Supplementary Table 1 Imputation accuracy. Results are based on simulated haplotypes under a coalescent model consistent with European haplotype diversity, and imputed into GWAS data.

Reference panel sample size	Imputation accuracy [mean r^2]			Probability an individual judged by imputation to carry the rare allele actually does carry that allele		
	Minor allele frequency bin					
	0.01-.1%	0.1-1%	>1%	0.01-0.1%	0.1-1%	>1%
1,000	0.41	0.64	0.96	0.57	0.66	0.95
2,000	0.49	0.71	0.97	0.64	0.72	0.96
5,000	0.59	0.79	0.98	0.71	0.79	0.97
10,000	0.69	0.84	0.98	0.78	0.85	0.98
20,000	0.79	0.89	0.99	0.84	0.89	0.98

Supplementary Table 2 Run time of minimac3 and Beagle4.1 for different reference panels and number of threads to impute 100 whole genomes (run time interpolated from analysis on chromosome 20). Both tools were run on 5 Mb chunks with 1 Mb overlap (13 chunks in serial or chromosome 20 yielding a total of 227,925 variants). Minimac3 and Beagle4.1 were run with their input file formats (m3vcf and bref, respectively).

		Time (in CPU-hours – wall clock)	
Reference panel	# Samples	minimac3 (6 threads)	Beagle 4.1 (6 threads)
1000 Genomes Phase 1	1,092	4 (1.0)	5 (2.4)
AMD	2,074	9 (1.7)	9 (3.3)
1000 Genomes Phase 3	2,504	6 (1.4)	9 (3.1)
Sardinia	3,489	7 (1.5)	11 (3.5)
Combined	9,341	17 (4.6)	31 (8.5)
Mega	11,845	21 (6.2)	40 (10.7)
HRC v1.1	32,390	31 (12.13)	128 (25.8)

Supplementary Table 3 Imputation in diverse populations.

1000 Genomes super population	Number of genomic segments	Average No of variants	Average (S.D.) number of unique haplotypes	Average time (in seconds) to impute single sample	Average time (in seconds) to impute whole genome
African American	49,584	36	14 (2.6)	11.65	582.5
Hispanic	43,522	41	13 (2.5)	10.83	541.5
South Asian	40,031	44	12 (2.4)	10.42	521.0
East Asian	35,822	49	10 (2.3)	9.91	495.5
European	35,367	50	10 (2.4)	9.74	487.0

Nature Genetics: doi:10.1038/ng.3656

Supplementary Table 4 Comparison of the VCF and the M3VCF file format for different reference panels.

Reference panel	Number samples	Number markers	Unzipped format [GB]			Zipped format [GB]			Time taken using 5 threads [minutes]
			VCF	m3vcf	% saving	VCF	m3vcf	% saving	Compression / Parameter estimation
1000 Genomes Phase 1	1,092	617,694	2.6	0.17	93.65	0.08	0.03	61.90	5 / 19
Sardinia	3,489	331,760	1.9	0.10	94.58	0.07	0.02	70.83	11 / 14
1000 Genomes Phase 3	2,004	1,047,613	9.9	0.34	96.56	0.22	0.06	70.73	30 / 66
HRC v1.1	32,390	885,404	108	1.8	98.33	1.8	0.28	84.44	246 / 432

Nature Genetics: doi:10.1038/ng.3656