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

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

The 1000 Genomes Project and disease-specific sequencing efforts are producing large collections of haplotypes that can be used as reference panels for genotype imputation in genome-wide association studies (GWAS). However, imputing from large reference panels with existing methods imposes a high computational burden. We introduce a strategy called 'pre-phasing' that maintains the accuracy of leading methods while reducing computational costs. We first statistically estimate the haplotypes for each individual within the GWAS sample (pre-phasing) and then impute missing genotypes into these estimated haplotypes. This reduces the computational cost because (i) the GWAS samples must be phased only once, whereas standard methods would implicitly repeat phasing with each reference panel update, and (ii) it is much faster to match a phased GWAS haplotype to one reference haplotype than to match two unphased GWAS genotypes to a pair of reference haplotypes. We implemented our approach in the MaCH and IMPUTE2 frameworks, and we tested it on data sets from the Wellcome Trust Case Control Consortium 2 (WTCCC2), the Genetic Association Information Network (GAIN), the Women's Health Initiative (WHI) and the 1000 Genomes Project. This strategy will be particularly valuable for repeated imputation as reference panels evolve.

© 2012 Nature America, Inc. All rights reserved.

Genotype imputation is a key step in the analysis of GWAS. The approach works by finding haplotype segments that are shared between study individuals, who are typically genotyped on a commercial array with 300,000–2,500,000 SNPs, and a reference panel of more densely typed individuals, such as those provided by the International HapMap Project[1,2] and the 1000 Genomes Project[3] or obtained by sequencing a subset of study individuals. Imputation methods can accurately estimate genotypes or genotype probabilities at markers that have not been directly examined in a GWAS. Imputed genotypes are now routinely used to increase the power of GWAS analyses, to guide fine-mapping efforts and to facilitate the meta-analysis of studies genotyped on different marker sets[4,5].

The maturation of high-throughput genotyping and sequencing technologies has led to a rapid increase in the size of publicly available reference data sets. For example, whereas HapMap Phase 2 included 210 unrelated individuals typed at ~4 million SNPs, the Phase 1 variant call set from the 1000 Genomes Project (released in March 2012) includes 1,092 individuals typed at >38 million polymorphic sites. The next phases of this project will extend this data set to over 2,000 individuals typed at an even greater number of sites, and other sequencing efforts are also producing large genetic variation resources.

These developments can provide immediate benefits to GWAS through imputation: a more complete catalog of variants increases the chances that causal or trait-associated variants will be imputed, and reference panels with more haplotypes increase imputation accuracy and power for downstream association analysis, especially for variants with low allele frequencies[4,5]. In contrast, many existing genotype imputation methods require substantial computing power when used with large reference data sets. This problem is compounded by the fact that reference collections are now regularly improved and expanded, such that investigators might benefit from imputing their samples multiple times over the course of a study.

Here, we propose a practical solution that maintains imputation accuracy while greatly reducing computational costs. Our approach is motivated by the observation that imputation methods spend much of their time accounting for the unknown phase of GWAS genotypes. Some methods do this through analytical calculations that integrate over all possible phase configurations for each study individual, whereas other methods average the imputation probabilities across multiple haplotypes sampled by a phasing algorithm[4]. Both approaches have limitations. Analytical phase integration becomes computationally expensive as reference panels grow and is only possible when the study individuals are treated independently, which sacrifices linkage disequilibrium (LD) information in the GWAS data. Sampling-based methods can scale better with reference panel size and capture LD information to improve imputation accuracy[6], but they may still have nontrivial computational costs because of the need to sample and impute into several haplotype configurations for each individual.

In our new approach, we first statistically estimate the haplotypes underlying the GWAS genotypes (in pre-phasing) and then impute into these haplotypes, as if they were correct; a schematic of a traditional workflow and the more efficient workflow proposed here are

[1]Department of Human Genetics, University of Chicago, Chicago, Illinois, USA. [2]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, USA. [3]Department of Statistics, University of Chicago, Chicago, Illinois, USA. [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. Correspondence should be addressed to G.R.A. (goncalo@umich.edu), J.M. (marchini@stats.ox.ac.uk) or M.S. (mstephens@uchicago.edu).

Received 13 September 2011; accepted 13 June 2012; published online 22 July 2012; doi:10.1038/ng.2354

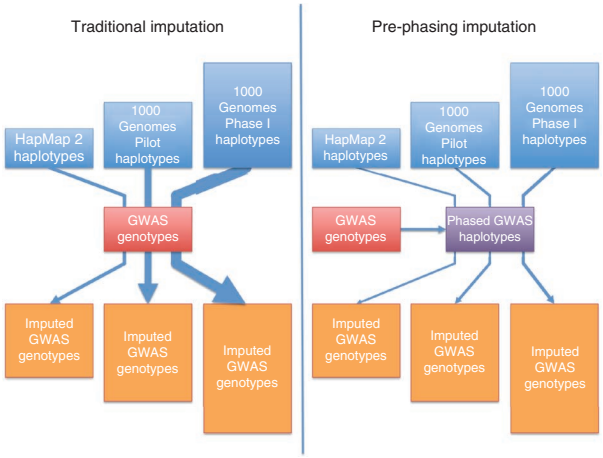

Figure 1 Imputation schematic. Each box represents a genetic data set and each arrow represents an analysis step. The sizes of the boxes reflect the relative numbers of genotypes they contain, and the widths of the arrows reflect the relative computational costs of the analyses. Given a single GWAS data set (red box), successively larger reference panels (blue boxes) lead to larger and more accurate imputed data sets (orange boxes). The computational cost of imputation is much lower when using pre-phased GWAS haplotypes (purple box) than when using traditional imputation approaches (left).

shown (**Fig. 1**). Imputing into pre-phased haplotypes is known to be fast, and it is highly accurate when the haplotypes are estimated through genotyped family members[7,8] or long segments of recent shared ancestry[9]. These two phasing techniques cannot be used on unrelated individuals from outbred populations (a common study design in GWAS), which means many data sets can only be phased by statistical algorithms that yield lower quality (but still reasonable) haplotypes. The central aims of this work are (i) to show that the GWAS haplotypes estimated by existing algorithms can produce accurate imputation and (ii) to quantify the efficiency gains from pre-phasing the study genotypes. We assume throughout that the reference genotypes were also phased before imputation, as is typical for public reference data sets.

RESULTS

Pre-phasing run-time performance

To show the computational advantages of pre-phasing, we analyzed a GWAS data set of 2,490 individuals from the 1958 British Birth Cohort of the WTCCC2 (ref. 10). We imputed this data set from a series of reference panels, using related imputation methods that account for phase uncertainty in different ways (**Table 1**). IMPUTE version 1 (IMPUTE1)[11] uses an analytical integration strategy. This was relatively efficient with a reference panel of 60 individuals (41 min per genome with 1000 Genomes Pilot data), but the computational burden grew quickly as haplotypes were added to the reference set. By contrast, IMPUTE version 2 (IMPUTE2)[6] uses a haplotype sampling strategy. This approach scaled more favorably with larger reference panels, but it still required 512 min per genome to impute from the latest 1000 Genomes panel. By comparison, an updated version of IMPUTE2 that uses our proposed approach required a one-time pre-phasing investment of 25 min per genome and then just 24 min to impute each sample from the largest reference panel. We observed similar trends with MaCH[12] (which typically uses a similar approach to IMPUTE1) and minimac (which performs imputation with pre-phased haplotypes in the MaCH framework) (**Supplementary Table 1**).

Haplotype estimation

These results show that pre-phasing can greatly speed up the imputation process, but the accuracy of imputation with this shortcut may depend on how well the GWAS haplotypes were estimated. The accuracy of computationally estimated haplotypes depends on a number of factors, including marker density, the relatedness of the sampled individuals, sample size and demography[12,13]. In founder populations[14], long-range haplotypes can be estimated very accurately, even with modest sample sizes[9]. For example, by comparing the results of population- and trio-based phasing in Finnish samples from the Finland–United States Investigation of Non-Insulin–Dependent Diabetes Mellitus Genetics (FUSION) study of type 2 diabetes[2,15], we estimated that population phasing produces <1 switch error[16] per 5.5 Mb of DNA. These results are highly accurate due to the large number of genotyped individuals (>2,000) and the fact that Finland is a founder population in which long haplotypes are shared by seemingly unrelated individuals. In more diverse populations, haplotype estimation may often be less accurate. For example, the distance between switches in European GWAS data sets is typically in the range of 0.6–1.4 Mb[17].

Genotype imputation accuracy in diverse populations

Here, we evaluate whether pre-phased haplotypes can be used to accurately impute missing genotypes in three GWAS data sets sampled from diverse populations: the WTCCC2 data described above, a European-American case-control data set from a psoriasis study by GAIN[18] and a set of data for African Americans from WHI[19]. In each data set, we masked and imputed a subset of the genotyped SNPs (details in **Table 2**). We measured imputation accuracy at these SNPs as the average squared correlation (mean R^2) between masked array genotypes and imputed allele dosages (posterior mean genotypes).

We used the GAIN data set to compare a well-benchmarked imputation method (MaCH) to a related method that uses pre-phasing (minimac). Encouragingly, both methods produced results with similar accuracy when applied to a common reference panel of 60 individuals from the Utah residents of Northern and Western European ancestry (CEU) population (**Table 2**). In fact, our pre-phasing strategy generated slightly better results, despite ignoring the uncertainty in the estimated GWAS haplotypes, possibly because

Table 1 Running times in WTCCC2 controls for different imputation methods and reference panels

	Imputation method		
Reference panel[a]	IMPUTE1	IMPUTE2 (sampling)[b]	IMPUTE2 (pre-phasing)[c]
HapMap 2 CEU (60 individuals, 2.5 million SNPs)	14	31	<1
1000 Genomes CEU (60 individuals, 7.3 million SNPs)	41	48	1
1000 Genomes EUR (283 individuals, 11.6 million SNPs)	1,287	144	6
1000 Genomes EUR (381 individuals, 37.4 million SNPs)	7,800[d]	512	24

Running times are shown as the central processing unit (CPU) minutes needed to impute across one whole individual genome. CEU, EUR: European populations. [a]Reference panels included HapMap 2 release 22, the 1000 Genomes low-coverage Pilot (June 2010), the 1000 Genomes interim release (August 2010) and the 1000 Genomes interim Phase 1 release (November 2010). [b]Version of the IMPUTE2 algorithm published by Howie *et al.*[6]. This method averages the imputation results across 20 sampled haplotype configurations per individual. [c]Running times do not include the initial investment required to phase the GWAS genotypes, which took 25 min per individual. [d]Projected running time extrapolated from existing benchmarks.

© 2012 Nature America, Inc. All rights reserved.

Table 2 Accuracy of different imputation methods and 1000 Genomes reference panels applied to various GWAS data sets

GWAS data set	Imputation method[a]	Reference panel[b]	Imputation accuracy (mean R^2)[c]		
			MAF 1–3%	MAF 3–5%	MAF >5%
GAIN psoriasis	MaCH or **minimac**	60 CEU individuals	0.67	0.76	0.91
(European American; $N = 2,759$)			**0.69**	**0.77**	**0.91**
		283 EUR individuals	**0.73**	**0.78**	**0.92**
		381 EUR individuals	**0.83**	**0.85**	**0.94**
WTCCC2	IMPUTE2	60 CEU individuals	0.66	0.78	0.88
(UK; $N = 2,490$)	(sampling or **pre-phasing**)		**0.65**	**0.77**	**0.87**
		283 EUR individuals	0.77	0.82	0.89
			0.75	**0.81**	**0.88**
		381 EUR individuals	0.84	0.88	0.92
			0.82	**0.86**	**0.91**
WHI	MaCH or **minimac**	60 CEU and 59 YRI individuals	0.51	0.73	0.83
(African-American; $N = 8,421$)			**0.49**	**0.70**	**0.80**
		283 EUR and 172 AFR individuals	**0.55**	**0.72**	**0.81**
		381 EUR and 174 AFR individuals	**0.61**	**0.75**	**0.83**
1000 Genomes EUR	IMPUTE2	380 EUR individuals	0.82	0.86	0.92
(European ancestry; $N = 381$)	(sampling or **pre-phasing**)	(WTCCC2 SNPs)	**0.81**	**0.85**	**0.91**
		380 EUR individuals (sequence SNPs)	0.66	0.79	0.91
			0.64	**0.78**	**0.90**

YRI, Yoruba from Ibadan, Nigeria; AFR, African population; CEU, EUR, European populations; from 1000 Genomes.
[a]We imputed each GWAS data set with an existing imputation method and its pre-phasing counterpart. [b]Reference panels used to impute each GWAS data set included the 1000 Genomes low-coverage Pilot (June 2010), the 1000 Genomes interim release (August 2010) and the 1000 Genomes interim Phase 1 release (November 2010). [c]Each cell shows the mean R^2 between true genotypes and imputed dosages for the specified MAF window and reference panel. For a given GWAS data set, all accuracy values within a MAF window were calculated on the same set of SNPs; the corresponding SNP counts are shown in **Supplementary Figure 1**. Accuracy values from pre-phasing are shown in bold (some analyses were performed only with pre-phasing).

pre-phasing captures joint LD information that is not used by the analytical phasing and imputation framework of methods like MaCH and IMPUTE1.

We then used the WTCCC2 data to compare pre-phasing to a haplotype sampling approach, both of which were implemented in IMPUTE2 (ref. 6). Our results again show that pre-phasing can provide comparable accuracy to that achieved with the existing imputation method, although, in this case, the pre-phasing results were slightly less accurate (**Table 2**). Both pre-phasing and haplotype sampling capture LD information in the GWAS data, but the sampling approach also accounts for some of the uncertainty in phasing the GWAS genotypes, which could explain why it was more accurate here.

Finally, as it is well established that phasing and imputation can be more challenging in individuals with recent African ancestry because of their reduced LD and higher genetic diversity, we evaluated our pre-phasing approach in the WHI GWAS of African Americans. In this comparison, pre-phasing was less accurate than the analytical approach in MaCH by the largest (but still small) margin (**Table 2**), which we interpret as evidence that accounting for phase uncertainty is more important when the haplotypes are harder to estimate.

The advantages of pre-phasing become particularly clear when considering successive reference panels that have been updated over time. Following a relatively modest pre-phasing investment, each new reference panel can be imputed at a low computational cost while improving the accuracy and completeness of the imputed genotypes. In agreement with this, adding haplotypes to the 1000 Genomes Project resource increased accuracy for all SNPs, especially those with minor allele frequency (MAF) of 1–3%, enhancing mean R^2 from 0.65 to 0.75 to 0.82 in the WTCCC2 data, from 0.69 to 0.73 to 0.83 in the GAIN data and from 0.49 to 0.55 to 0.61 in the WHI data (**Table 2**). Beyond the accuracy increase at known variants, each new panel also introduces many novel variants that could lead to additional association signals and biological insights.

Evaluation of imputation accuracy using sequence data

One caveat to the comparisons above is that SNPs on GWAS arrays tend to be more common (example in **Supplementary Fig. 1**) and are easier to impute than unascertained SNPs[2]. We addressed this issue by performing a cross-validation in the European (EUR) panel of 1000 Genomes Phase 1, which includes a more complete set of SNPs discovered by low-pass whole-genome and high-pass exome sequencing in >1,000 individuals. For each of the 381 EUR individuals in turn, we masked genotypes on chromosome 10 at all sites except those included on the Affymetrix 500k SNP array and then imputed the missing sites using the Affymetrix 500k scaffold and the remaining 760 EUR haplotypes. To mimic pre-phasing in a GWAS, we reduced the EUR data set to sites present on the array scaffold, phased the genotypes again and then used these estimated haplotypes when imputing masked genotypes for a given individual (bottom rows in **Table 2**).

To provide a point of comparison with the GWAS results in **Table 2**, we initially imputed only the SNPs that were used in the WTCCC2 analysis (WTCCC2 SNPs). The imputation accuracy at these SNPs was slightly lower in the EUR cross-validation than in the WTCCC2 analysis; for example, pre-phasing in EUR produced mean R^2 values of 0.81, 0.85 and 0.91 for SNPs in ascending MAF bins, compared to 0.82, 0.86 and 0.91 for the WTCCC2 experiment with the same scaffold SNPs, reference panel and phasing approach (**Table 2**). These differences in pre-phasing accuracy may reflect the relative amount of phase information in a sample of 381 individuals (1000 Genomes EUR) and a sample of nearly 2,500 individuals (WTCCC2). Nonetheless, the overall similarity in results suggests that our EUR cross-validation provides a good approximation to a European GWAS.

We next extended the experiment by imputing the full set of SNPs in the EUR sequence data (sequence SNPs). As expected, the sequence SNPs were imputed less accurately than the WTCCC2 SNPs within each frequency bin. For example, haplotype sampling produced mean R^2 values of 0.82, 0.86 and 0.92 (for MAFs of 1–3%, 3–5% and >5%, respectively) in the array SNP analysis, but the accuracy dropped

© 2012 Nature America, Inc. All rights reserved.

to 0.66, 0.79 and 0.91, respectively, when evaluating all sequence SNPs in the same frequency ranges (**Table 2**). Despite the added difficulty of imputing low-frequency and unascertained variants, pre-phasing was still nearly as effective as haplotype sampling at these SNPs (mean R^2 of 0.64 versus 0.66 for MAFs of 1–3%; **Table 2**). This analysis also allows us to measure the accuracy at SNPs with MAFs of < 1%, for which we observed mean R^2 values of 0.42 and 0.44 for pre-phasing and haplotype sampling, respectively. Hence, although all methods have lower imputation accuracy at unascertained and low-frequency SNPs, pre-phasing still achieves competitive accuracy at such variants.

Multiple imputations

The examples in **Table 2** show that imputation accuracy may sometimes decrease when using the most likely haplotype pair for each GWAS individual rather than integrating over the phase uncertainty. We also note that, over the span of entire chromosomes and in the data sets examined here, haplotype estimates will almost never match the true underlying haplotypes. These considerations led us to assess whether we could improve accuracy for a reasonable increment in computing time by saving multiple sampled haplotype configurations at the pre-phasing stage and then imputing into each of these (details in the **Supplementary Note**). Imputing into 4–10 sampled haplotypes per individual provided a small increase in accuracy while increasing computational costs by 4–10× (**Supplementary Figs. 2** and **3**). Using a much larger number of sampled haplotypes (up to 500 per individual, for a 500× increase in computing time) provided only a modest additional increase in accuracy (**Supplementary Fig. 2**), which confirms that a single pre-phased configuration provides nearly as much accuracy as much more computationally intensive methods for capturing haplotype uncertainty. These results suggest that pre-phasing is a good general strategy for genome-wide imputation, whereas slower but more accurate approaches may be useful for follow-up analyses near putative disease-causing loci.

DISCUSSION

We have described a practical strategy for imputing genotypes from the large reference panels that are now emerging from sequencing efforts, such as the 1000 Genomes Project. These panels are regularly updated, both to incorporate newly sequenced individuals and to take advantage of improved methods for analyzing next-generation sequencing data that can handle increasingly diverse variant types, including insertion and/or deletion polymorphisms and copy-number variants. New reference data sets may provide substantial benefits for disease studies, but imputing them into large-scale GWAS currently requires substantial computational resources. The pre-phasing strategy introduced here will allow investigators to routinely impute from these emerging reference panels at a reasonable computational cost and will thereby enhance studies of complex disease genetics.

Overall, our results show that pre-phasing provides comparable accuracy to state-of-the-art imputation methods. Although we focus on selected combinations of data and methods (**Tables 1** and **2**), we also find that minimac and IMPUTE2 produce very similar trends in accuracy and running time when applied to the same data set (for details, see **Supplementary Fig. 4** and compare **Table 1** and **Supplementary Table 1**).

It is somewhat unexpected that pre-phasing remains competitive with other methods when imputing rare variants (MAF < 1%; **Table 2**). Such variants should require longer flanking haplotypes for successful imputation, and a single pre-phasing solution may include errors that break up long-range haplotypes. One possible explanation is that existing methods also struggle to infer very long haplotypes, such that pre-phasing still seems accurate by comparison. Conversely, it is important to realize that imputation accuracy is affected by phasing accuracy in the reference panel and by the GWAS SNPs used to drive imputation[4,5]. Imperfections in the reference haplotypes would limit imputation accuracy, even with perfectly phased GWAS haplotypes, and it may be difficult to impute rare variants with any method when using sparse GWAS scaffolds. These factors may outweigh the differences between methods that use pre-phasing and those that integrate over phase uncertainty.

In the **Supplementary Note**, we consider extensions of the pre-phasing approach, including an exploration of haplotype sampling approaches and an example of how imputation accuracy can be improved by pre-phasing with other haplotyping engines, such as SHAPEIT[17]. We also note that when genotypes from family members are available, it may be particularly attractive to use our imputation software with haplotypes informed by transmission patterns in pedigrees, where the best phasing tool may depend on family structure.

Our results show that pre-phasing is a highly generalizable strategy that can be adapted to most imputation engines, and we expect that it will be combined with future methodological developments to make imputation even faster and more flexible. Software implementing our approach, using either the IMPUTE2 or MaCH and minimac framework, is available from the authors' websites (see URLs).

URLs. Minimac and instructions on how to implement the pre-phasing approach describe here using MaCH, http://genome.sph.umich.edu/wiki/minimac; separate pre-phasing and imputation with IMPUTE 2.0, http://mathgen.stats.ox.ac.uk/impute/impute_v2.html#prephasing_gwas; the data sets analyzed here available from the database of Genotypes and Phenotypes (dbGaP), http://www.ncbi.nlm.nih.gov/gap, and WTCCC, http://www.wtccc.org.uk/.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank M. Boehnke for critical reading, advice and suggestion, Y. Li for aid with cleaning the WHI data and the two anonymous reviewers for their helpful comments. B.H. and M.S. were supported by a grant from the National Human Genome Research Institute (NHGRI; HGO2585) to M.S. J.M. was supported by a grant from the UK Medical Research Council (G0801823). C.F. and G.R.A. were supported by grants from the US National Institutes of Health (NIH; DK0855840, HG005552 and HG005581). This study makes use of data generated by the WTCCC, GAIN and WHI. A full list of the investigators who contributed to the generation of the WTCCC data is available from the WTCCC web site (see URLs). The WTCCC was partially funded by the Wellcome Trust under awards 076113 and 085475. For details of contributors to the GAIN and WHI studies, please see the corresponding dbGaP accessions.

AUTHOR CONTRIBUTIONS

B.H., C.F., M.S., J.M. and G.R.A. designed the methods and experiments. B.H. and C.F. ran the experiments and wrote the first draft; all authors contributed critical reviews of the manuscript during its preparation.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/ng.2354.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

© 2012 Nature America, Inc. All rights reserved.

1. International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
2. Altshuler, D.M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
3. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
4. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).
5. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu. Rev. Genomics Hum. Genet.* **10**, 387–406 (2009).
6. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
7. Burdick, J.T., Chen, W.M., Abecasis, G.R. & Cheung, V.G. *In silico* method for inferring genotypes in pedigrees. *Nat. Genet.* **38**, 1002–1004 (2006).
8. Chen, W.M. & Abecasis, G.R. Family-based association tests for genomewide association scans. *Am. J. Hum. Genet.* **81**, 913–926 (2007).
9. Kong, A. *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat. Genet.* **40**, 1068–1075 (2008).
10. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
11. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007).
12. 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**, 816–834 (2010).
13. Varilo, T. & Peltonen, L. Isolates and their potential use in complex gene mapping efforts. *Curr. Opin. Genet. Dev.* **14**, 316–323 (2004).
14. Peltonen, L., Palotie, A. & Lange, K. Use of population isolates for mapping complex traits. *Nat. Rev. Genet.* **1**, 182–190 (2000).
15. Scott, L.J. *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316**, 1341–1345 (2007).
16. Marchini, J. *et al.* A comparison of phasing algorithms for trios and unrelated individuals. *Am. J. Hum. Genet.* **78**, 437–450 (2006).
17. Delaneau, O., Marchini, J. & Zagury, J.F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2012).
18. Manolio, T.A. *et al.* New models of collaboration in genome-wide association studies: the Genetic Association Information Network. *Nat. Genet.* **39**, 1045–1051 (2007).
19. Women's Health Initiative Study Group. Design of the Women's Health Initiative clinical trial and observational study. The Women's Health Initiative Study Group. *Control. Clin. Trials* **19**, 61–109 (1998).

© 2012 Nature America, Inc. All rights reserved.

ONLINE METHODS

FUSION data set. The FUSION Study consists of 1,161 Finnish individuals with type 2 diabetes (T2D) and 1,174 normal, glucose-tolerant Finnish controls. Samples were genotyped with the Illumina Human-Hap300 BeadChip (v1.1). In total, 306,222 autosomal SNPs passed quality control (Hardy-Weinberg equilibrium $P \geq 1 \times 10^{-6}$ in the total sample; call frequency ≥ 0.90; and MAF > 0.01)[15]. In addition, 120 trios were genotyped with the same chip, and haplotypes were estimated on the basis of the most likely pattern of gene flow using Merlin[20] and were compared with haplotypes estimated statistically using population information and MaCH[12].

GAIN psoriasis data set. GAIN[21] supported a series of GWAS designed to identify specific positions of DNA variation associated with the occurrence of a particular common disease. Data used for this study were from 1,359 psoriasis cases and 1,400 controls genotyped at Perlegen Sciences using a custom genotyping array. In total, 438,670 autosomal SNPs passed the quality control filters (Hardy-Weinberg equilibrium $P \geq 1 \times 10^{-6}$ in the total sample; call frequency ≥ 0.95; and MAF > 0.01)[21]. In this study, 88 individuals were also genotyped using Affymetrix 6.0 arrays and these genotypes were used to evaluate 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).

WTCCC2 data set. We used genotype data from the WTCCC2 (ref. 10) on members of the 1958 British Birth Cohort, which is comprised of controls sampled from the UK. These individuals were genotyped on Affymetrix 6.0 and Illumina 1.2 M SNP arrays. The WTCCC2 merged genotypes across platforms and applied standard quality control filters, which resulted in data from 2,490 individuals at 71,190 SNPs on chromosome 10. For our imputation experiments, we masked the SNPs not found on the Affymetrix 500k array, imputed the masked SNPs and compared the imputed dosages to the original array genotypes.

WHI data set. We obtained genotype data for the WHI[19] study from dbGaP (see URLs). The data set included 8,421 African Americans genotyped on Affymetrix 6.0. We removed SNPs with genotype call rate of <90%, Hardy-Weinberg equilibrium P value of $<1 \times 10^{-6}$ or MAF of <1%, resulting in 829,370 SNPs passing quality control criteria. For our imputation experiments, we masked every tenth SNP and repeated in sliding windows, such that each analysis was informed by ~90% of the array SNPs and every array SNP was imputed exactly once.

Phasing. Haplotyping approaches, such as those implemented in MaCH and IMPUTE2, proceed through a series of iterative steps. In each step, a new pair of haplotypes is sampled for each individual as an imperfect mosaic[22] of the estimated haplotypes (templates) for other individuals in the data set. After a number of iterations, best-guess haplotypes are constructed for each individual by combining information across the sampled haplotype configurations; both MaCH and IMPUTE2 perform this step by minimizing the expected switch error rate[23]. The computational cost of phasing with these methods depends on the number of iterations performed and the number of template haplotypes that are used in each update. For the experiments described here, we used 20 iterations and 200–400 templates for MaCH and 30 iterations (first 10 discarded as burn-in) and 80 templates for IMPUTE2. These methods differ in various details, such as how they fit the parameters of their models and how they choose templates for each haplotype sampling step; further information is provided in the original papers[6,12].

Imputation into phased haplotypes. When GWAS genotypes have been phased before imputation, each haplotype can be imputed separately, if we assume that the GWAS haplotypes are conditionally independent, given a reference panel. The reference panel provides template haplotypes for the imputation model, and marginal probabilities for the untyped alleles in each GWAS haplotype are estimated via standard hidden Markov model (HMM) calculations (the forward-backward algorithm)[24]. The parameters of HMM are estimated in different ways by minimac and IMPUTE2; see elsewhere for details[6,12]. Allelic probabilities are converted to genotypic probabilities for each individual by assuming Hardy-Weinberg equilibrium; these genotypic probabilities can be directly compared with those produced by other imputation approaches.

Computational costs. Many existing imputation methods (for example, MaCH and IMPUTE1) use analytical integration to account for the unknown phase of GWAS genotypes. The computational cost of this approach is proportional to the number of GWAS individuals (N), the number of genotyped markers in the reference panel (M_{REF}) and the square of the number of reference haplotypes (H^2), or $O(N \times M_{REF} \times H^2)$. Some methods, such as fastPHASE[23] and Beagle[25], reduce H by grouping similar haplotypes into clusters. The quadratic term affects all markers, whether they are typed in a GWAS or only in the reference panel. Consequently, the computational cost grows quickly with reference panel size, and it can be time-consuming to run these methods on modern reference data sets.

IMPUTE2 aims to reduce the computing burden through a Monte Carlo algorithm that separates the phasing and imputation tasks. This approach alternately samples phase configurations for genotyped markers and imputes allele probabilities for markers not typed in the GWAS. The cost of the phasing component is proportional to the number of GWAS individuals (N), the number of genotyped markers in the GWAS data (M_{GWAS}), the number of iterations (I) and the square of the number of templates used in each phasing update (K^2), or $O(N \times M_{GWAS} \times I \times K^2)$. The cost of the imputation component is proportional to the number of GWAS haplotypes ($2N$), the number of markers in the reference panel (M_{REF}), the number of iterations (I) and the number of haplotypes in the reference panel (H), or $O(N \times M_{REF} \times I \times H)$. Partitioning the analysis in this way allows better scaling with reference panel size, but it requires I repetitions of the imputation step (one for each sampled phase configuration).

Like the IMPUTE2 Monte Carlo algorithm, pre-phasing separates the phasing and imputation steps when imputing a GWAS data set. The computational cost of pre-phasing in our framework is $O(N \times M_{GWAS} \times I \times K^2)$. This is the same as the phasing cost for Monte Carlo integration, although, in this context, the phasing must be performed only once per GWAS data set. Given a set of pre-phased GWAS haplotypes, the cost of imputation is then $O(N \times M_{REF} \times H)$; the efficiency of this step makes imputation from pre-phased haplotypes very fast. The cost of each step in our current computing system, in CPU hours, is approximately $N \times M_{GWAS} \times I \times K^2 \times 10^{-11}$ for phasing and $N \times M_{REF} \times H \times 10^{-11}$ for imputation.

20. Abecasis, G.R. & Wigginton, J.E. Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. *Am. J. Hum. Genet.* **77**, 754–767 (2005).
21. Nair, R.P. *et al.* Genome-wide scan reveals association of psoriasis with IL-23 and NF-κB pathways. *Nat. Genet.* **41**, 199–204 (2009).
22. Stephens, M. & Donnelly, P. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.* **73**, 1162–1169 (2003).
23. Scheet, P. & Stephens, M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* **78**, 629–644 (2006).
24. Baum, L.E., Petrie, T., Soules, G. & Weiss, N. A maximization technique occurring in statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.* **41**, 164–171 (1970).
25. Browning, B.L. & Browning, S.R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **84**, 210–223 (2009).

© 2012 Nature America, Inc. All rights reserved.

doi:10.1038/ng.2354