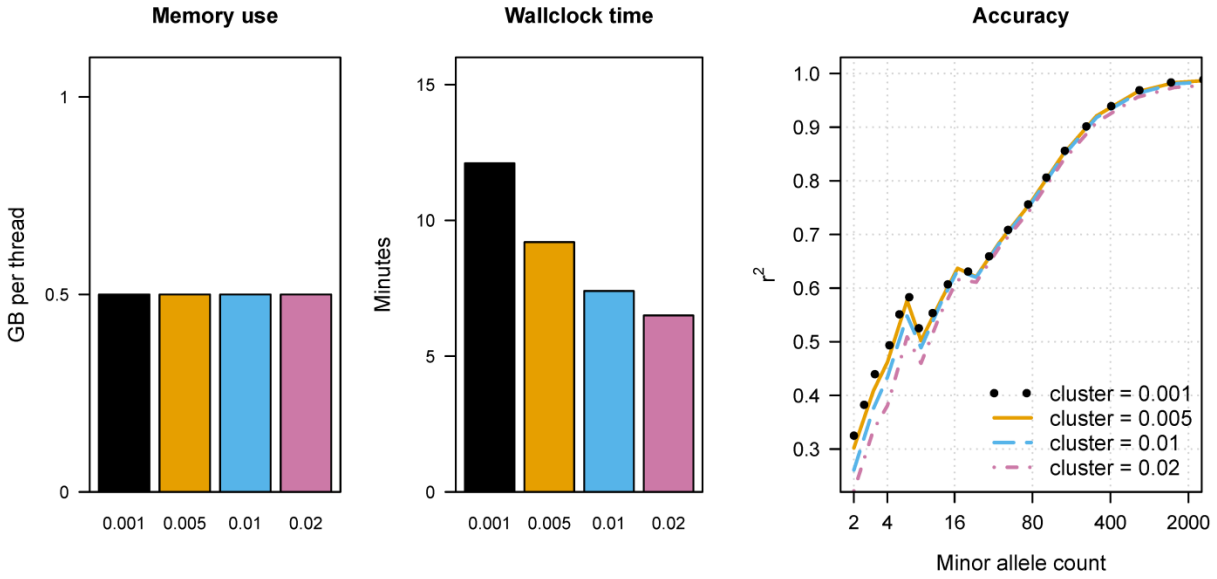

Figure S5. Beagle 4.1's performance as the cluster parameter is varied

Memory use, wallclock time, and genotype imputation accuracy when using Beagle 4.1 to impute chromosome 20 variants from a UK10K reference panel (n = 3781). The imputation target was the 1000 Genomes Project European samples. Genotypes that were not on the Illumina Omni2.5 array were masked and imputed in the target samples. Beagle 4.1 was run with four different values of the cluster parameter (0.005, 0.001, 0.1, and 0.02). All Beagle runs used 12 computational threads. Imputed genotypes were binned according to the minor allele count of the marker in the reference panel. The squared correlation between the imputed minor allele dose and the true minor allele dose is reported for each minor allele count bin. The horizontal axis in the genotype imputation accuracy plot is on a log scale.

Method	Reference format	Memory per thread (GB)	Wallclock time (min)	CPU time (min)
Beagle 4.1	vcf	0.5	2.1	8.9
Beagle 4.1	bref	0.5	1.9	5.9
Minimac3	vcf	2.6	112.7	112.4
Minimac3	m3vcf	2.6	5.0	5.0
Impute2	Impute	74.0	92.8	92.5

Table S1: Memory use and computation time for 1000 Genomes Project reference samples
Memory use and computation time for Beagle 4.1, Minimac3, and Impute2 when imputing chromosome 20 markers from a 1000 Genomes Project reference panel with 957,209 markers. The 1000 Genomes Project data for chromosome 20 were divided into a reference panel with 2452 sequenced individuals and an imputation target with 52 individuals genotyped on the Illumina Omni2.5 array and having all other sequenced variants masked. CPU time includes the sum of the computation time consumed by each computational thread. The Beagle analyses used 12 computational threads.

Method	Reference format	Memory per thread (GB)	Wallclock time (min)	CPU time (min)
Beagle 4.1	vcf	0.5	10.4	79.6
Beagle 4.1	bref	0.5	8.7	63.8
Minimac3	vcf	1.9	111.8	111.4
Minimac3	m3vcf	1.9	26.6	26.5
Impute2	Impute	54.4	533.5	531.9

Table S2: Memory use and computation time for UK10K Project reference samples
Memory use and computation time for Beagle 4.1, Minimac3, and Impute2 when imputing chromosome 20 markers from a UK10K Project reference panel with 406,878 markers into the 503 designated European samples from 1000 Genomes Project. The target samples were genotyped on the Illumina Omni2.5 array and had all other sequenced variants masked. CPU time includes the sum of the computation time consumed by each computational thread. The Beagle analyses used 12 computational threads.

Method	Reference format	Memory per thread (GB)	Wallclock time (min)	CPU time (min)
Beagle 4.1	vcf	0.5	15.9	96.7
Beagle 4.1	bref	0.6	12.3	76.8
Minimac3	vcf	6.0	752.4	750.4
Minimac3	m3vcf	3.4	108.6	108.2
Impute2 k_hap=100k	Impute	122.4	13251.9	13197.1
Impute2 k_hap=10k	Impute	121.3	2123.8	2101.5
Impute2 k_hap=3k	Impute	121.3	1624.2	1599.7
Impute2 k_hap=1k	Impute	121.3	728.9	711.5

Table S3: Memory use and computation time for 50k simulated reference samples
Memory use and computation time for Beagle 4.1, Minimac3, and Impute2 when imputing 10 Mb of simulated sequence data from 50,000 reference samples with 382,425 markers. The simulated imputation target was 1000 individuals genotyped on a 1M SNP array (3,333 markers in the 10 Mb region). CPU time includes the sum of the computation time consumed by each computational thread. The Beagle analyses used 12 computational threads.

Method	Reference format	Memory per thread (GB)	Wallclock time (min)	CPU time (min)
Beagle 4.1	vcf	0.7	34.7	214.6
Beagle 4.1	bref	0.7	22.2	142.7
Minimac3	vcf	12.0	2470.8	2463.9
Minimac3	m3vcf	7.2	259.0	258.1

Table S4: Memory use and computation time for 100k simulated reference samples
Memory use and computation time for Beagle 4.1 and Minimac3 when imputing 10 Mb of simulated sequence data from 100,000 reference samples with 650,561 markers. The simulated imputation target was 1000 individuals genotyped on a 1M SNP array (3,333 markers in the 10 Mb region). CPU time includes the sum of the computation time consumed by each computational thread. The Beagle analyses used 12 computational threads.

Method	Reference format	Memory per thread (GB)	Wallclock time (min)	CPU time (min)
Beagle 4.1	vcf	0.9	83.1	498.7
Beagle 4.1	bref	0.9	40.3	289.0
Minimac3	vcf	25.1	7523.6	7503.3
Minimac3	m3vcf	16.5	551.5	549.6

Table S5: Memory use and computation time for 200k simulated reference samples
Memory use and computation time for Beagle 4.1 and Minimac3 when imputing 10 Mb of simulated sequence data from 200,000 reference samples with 1,059,310 markers. The simulated imputation target was 1000 individuals genotyped on a 1M SNP array (3,333 markers in the 10 Mb region). CPU time includes the sum of the computation time consumed by each computational thread. The Beagle analyses used 12 computational threads.