



# Genotype imputation for genome-wide association studies

Jonathan Marchini\* and Bryan Howie†

**Abstract** | In the past few years genome-wide association (GWA) studies have uncovered a large number of convincingly replicated associations for many complex human diseases. Genotype imputation has been used widely in the analysis of GWA studies to boost power, fine-map associations and facilitate the combination of results across studies using meta-analysis. This Review describes the details of several different statistical methods for imputing genotypes, illustrates and discusses the factors that influence imputation performance, and reviews methods that can be used to assess imputation performance and test association at imputed SNPs.

## Hidden Markov model

A class of statistical model that can be used to relate an observed process across the genome to an underlying, unobserved process of interest. Such models have been used to estimate population structure and admixture, for genotype imputation and for multiple testing.

Genotype imputation is the term used to describe the process of predicting or imputing genotypes that are not directly assayed in a sample of individuals. There are several distinct scenarios in which genotype imputation is desirable, but the term now most often refers to the situation in which a reference panel of haplotypes at a dense set of SNPs is used to impute into a study sample of individuals that have been genotyped at a subset of the SNPs. An overview of this process is given in BOX 1. Genotype imputation can be carried out across the whole genome as part of a genome-wide association (GWA) study or in a more focused region as part of a fine-mapping study. The goal is to predict the genotypes at the SNPs that are not directly genotyped in the study sample. These ‘*in silico*’ genotypes can then be used to boost the number of SNPs that can be tested for association. This increases the power of the study, the ability to resolve or fine-map the causal variant and facilitates meta-analysis. BOX 2 discusses these uses of imputation as well as the imputation of untyped variation, human leukocyte antigen (HLA) alleles, copy number variants (CNVs), insertion–deletions (indels), sporadic missing data and correction of genotype errors.

The HapMap 2 haplotypes<sup>1</sup> have been widely used to carry out imputation in studies of samples that have ancestry close to those of the HapMap panels. The CEU (Utah residents with northern and western European ancestry from the CEPH collection), YRI (Yoruba from Ibadan, Nigeria) and JPT + CHB (Japanese from Tokyo, Japan and Chinese from Beijing, China) panels consist of 120, 120 and 180 haplotypes, respectively, at a very dense set of SNPs across the genome. Most studies have used a two-stage procedure that starts by imputing the missing

genotypes based on the reference panel without taking the phenotype into account. Imputed genotypes at each SNP together with their inherent uncertainty are then tested for association with the phenotype of interest in a second stage. The advantage of the two-stage approach is that different phenotypes can be tested for association without the need to redo the imputation.

This Review provides an overview of the different methods that have been proposed for genotype imputation, discusses and illustrates the factors that affect the accuracy of genotype imputation, discusses the use quality-control measures on imputed data and methods that can be employed in testing for association using imputed genotypes.

## Genotype imputation methods

We assume that we have data at  $L$  diallelic autosomal SNPs and that the two alleles at each SNP have been coded 0 and 1. Let  $H$  denote a set of  $N$  haplotypes at these  $L$  SNPs and let  $G$  denote the set of genotype data at the  $L$  SNPs in  $K$  individuals with  $G_i = \{G_{i1}, \dots, G_{iL}\}$  denoting the genotypes of the  $i$ th individual. The individual genotypes are either observed so that  $G_{ik} \in \{0, 1, 2\}$  or they are missing so that  $G_{ik} = \text{missing}$ . The main focus here is in predicting the genotypes of those SNPs that have not been genotyped in the study sample at all but there are usually sporadic missing genotypes as well. We assume that strand alignment between data sets has been carried out ([Supplementary information S1](#) (box)).

**IMPUTE v1.** IMPUTE v1 (REF. 2) is based on an extension of the hidden Markov models (HMMs) originally developed as part of importance sampling schemes for

\*Department of Statistics, University of Oxford, Oxford, UK.

†Department of Human Genetics, University of Chicago, Chicago, USA. Correspondence to J.M. e-mail: [marchini@stats.ox.ac.uk](mailto:marchini@stats.ox.ac.uk) doi: 10.1038/nrg2796 Published online 2 June 2010

### Linkage disequilibrium

The statistical association within gametes in a population of the alleles at two loci. Although linkage disequilibrium can be due to linkage, it can also arise at unlinked loci — for example, because of selection or non-random mating.

simulating coalescent trees<sup>3,4</sup> and for modelling linkage disequilibrium (LD) and estimating recombination rates<sup>5</sup>. The method is based on an HMM of each individual's vector of genotypes,  $G_i$ , conditional on  $H$ , and a set of parameters. This model can be written as

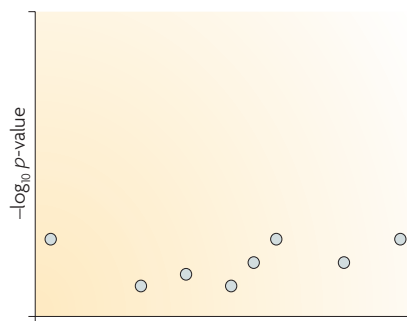
$$P(G_i|H, \theta, \rho) = \sum_z P(G_i|Z, \theta), P(Z|H, \rho) \quad (1)$$

in which  $Z = \{Z_1, \dots, Z_L\}$  with  $Z_j = \{Z_{j1}, Z_{j2}\}$  and  $Z_{jk} = \{1, \dots, N\}$ . The  $Z_j$  can be thought of as the pair of haplotypes from the reference panel at SNP  $j$  that are being copied to

form the genotype vector. The term  $P(Z|H, \rho)$  models how the pair of copied haplotypes changes along the sequence and is defined by a Markov chain in which switching between states depends on an estimate of the fine-scale recombination map ( $\rho$ ) across the genome. The term  $P(G_i|Z, \theta)$  allows each observed genotype vector to differ through mutation from the genotypes determined by the pair of copied haplotypes and is controlled with the mutation parameter  $\theta$ . Estimates of the fine-scale recombination map ( $\rho$ ) are provided on the IMPUTE v1 webpage and  $\theta$  is fixed internally by the

### Box 1 | How genotype imputation works

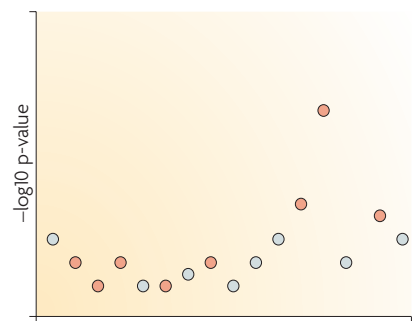
**b** Testing association at typed SNPs may not lead to a clear signal



**d** Reference set of haplotypes, for example, HapMap

0	0	0	0	1	1	1	0	0	1	1	1	1	1	0
1	1	1	1	1	1	1	0	0	1	0	0	1	1	1
1	1	1	1	1	0	1	0	0	1	0	0	0	1	0
0	0	1	0	1	1	1	0	0	1	1	1	1	1	0
1	1	1	0	1	1	0	0	1	1	1	0	1	1	1
0	0	1	0	1	1	1	0	0	1	1	1	1	1	0
1	1	1	1	1	0	1	0	0	1	0	0	0	1	0
1	1	1	0	0	1	0	0	1	1	1	0	1	1	1
0	0	0	0	1	1	1	0	0	1	1	1	1	1	0
1	1	1	0	0	1	0	0	1	1	1	0	1	1	0

**f** Testing association at imputed SNPs may boost the signal



**a** Genotype data with missing data at untyped SNPs (grey question marks)

1	?	?	?	?	1	?	1	?	0	2	2	?	?	?	?	0
0	?	?	?	?	2	?	2	?	0	2	2	?	?	?	?	0
1	?	?	?	?	2	?	2	?	0	2	1	?	?	?	?	0
1	?	?	?	?	2	?	1	?	1	2	2	?	?	?	?	0
2	?	?	?	?	2	?	2	?	1	2	1	?	?	?	?	0
1	?	?	?	?	1	?	1	?	1	2	2	?	?	?	?	0
1	?	?	?	?	2	?	2	?	0	2	1	?	?	?	?	1
2	?	?	?	?	1	?	1	?	1	2	1	?	?	?	?	1
1	?	?	?	?	0	?	0	?	2	2	2	?	?	?	?	0

**c** Each sample is phased and the haplotypes are modelled as a mosaic of those in the haplotype reference panel

0	?	?	?	?	1	?	1	?	0	1	1	?	?	?	?	0
1	?	?	?	?	1	?	1	?	0	1	1	?	?	?	?	0
1	?	?	?	?	1	?	1	?	0	1	0	?	?	?	?	0
1	?	?	?	?	1	?	1	?	1	1	1	?	?	?	?	0
1	?	?	?	?	0	?	0	?	1	1	1	?	?	?	?	0
0	?	?	?	?	0	?	0	?	1	1	1	?	?	?	?	0

**e** The reference haplotypes are used to impute alleles into the samples to create imputed genotypes (orange)

1	1	1	1	1	2	1	0	0	2	2	0	2	2	2	2	0
0	0	1	0	2	2	2	0	0	2	2	2	2	2	2	2	0
1	1	1	1	2	2	2	0	0	2	1	1	2	2	2	2	0
1	1	2	0	2	2	1	0	1	2	2	1	2	2	2	2	0
2	2	2	2	2	1	2	0	1	2	1	1	2	2	2	2	0
1	1	1	0	1	2	1	0	1	2	2	1	2	2	2	2	0
1	1	2	1	2	1	2	0	0	2	1	1	1	2	1	1	1
2	2	2	1	1	1	1	0	1	2	1	0	1	2	1	1	1
1	2	2	0	0	2	0	0	2	2	2	1	2	2	2	2	0

In samples of unrelated individuals, the haplotypes of the individuals over short stretches of sequence will be related to each other by being identical by descent (IBD). The local pattern of IBD can be described by an (unobserved) genealogical tree, which will differ at different loci throughout the genome owing to recombination. Imputation methods attempt to identify sharing between the underlying haplotypes of the study individuals and the haplotypes in the reference set and use this sharing to impute the missing alleles in study individuals. For this reason, there are strong connections between the models and methods used to infer haplotype phase and those used to perform genotype imputation<sup>22,37</sup>, as well as strong connections to tagging SNP-based approaches<sup>19,21,38</sup> and methods used in linkage studies<sup>39,40</sup>.

The figure above illustrates imputation for a sample of unrelated individuals. The raw data consist of a set of genotyped SNPs that has a large number of SNPs without any genotype data (part **a**). Testing for association at just these SNPs may not lead to a significant association (part **b**).

Imputation attempts to predict these missing genotypes. Algorithms differ in their details but all essentially involve phasing each individual in the study at the typed SNPs. The figure highlights three phased individuals (part **c**). These haplotypes are compared to the dense haplotypes in the reference panel (part **d**). Strand alignment between data sets must be done before this comparison takes place (Supplementary information S1 (box)). The phased study haplotypes have been coloured according to which reference haplotypes they match. This highlights the idea implicit in most phasing and imputation models that the haplotypes of a given individual are modelled as a mosaic of haplotypes of other individuals. Missing genotypes in the study sample are then imputed using those matching haplotypes in the reference set (part **e**). In real examples, the genotypes are imputed with uncertainty and a probability distribution over all three possible genotypes is produced. It is necessary to take account of this uncertainty in any downstream analysis of the imputed data. Testing these imputed SNPs can lead to more significant associations (part **f**) and a more detailed view of associated regions.

## Box 2 | Uses of imputation

**Boosting power**

Imputation can lead to a boost in power of up to 10% over testing only genotyped SNPs in a genome-wide association (GWA) study<sup>26</sup>. Other simulations have shown that the most benefit occurs for rare SNPs, which are harder to tag<sup>2</sup>.

**Fine-mapping**

Imputation provides a high-resolution view of an associated region and increases the chance that a causal SNP can be directly identified. The figure below shows imputation in the transcription factor 7-like 2 (*TCF7L2*) gene in the Wellcome Trust Case-Control Consortium (WTCCC) type 2 diabetes scan (image reproduced, with permission, from *Nature Genetics* REF. 2 © (2007) Macmillan Publishers Ltd). The upper part of the plot shows the  $-\log_{10}$   $p$ -values using called genotypes (black circles) and imputed genotypes (red circles). The largest  $-\log_{10}$   $p$ -value at a genotyped SNP (rs4506565) is 12.25, whereas the largest  $-\log_{10}$   $p$ -value at an imputed SNP (rs7903146) is 13.57. Imputed SNPs that show large associations can be better candidates for replication studies. The limiting factor that determines precision of fine-mapping is effectively the number of recombination events that have occurred in a region, or the amount of linkage disequilibrium (LD) in the region. Increasing sample size or mapping in populations with lower levels of LD will both act to increase the chance that a true causal variant can be identified.

**Meta-analysis**

If different cohorts have used different genotyping chips, imputation can be used to equate the set of SNPs in each study. The studies can then be combined in a meta-analysis at each SNP to boost power. This approach has been successful in identifying several new loci for a range of different traits<sup>31,41–44</sup>. A good practical guide to meta-analysis is provided by de Bakker *et al.*<sup>45</sup>. Usually each cohort is analysed separately as this allows conditioning on cohort-specific covariates. The results are then combined using fixed-effects or random-effects models.

**Imputation of untyped variation**

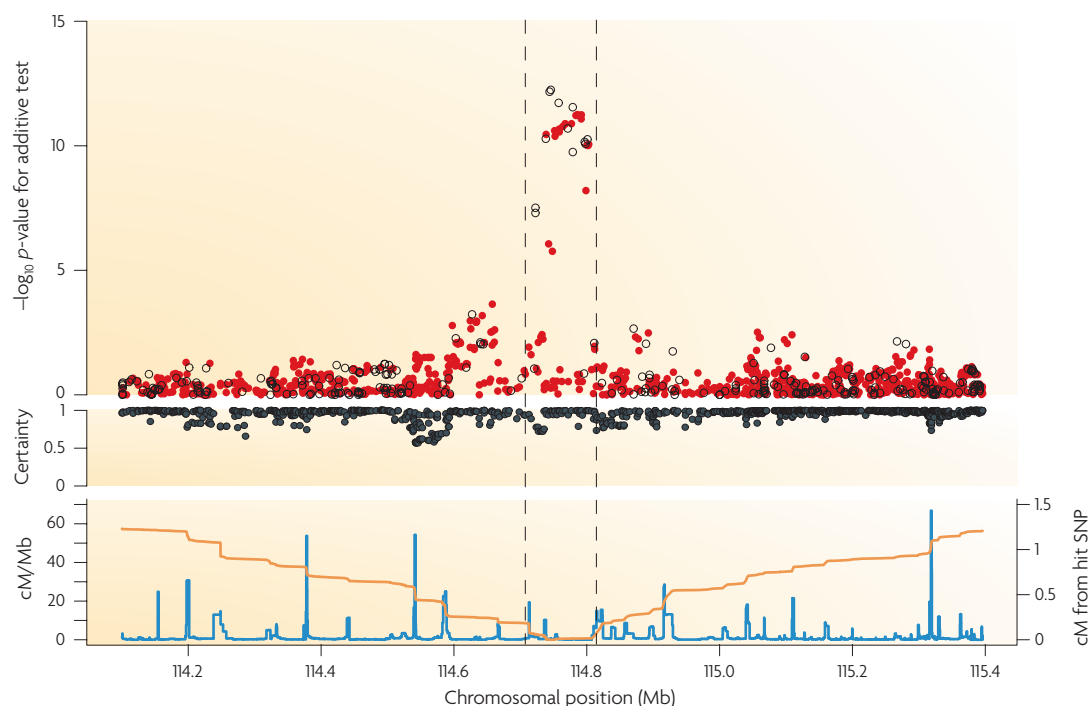
Imputation of SNPs which have not been typed in the haplotype reference panel or the study sample is also possible. Some methods do this through inference of the genealogy between study sample haplotypes<sup>2,46–48</sup> whereas others aim to identify haplotype effects more directly<sup>49</sup>. These methods can lead to a boost in power, especially when the causal variant is rare or there is local heterogeneity in the signal of association<sup>48</sup>.

**Imputation of non-SNP variation**

The general idea of imputation is readily extended to other types of genetic variation such as copy number variants and classical human leukocyte antigen alleles<sup>50</sup>. Looking ahead, the imputation of large numbers of small insertions and deletions (indels) that will be discovered from sequencing-based projects such as the 1000 Genomes Project are likely to be widely adopted in GWA studies.

**Sporadic missing data imputation and correction of genotyping errors**

Many of the widely used imputation programs allow imputation of sporadic missing genotypes that can occur when calling genotypes from genotyping chips (TABLE 1). Genotyping error rates are often very low (0.2% in the WTCCC study) so this type of imputation will not greatly boost power but can help control false positives at SNPs for which genotype calling is challenging. Recently, the BEAGLE model<sup>51</sup> has been extended to handle genotype intensity data so that genotypes can be called by using LD information between SNPs and this offers a small improvement in genotyping error rates.



### Box 3 | Chromosome X imputation

Imputation on chromosome X is fundamentally similar to imputation on the autosomes, but it is complicated by the fact that males carry only one copy of the chromosome (that is, they are hemizygous), in contrast to the two copies carried by females. This is mainly a practical issue: most imputation models could in principle accommodate a mixture of haploid (male) and diploid (female) genotypes, but few methods have implemented this function in their software (TABLE 1). One way around this is to treat male genotypes as diploids that are homozygous at every SNP. However, 'double counting' the male haplotypes in this way could decrease accuracy for some methods.

Performing imputation on chromosome X also requires care because the hemizygosity in males reduces the effective population size ( $N_e$ ) on this chromosome to about three-quarters of the autosomal value. Imputation algorithms work by copying stretches of shared haplotypes between individuals, and a lower  $N_e$  causes the copied haplotype segments to be longer and exhibit fewer base pair differences. Non-parametric methods (like BEAGLE) and methods that fit population genetic parameters in each data set (like fastPHASE and MACH) can automatically adjust to these characteristics of chromosome X, although they still need special functions to treat males and females differently. By contrast, IMPUTE specifies the population genetic parameters of its model ahead of time, so these must be modified for use on chromosome X.

IMPUTE v1 and v2 model historical recombination by reading in a fine-scale genetic map and scaling it by  $4N_e$ . This scaling is reduced to  $3N_e$  on chromosome X, which captures the idea that transitions between copied haplotypes should occur less often. It is tempting to reduce the 'mutation rate'  $\theta$  by the same factor, but this turns out to be wrong because the  $\theta$  parameter in the Li and Stephens model is defined conditional on a site carrying a mutation in the sample. This formulation assumes that there is already a mutation in the genealogical tree, and that we want to know the probability that a new haplotype will coalesce onto (or 'copy') a lineage that carries the mutation. This probability is a property of the bipartitions of a neutral coalescent tree, and it is invariant to constant scalings of the branch lengths. Hence, although the genealogical branches are shorter on chromosome X,  $\theta$  remains the same in IMPUTE v1 and v2.

program. The effective population size parameter ( $N_e$ ) must be specified by the user but estimates of this parameter are available for a wide range of human populations and our experience is that performance is robust to variation in this parameter. More details of these terms and parameters are given in REFS 2, 5.

Exact marginal probability distributions for the missing genotypes that are conditional on the observed genotype data in the vector  $G_i$  are obtained using the forward-backward algorithm for HMMs<sup>6</sup>. Using a simple modification to the algorithm it is also possible to obtain a marginal distribution for genotypes that are not missing. This provides a useful method of validating observed genotypes and allows quality assessment of imputation runs. IMPUTE v1 can also carry out imputation on the X chromosome and this is described in BOX 3.

**IMPUTE v2.** IMPUTE v2 (REF. 7) uses a related but more flexible approach than IMPUTE v1. SNPs are first divided into two sets: a set  $T$  that is typed in both the study sample and reference panel, and a set  $U$  that is untyped in the study sample but typed in the reference panel. The algorithm involves estimating haplotypes at SNPs in  $T$  (using the IMPUTE v1 HMM) and then imputing alleles at SNPs in  $U$  conditional on the current estimated haplotypes. As the imputation step is haploid imputation, it is very fast ( $O(N)$ ) compared with diploid imputation ( $O(N^2)$ ) carried out in IMPUTE v1. Phase uncertainty is accounted for by iterating these steps using a Markov chain Monte Carlo (MCMC)

approach. As imputation performance is driven by accurate matching of haplotypes, the method focuses on accurate haplotype estimation at the SNPs in  $T$  using as many individuals as possible.

Alternating between phasing and haploid imputation at a carefully chosen subset of SNPs is particularly suited to study designs in which different amounts of genotype data are available in different cohorts of a study. For example, IMPUTE v2 can use both the set of haplotypes from the pilot data of the [1000 Genomes Project](#) (see Further information for a link) and haplotype sets from the HapMap3 data set as reference panels for imputation. Compared with imputation from HapMap2, this provides a much larger set of imputed SNPs and a notable boost in accuracy at those SNPs included in the HapMap3 SNP set. Other methods can be made to handle this imputation scenario but IMPUTE v2 has been shown to be the most accurate approach<sup>7</sup> and the program makes it straightforward to apply.

When phenotype is strongly correlated with genotyping platform Howie *et al.*<sup>7</sup> found that imputing untyped SNPs in cases from SNPs that are present in a dense set of genotype data from controls did not lead to increased false-positive rates. However, if cases and controls are typed on different chips, then imputing SNPs that are untyped in both cases and controls from a haplotype panel can lead to false-positive associations. SNPs that are imputed accurately from one chip but poorly from another chip may lead to differences in allele frequency that just reflect allele frequency differences between the haplotype reference panel and the study population. This is similar to the way that population structure can cause problems in GWA studies. Ideally, situations like this are best avoided by sensible study design. If this isn't possible, we recommend quality-control measures to ensure only the most accurately imputed SNPs are used.

**fastPHASE and BIMBAM.** The fastPHASE<sup>8</sup> method can be used to estimate haplotypes and carry out imputation and has recently been incorporated into an association-testing program called BIMBAM<sup>9,10</sup>. The method uses the observation that haplotypes tend to cluster into groups of closely related or similar haplotypes. The model specifies a set of  $K$  unobserved states or clusters that are meant to represent common haplotypes. The  $k$ th cluster is assigned a weight ( $\alpha_{kl}$ ) that denotes the fraction of haplotypes it contains at site  $l$ , with

$$\sum_k \alpha_{kl} = 1 \quad (2)$$

Each cluster also has an associated frequency ( $\theta_{kl}$ ) of allele 1 at each site. Each individual's genotype data is then modelled as an HMM on this state space with transitions between states controlled by a further set of parameters ( $r$ ) at each SNP,

$$P(G_i | \alpha, \theta, r) = \sum_z P(G_i | Z_i, \theta) P(Z_i | \alpha, r) \quad (3)$$

This equation is similar to equation 1 above with  $P(G_i | Z, \theta)$  modelling how likely the observed genotypes are given the underlying states and  $P(Z | \alpha, r)$  modelling

patterns of switching between states, but the states represent clusters rather than reference haplotypes. An analogous model can be used for a set of observed haplotypes so that a likelihood can be written as

$$L(G, H|\alpha, \theta, r) = \prod_i P(G_i|\alpha, \theta, r) \prod_i P(H_i|\alpha, \theta, r) \quad (4)$$

An expectation-maximization algorithm (EM algorithm) is used to fit the model and missing genotypes are imputed conditional on the parameter estimates using the forward-backward algorithm. The authors found that averaging over a set of estimates produced much better results than choosing a single best estimate. Empirical experiments<sup>10</sup> suggest that using  $K = 20$  clusters and  $E = 10$  start points for the EM algorithm represents a practical compromise between speed and accuracy. The model underlying the GEDI<sup>11</sup> approach is very similar to that of fastPHASE.

When imputing untyped SNPs from a reference panel, it was discovered (B.H., unpublished observations) that maximizing the full likelihood  $L(G, H|\alpha, \theta, r)$  resulted in relatively high error rates compared to other methods. Subsequently, it was shown that fixing parameter estimates based only on the likelihood for the set of haplotypes produces lower error rates<sup>11</sup>. This is a similar strategy to that used by IMPUTE v1 in which each cohort individual is independently imputed conditional only on the panel data. IMPUTE v1 has the advantage of not needing to estimate any parameters by using real haplotypes as the models of underlying states. By contrast, fastPHASE uses a much smaller set of states, which speeds up the required HMM calculations but the need to estimate the many parameters of this method can counteract this effect.

**MACH.** MACH uses an HMM model very similar to that used by HOTSPOTTER<sup>5</sup> and IMPUTE. The method can carry out phasing and as a consequence it can be used for imputation. The method works by successively updating the phase of each individual's genotype data conditional on the current haplotype estimates of all the other samples. The model used can be written as

$$P(G_i|D_{-i}, \theta, \eta) = \sum P(G_i|Z, \eta)P(Z|D_{-i}, \theta) \quad (5)$$

in which  $D_{-i}$  is the set of estimated haplotypes of all individuals except  $i$ ,  $Z$  denotes the hidden states of the HMM,  $\eta$  is an 'error' parameter that controls how similar  $G_i$  is to the copied haplotypes and  $\theta$  is a 'crossover' parameter that controls transitions between the hidden states. The parameters  $\eta$  and  $\theta$  are also updated during each iteration based on counts of the number and location of the change points in the hidden states  $Z$  and counts of the concordance between the observed genotypes to those implied by the sampled hidden states.

Imputation of unobserved genotypes using a reference panel of haplotypes,  $H$ , is naturally accommodated in this method by adding  $H$  to the set of estimated haplotypes  $D_{-i}$ . The marginal distribution of the unobserved genotypes can then be estimated from the haplotypes sampled at each iteration. An alternative two-step

approach is also recommended that estimates  $\eta$  and  $\theta$  using a subset of individuals and then carries out maximum-likelihood genotype imputation based on the estimated parameters. By contrast, IMPUTE v1 uses fixed estimates of its mutation rates and recombination maps. Estimating the parameters allows more flexibility to adapt to the data set being analysed. However, it is likely that some parameters will not be estimated well and this will reduce imputation accuracy.

**BEAGLE.** The BEAGLE method<sup>12–14</sup> is based on a graphical model of a set of haplotypes. The method works iteratively by fitting the model to the current set of estimated haplotypes and then resampling new estimated haplotypes for each individual based on the model of fit. The probabilities of missing genotypes are calculated from the model that is fitted at the final iteration. The model is empirical in the sense that it has no parameters that need to be estimated and is applied to a given set of haplotypes in two steps. In the first step, a bifurcating tree that describes a set of haplotypes is constructed from left to right across the set of haplotypes. Once completed, each edge of the tree is weighted by the number of haplotypes that pass along it. In the second step, the tree is pruned to produce a more parsimonious characterization of the data set. At each level of the tree, pairs of nodes are compared in terms of their downstream haplotype frequencies by summing the squared differences of their downstream partial haplotype frequencies; if this number exceeds a threshold, then the nodes are not similar enough to combine. The current recommended threshold was determined empirically from simulated data<sup>12</sup>. Possibly the best way to understand the model is by looking at the small example given in Figure 2 and Table 1 of REF. 12. REF. 15 provides a useful review that contrasts various methods for phasing and imputation.

The BEAGLE model has the property that the graph will have few or many edges in regions in which there is low or high LD respectively. In this way, the model has the attractive property that it can adapt to the local haplotype diversity that occurs in the data. In some sense it can be thought of as a local haplotype-clustering model, similar to fastPHASE, but with a variable number of clusters across a region.

**SNP tagging-based approaches.** Some methods (PLINK<sup>16</sup>, SNPSTAT<sup>17</sup>, UNPHASED and TUNA<sup>18</sup>) carry out imputation using methods based on tag SNP approaches<sup>19–21</sup>. For each SNP to be imputed, the reference data set is used to search for a small set of flanking SNPs that, when phased together with the SNP, leads to a haplotype background that has high LD with the alleles at the SNP. The genotype data from the study and the reference panel are then jointly phased at these SNPs and the missing genotypes in the study are imputed as part of the phasing. The advantage of this approach is that it is simple and quick. The downside is that these approaches generally don't provide as accurate results as other methods because they don't use all the data and the phasing is carried out through a simple multinomial model of haplotype frequencies<sup>22</sup>.

#### Expectation-maximization algorithm

A method for finding maximum-likelihood estimates of parameters in statistical models, in which the model depends on unobserved latent variables. It is an iterative method which alternates between performing an expectation (E) step and a maximization (M) step.



**Imputation in related samples.** The UNPHASED program implements an unpublished method for genotype imputation in nuclear families. This approach has been used to impute sporadic missing SNP genotype data in a study of nuclear families and unrelated individuals with a mixture of HLA and SNP genotypes<sup>23</sup>. A more focused method in which genotypes in founders are imputed down to descendants has also been proposed<sup>24</sup>. As close relatives will share long stretches of haplotypes, the descendants need only be typed at a relatively sparse set of markers for this to work well. Kong *et al.*<sup>25</sup> proposed a related approach in which surrogate parents are used instead of real parents. For each individual, surrogate parents are identified as those who share long stretches of sequence with at least one allele that is identical by state (IBS). Regions in which this occurs are assumed to be identical by descent (IBD) and this estimated relatedness is used to help phase the individuals accurately over long stretches. This approach only works when a sufficient proportion of the population (>1% as a rule of thumb) has been genotyped, but may have useful applications when carrying out imputation if large, densely typed or sequenced cohorts become available. A related idea is used in IMPUTE v2 (REF. 7) in which a 'surrogate family' of individuals is used when updating the phase of a given individual over reasonably long stretches of sequence (typically 5 Mb in practice).

**Comparison between methods.** TABLE 1 summarizes the properties of each of the most popular imputation methods divided into subsections that deal with properties of the reference panels the methods can handle, properties of the study samples, relevant program options and features, computational performance, error rates, and properties and ways of using the output files. [Supplementary information S2](#) (table) is a fuller version of this table, which includes all the methods discussed above.

The sections of TABLE 1 on computational performance and error rate include an updated version of the comparison of the methods IMPUTE (v1 and v2), MACH, fastPHASE and BEAGLE carried out by Howie *et al.*<sup>7</sup>. IMPUTE v2 is the most accurate approach in all of the settings examined but all the methods produce broadly similar performances. The methods are also broadly comparable in terms of computational performance. Several authors<sup>7,14</sup> have noted that the HMM models used by IMPUTE and MACH scale quadratically as the number of haplotypes in the panel increases, but the adaptive haplotype selection approach in IMPUTE v2 (REF. 7) scales linearly with the number of haplotypes in the panel and overcomes this problem.

To examine how the methods might perform on a large reference panel of haplotypes, such as that being generated by the 1000 Genomes Project, we timed IMPUTE v2, fastPHASE and BEAGLE when imputing genotypes using a reference panel of 1,000 haplotypes into a study of samples consisting of 500 and 1,000 individuals. We used HAPGEN<sup>26</sup> to simulate these data sets based on some of the pilot CEU haplotypes from the 1000 Genomes Project in a 5 Mb region on chromosome 10. The haplotype reference contains 8,712 SNPs and the

study sample has genotype data at 872 of these SNPs. The results in TABLE 1 show that IMPUTE v2 is at least twice as fast as both BEAGLE and fastPHASE on this data set.

### Factors that affect imputation accuracy

Most imputation methods produce a probabilistic prediction of each imputed genotype of the form

$$p_{ijk} = P(G_{ij} = k | H, G), k \in \{0, 1, 2\}, \sum_k p_{ijk} = 1 \quad (6)$$

in which  $G_{ij} \in \{0, 1, 2\}$  denotes the genotype of the  $i$ th individual at the  $j$ th SNP.

To assess the quality of predictions and compare methods, genotypes can be masked and then predicted. The most likely predicted genotype above some threshold can be compared with the true genotype and a plot of the percentage discordance versus the percentage of missing genotypes can be constructed for a range of thresholds to illustrate performance. This method was recently used to compare methods using 1,377 UK individuals genotyped on both the Affymetrix 500k SNP chip and the Illumina 550k chip. Genotypes on the Affymetrix chip were combined with the 120 CEU haplotypes to predict the 22,270 HapMap SNPs on chromosome 10 that were on the Illumina chip but not the Affymetrix chip. The error rate of the best-guess genotype for various methods was: BEAGLE (default), 6.33%; BEAGLE (50 iterations), 6.24%; fastPHASE ( $k = 20$ ), 6.07%; fastPHASE ( $k = 30$ ), 5.92%; IMPUTE v1, 5.42%; IMPUTE v2 ( $k = 40$ ), 5.23%; IMPUTE v2 ( $k = 80$ ), 5.16%; MACH, 5.46%, and these results are consistent with other comparisons<sup>27,28</sup>. For the best methods an error rate of 2–3% can be achieved but at the expense of 10% of missing genotypes. Another option involves measuring the squared correlation between the best-guess genotype and the true genotype<sup>14</sup> which can be averaged across SNPs to give a single measure. Another desirable property of imputation methods is that the predicted probabilities they produce should be well calibrated. Most methods in common use have been shown to produce well-calibrated probabilities<sup>2,8,14</sup>.

The imputation accuracy results from Howie *et al.*<sup>7</sup> are specific to a UK population using the CEU HapMap and the Affymetrix 500k chip. The study population, properties of the reference panel and genotyping chip will all influence performance, and performance may vary between rare and common alleles. To illustrate the way in which these factors affect imputation accuracy we took the CEU, YRI and JPT + CHB HapMap 2 haplotype panels and removed a single individual from each. We then used genotypes at SNPs on four chips (Affymetrix 500k, Affymetrix 6.0, Illumina Human660W and Illumina Human1M) to impute masked genotypes not on each chip in that individual, based on the remaining haplotypes in their panel of origin. We also assessed four other panels of haplotypes: a combined CEU + YRI + JPT + CHB panel of 414 haplotypes, which can be used to assess how a larger more diverse set of haplotypes compares with a small, more homogeneous set of haplotypes; a CEU panel rephased without using trio information, using fastPHASE<sup>8</sup> (denoted CEU\_FP) to assess

#### Identical by state

Two or more alleles are identical by state if they are identical. Alleles which are identical by state may or may not be identical by descent owing to the possibility of multiple mutation events.

#### Identical by descent

Two or more alleles are identical by descent if they are identical copies of the same ancestral allele.

#### Best-guess genotype

Most imputation methods provide a probabilistic prediction of the missing genotypes. The best guess genotype is that genotype which has the largest probability.

#### Calibration

The probabilities of events predicted by a probability model are said to be well calibrated if they accurately estimate the proportion of times the events occur. For imputation, a method is well calibrated if genotypes that are predicted with probability  $p$  are correct  $\sim 100p\%$  of the time.

Table 1 | Comparison of imputation methods

Properties	Imputation method				
	IMPUTE v1	IMPUTE v2.2	MACH v1.0.16	fastPHASE v1.4.0 BIMBAM v0.99	BEAGLE v3.2
<b>Reference panels</b>					
Can use a haplotype reference panel?	Yes	Yes	Yes	Yes	Yes
Can use a genotyped reference panel?	No	Yes	Yes	Yes	Yes
Can two haplotype or genotype reference panels be used in the same run?	No	Yes	No	No	No
Reference panels available in correct format	HapMap2 HapMap3 1KGP pilot data	HapMap2 HapMap3 1KGP pilot data	HapMap2 HapMap3 1KGP pilot data	HapMap2	No
<b>Study samples</b>					
Can take genotypes specified with uncertainty?	No	Yes	No	No	Yes
Can accommodate trios and related samples?	No	No	No	No	Trios and duos
Can impute into a study sample of autosomal haplotypes?	Yes	Yes	No	No	Yes
Can impute on the X chromosome?	Yes	Yes	No	No	Yes
<b>Program options and features</b>					
Does phasing as well as imputation?	No	Yes	Yes	Yes	Yes
Can impute sporadic missing genotypes?	No	Yes	Yes	Yes	Yes
Has internal performance assessment?	Yes	Yes	Yes	No	No
Can impute only in a specified interval?	Yes	Yes	No	No	No
Can handle strand alignment between data sets?	Yes	Yes	Yes	No	No
SNP and sample inclusion and exclusion options?	Yes	Yes	No	Yes	Yes
Joint model for imputation and association testing?	No	No	No	No	No
Operating system requirements	Linux, Solaris, Windows, Mac	Linux, Solaris, Windows, Mac	Linux, Windows, Mac	BIMBAM (source code + Windows) fastPHASE (Linux, Solaris, Windows, Mac)	Java executable
<b>Computational performance</b>					
Assessment 1*	43m (1000 Mb)	75m (180 Mb)	105m (80 Mb)	855m (16 Mb)	56m (3100 Mb)
Assessment 2†	---	48m (115m)	---	157m (211m)	104m (234m)
<b>Error rates<sup>§</sup></b>					
Rows correspond to the Scenario A, Scenario B (restricted) and Scenario B (full) data sets	5.42%	5.16%	5.46%	5.92%	6.33%
	---	3.4% (0.86%)	---	5.33% (1.32%)	3.46% (0.93%)
	---	3.4% (0.86%)	---	---	4.01% (1.04%)
<b>Output files</b>					
Genotype posteriors produced?	Yes	Yes	Yes	Yes	Yes
Information measures?	Yes	Yes	Yes	No	Yes
Easiest use of output files to test association	Feed files directly into SNPTEST. Test based on genotype posteriors, dosages or thresholded genotypes	Feed files directly into SNPTEST. Test based on genotype posteriors, dosages or thresholded genotypes	Genotype dosage files can be fed into MACH2DAT or MACH2QTL	BIMBAM can produce file formats used by BIMBAM. fastPHASE out files need to be processed	Best-guess phased haplotypes can be tested in BEAGLE. Processing required to use genotype posteriors or dosage

Properties are shown for the version of the method as listed. \* Imputation of 1377 samples on the Affy500k chip from 120 CEU HapMap2 haplotypes; 7.5 Mb region. Data from REF. 7. † Imputation of 500 (1000) samples genotyped at 872 SNPs from 1000 haplotypes at 8712 SNPs in a 5 Mb region. Timings based on data sets simulated using HAPGEN and the pilot CEU haplotypes from the 1000 Genomes project in a 5 Mb region on chromosome 10. § Error rates from REF. 7, except results for IMPUTE v2 have been updated. Scenario B error rates given are for Illumina SNPs imputed from Affymetrix SNPs. Error rates for Affymetrix SNPs imputed from Illumina SNPs are given in brackets.

the effects of trio phasing on imputation and to assess the effect of reference panel size; a subset of 60 haplotypes from the CEU panel (denoted CEU\_60); and a subset of 120 haplotypes from the JPT + CHB panel (denoted JPT + CHB\_120).

The results of this analysis are described in detail in BOX 4. They show that: across all imputation panels and genotyping chips, imputation error rate increases as the minor allele frequency decreases, which is in line with previous observations<sup>1</sup> that have shown that rare SNPs are more difficult to tag than common SNPs; using a reference panel phased using trio information boosts imputation performance, compared with using a reference panel phased without trio information; and the Illumina chips outperform the two Affymetrix chips in the CEU population, but in the YRI population, the performance of all chips decreases, more so for the Illumina chips. Therefore, the use of tagging methods for chip design can influence the imputation performance. The results also show that the error rate decreases as reference panel size increases<sup>7,14</sup>, and using a combination of CEU, YRI and JPT + CHB haplotype panels can boost the performance of imputation, especially at rare SNPs, compared with using a single haplotype panel.

It is also important to consider the performance of imputation in individuals from populations other than the three main HapMap panels. Huang *et al.*<sup>29</sup> examined the ‘portability’ of the HapMap reference panels for imputation using genome-wide SNP data collected on samples from 29 worldwide populations. When a single HapMap panel was used as the basis for imputation, they found that European populations had the lowest imputation error rates, followed by populations from east Asia, central and south Asia, the Americas, Oceania, the Middle East and Africa. Within Africa, which has high levels of genetic diversity, imputation accuracy using the YRI panel varied substantially. These results indicate that differences in genetic diversity between the study population and the reference panel also influence imputation accuracy.

Huang *et al.*<sup>29</sup> also found that imputation-based mixtures of at least 2 HapMap panels reduced imputation error rates in 25 of the populations. In 11 of the populations, the optimal choice of mixtures was to combine all 3 HapMap populations together as a reference panel. Of these 11 groups (Bedouin, Mozabite, Druze, Basque, Burusho, Daur and Yi), 7 were from Eurasia with some degree of dissimilarity from the HapMap CEU and JPT + CHB panels. The remaining four groups (Melanesian, Papuan, Pima and Colombian) were from Oceania and the Americas. These results can guide the choice of HapMap panels to use, with the caveat that they are specific to the HumanHap550 chip. A related point concerns imputation of admixed individuals. Pasiuic *et al.*<sup>30</sup> have shown that imputation conditional on a local ancestry estimate can be more accurate than unconditional imputation, but the biggest gains in accuracy will occur in admixed individuals from genetically dissimilar populations.

More recently, sets of haplotypes from the HapMap3 project and from the pilot phase of the 1000 Genomes

Project (1KGP) have been made available. The HapMap3 has ten distinct sets of haplotypes and larger numbers of haplotypes in each set. For example, there are 330 CEU haplotypes. This allows more accurate imputation of rarer SNPs but HapMap3 has a smaller set of SNPs than HapMap2. At the time of publication of this Review, there are ~7.7 million SNPs after filtering in the CEU panel of the 1KGP pilot project. This large boost in the number of SNPs allows finer resolution of signals in associated regions<sup>55</sup>. When the 1KGP data is complete, it is likely that this will become the reference set of choice for imputation into GWA study data sets. The large increase in both the number of SNPs and samples will allow more accurate imputation of most SNPs, indels and other structural variants that occur at a frequency above 1%.

**Post-imputation information measures.** Once imputation has been carried out, it is useful to assess the quality of imputed genotypes at SNPs in the absence of any true set of genotypes to compare them to. If the imputation quality is low at a SNP, it may be wise to filter out such SNPs before association testing is performed<sup>31</sup>. Four metrics have been proposed in the literature to assess quality that are designed to lie in the range (0,1) ([Supplementary information S3](#) (box)). A value of 1 indicates that there is no uncertainty in the imputed genotypes whereas a value of 0 means that there is complete uncertainty about the genotypes. All of these measures can be interpreted in the following way: an information measure of  $\alpha$  on a sample of  $N$  individuals indicates that the amount of data at the imputed SNP is approximately equivalent to a set of perfectly observed genotype data in a sample size of  $\alpha N$ .

The MACH  $\hat{r}^2$  measure is the ratio of the empirically observed variance of the allele dosage to the expected binomial variance at Hardy–Weinberg equilibrium. BEAGLE advocates using the  $R^2$  between the best-guess genotype and the allele dosage as an approximation to the  $R^2$  between the best guess-genotype and the true genotype<sup>14</sup>. The IMPUTE software calculates a measure of the relative statistical information about the SNP allele frequency from the imputed data. The SNPTEST program, which is primarily a package to carry out tests of association at SNPs, also calculates a similar relative information measure, but where the parameter of interest is the relevant association parameter of the model of association being fitted. When an additive model is fitted, this measure then has a very strong correlation to the IMPUTE information measure ([Supplementary information S4](#) (figure)).

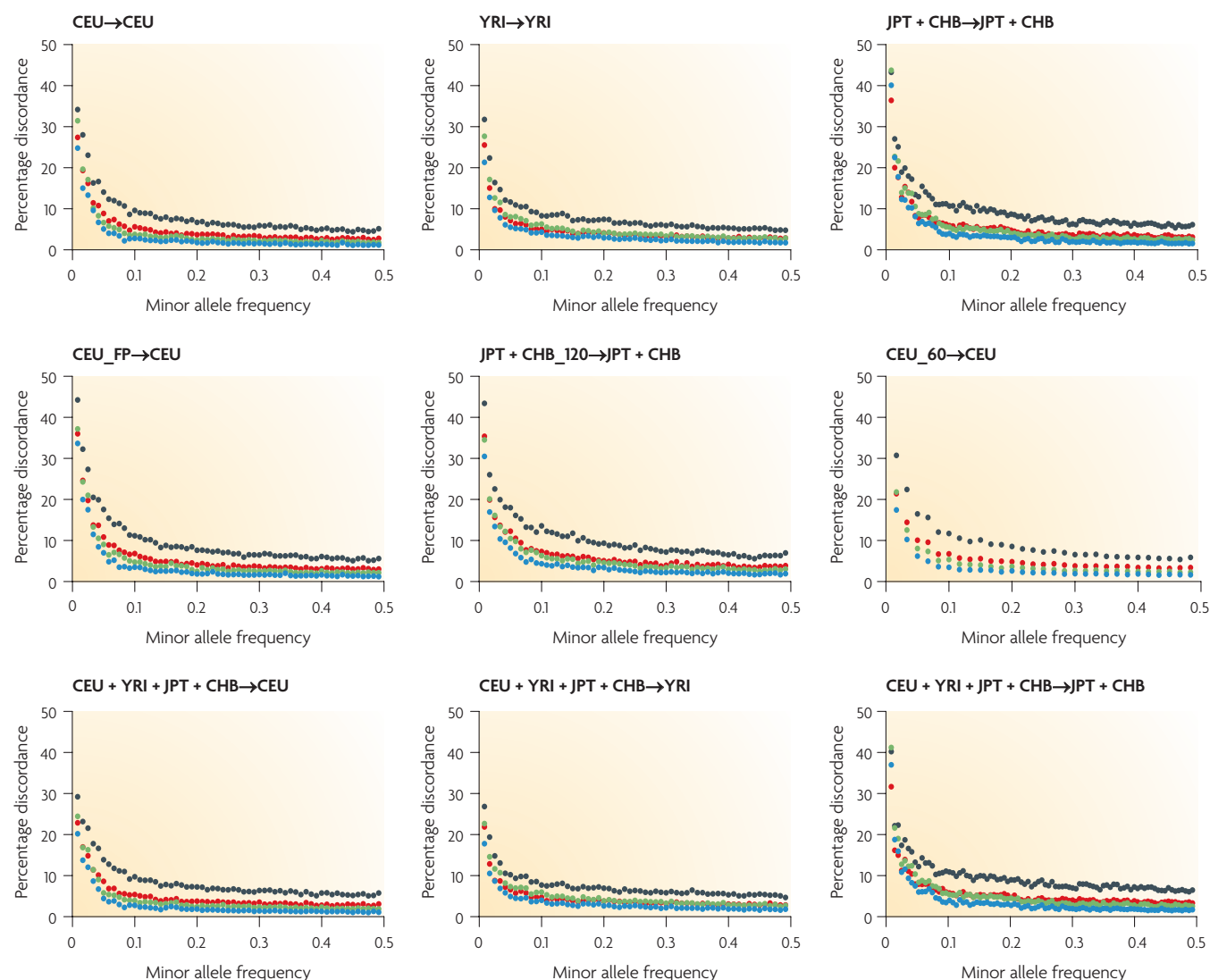
FIGURE 1 shows the MACH, BEAGLE and IMPUTE information measures applied to a simulated imputed data set across a 7 Mb interval on chromosome 22 and shows that the measures are highly correlated, although the MACH measure often goes above 1 and the BEAGLE measure is undefined at almost 3% of SNPs (see also [Supplementary information S4–S6](#) (figures)).

### Association testing using imputed data

The probabilistic nature of imputed SNPs means that testing for association at these SNPs requires some care.



# Box 4 | Factors that affect imputation accuracy

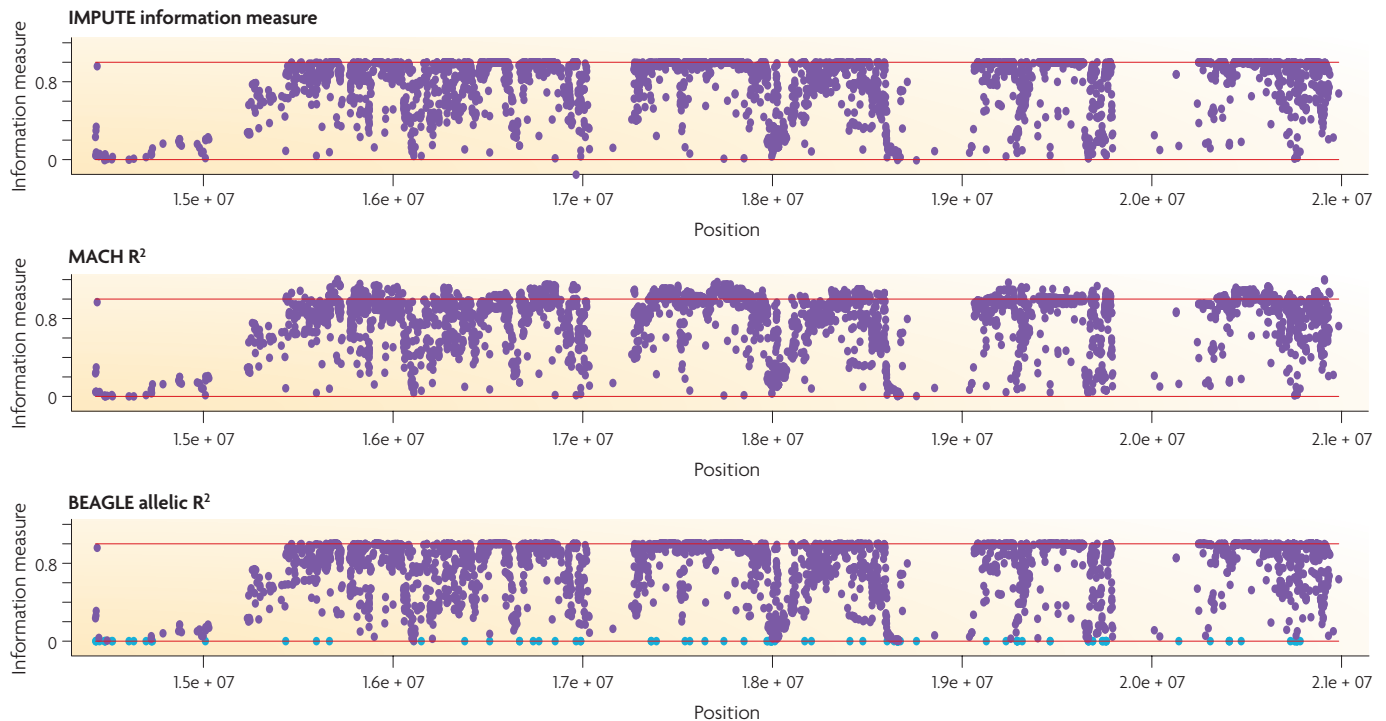


The figures above show an analysis of factors that affect imputation accuracy. Results are stratified by allele frequency and genotyping chip as the percentage discordance of the best-guess imputed genotype call, but only using genotypes that contain at least one copy of the rare allele, to increase focus on performance of imputing the rare allele. Affymetrix 500k results are shown in black, Affymetrix 6.0 in red, Illumina Human660W in green and Illumina Human1M in blue. The top row of graphs shows the results when imputing individuals from the HapMap panels CEU (Utah residents with northern and western European ancestry from the CEPH collection), YRI (Yoruba from Ibadan, Nigeria) and JPT + CHB (Japanese from Tokyo, Japan and Chinese from Beijing, China). The middle row shows the results from the CEU\_FP (CEU repurposed using fastPHASE), JPT + CHB\_120 (a subset of 120 haplotypes from the JPT + CHB panel) and the CEU\_60 (a subset of 60 haplotypes from the CEU panel). The bottom row shows the performance of the combined CEU + YRI + JPT + CHB panel when used to impute genotypes into the single individuals extracted from the CEU, YRI and JPT + CHB panel. Discordance versus missing genotype plots are given in [Supplementary information S8](#) (figure).

The Illumina chips outperform the two Affymetrix chips in the CEU population. In the YRI panel, performance of all chips decreases, more so for the Illumina chips. The difference between chips is larger at rarer SNPs. The use of tagging methods in the design of the various chips is expected to cause these differences.

Comparisons of the JPT + CHB panel with the CEU and YRI panels are confounded by several factors that may affect performance in different directions. First, the JPT + CHB panel was phased without the use of trio information, so will contain more phasing errors than the CEU and YRI panels<sup>52</sup> and may act to increase imputation error rates. In addition, we might expect that increasing SNP density on a chip and the number of SNPs chosen using tagging-based methods will ameliorate the effects of phasing errors in the panel. Also, the JPT + CHB panel contains more haplotypes than the CEU and YRI panels so this should act to improve accuracy. These assertions are supported by the results of the CEU\_FP panel, which show an increase in error rates for the Affymetrix 500k chip at common SNPs. Affymetrix 6.0 and Illumina Human660W chips show smaller increases and the Illumina Human1M chip shows no noticeable change in performance. The JPT + CHB\_120 panel clearly shows the effect of reference panel size. When focusing on imputing the minor allele at rare SNPs, all of the chips show an increase in error rate (of 5–10% for the most rare SNPs) when using the CEU\_FP panel.

The performance of the combined CEU + YRI + JPT + CHB panel is also illustrated in the figure and in [Supplementary information S9, S10](#) (figures). Supplementary information S10 shows that there is a notable boost in performance across all the chips and populations studied here when imputing genotypes at SNPs with a minor allele frequency <5%.



**Figure 1 | Post-imputation information measures.** The plot shows the IMPUTE, MACH and BEAGLE information measures applied to a data set simulated of 1,000 cases and 1,000 controls on chromosome 22 using HAPGEN<sup>26</sup> and the CEU (Utah residents with northern and western European ancestry from the CEPH collection) HapMap2 haplotypes (release 22) in the interval 14–21 Mb. Only genotypes at SNPs on the Affymetrix 500k chip were simulated. IMPUTE was then used to impute all ungenotyped SNPs from the CEU HapMap2 haplotypes. Each of the three metrics are plotted against the base pair position for each imputed SNP. The blue dots in the BEAGLE plot indicate the position of all those SNPs for which the allelic  $R^2$  metric is undefined owing to the most likely genotype call resulting in a monomorphic SNP. Red lines are shown at 0 and 1.

#### Proportional hazards model

A class of survival models in statistics. Survival models relate the time that passes before some event occurs to one or more covariates that may influence that quantity. In a proportional hazards model, the unique effect of a unit increase in a covariate is multiplicative with respect to the hazard rate.

#### Bayesian

A statistical school of thought in which the posterior probability distribution for any unknown parameter or hypothesis given the observed data is used to carry out inference. Bayes theorem is used to construct the posterior distribution using the observed data and a prior distribution, often allowing the incorporation of useful knowledge into the analysis.

Using only imputed genotypes that have a posterior probability above some threshold (or using the best-guess genotype) is a reasonable method of comparing the accuracy across methods but it is not recommended when carrying out association tests at imputed SNPs. Removing genotypes in this way can lead to both false positives and loss of power.

**Frequentist tests.** To fully account for the uncertainty in imputed genotypes, well-established statistical theory for missing data problems can be used<sup>2</sup> (BOX 5). An observed data likelihood is used in which the contribution of each possible genotype is weighted by its imputation probability. A Score test (implemented in SNPTEST) is the quickest way to use this likelihood to test for association, as it attempts to maximize the likelihood in one step by evaluating the first and second derivatives of the likelihood under the null hypothesis and works well when the log-likelihood is close to a quadratic. Small sample size, low allele frequency and increasing genotype uncertainty from imputation all act to degrade this assumption and can lead to the test reporting a spuriously low  $p$ -value. In practice, thresholds on information metrics and allele frequencies to filter out SNPs at which this happens have been used and work well<sup>31–33</sup>. As such SNPs are those likely to have very low power

to detect effects, it is unlikely that has a negative effect on the study. SNPTEST v2 implements an iterative Newton–Raphson scheme and an EM algorithm to maximize the likelihood and improves performance at SNPs for which the Score test does poorly. SNPTEST allows both quantitative and binary traits and can condition on user-specified covariates.

A simpler approach involves using the expected allele count  $e_{ij} = p_{ij1} + 2p_{ij2}$  (also called the posterior mean or allele dosage). These expected counts can be used to test for association with a binary or quantitative phenotype, using a standard logistic or linear regression model, respectively. This method has been shown to provide a good approximation to methods that take the genotype uncertainty into account when the effect size of the risk allele is small<sup>10</sup>, which is the case for most of the common variants found in recent GWA studies. This approach is implemented in the programs MACH2DAT/MACH2QTL, SNPTEST, PLINK and the R package ProbABEL. The ProbABEL package also allows time-to-event phenotypes to be considered using Cox proportional hazards models.

**Bayesian approaches.** Bayesian methods for analysing SNP associations have recently been proposed<sup>2,9,10,34,35</sup> and have advantages over the use of  $p$ -values in power

# Frequentist

A name for the school of statistical thought in which support for a hypothesis or parameter value is assessed using the probability of the observed data (or more extreme data sets) given the hypothesis or value. These theories are usually contrasted with Bayesian models.

and interpretation. Within the Bayesian framework, focus centres on calculation of a Bayes factor (BF), which is the ratio of marginal likelihoods between a model of association ( $M_1$ ) and a null model of no association ( $M_0$ ),

$$BF = \frac{P(Data|M_1)}{P(Data|M_0)} \quad (7)$$

in which the marginal likelihoods are defined by

$$P(Data|M_i) = \int \left( \prod_{j=1}^N \sum_{k=0}^2 P(\Phi|G_{ij} = k, \theta) p_{ijk} \right) P(\theta|M_i) d\theta \quad (8)$$

and  $\theta$  denotes the regression parameters. This can be approximated using a Laplace approximation and a straightforward modification of the likelihood maximization used by frequentist methods ([Supplementary information S7](#) (box)). We have found that this approach (implemented in SNPTEST) is much more stable than when maximizing the likelihood, as the prior acts to regularize the parameter estimation. The expected genotype count can also be used to calculate Bayes factors<sup>10</sup> and is implemented in both BIMBAM and SNPTEST. Stephens and Balding<sup>35</sup> provide an excellent Review of the use of Bayes factors and include a good discussion on the choice of priors. In particular, they discuss the idea of using a mixture of priors to more precisely control beliefs about large effect sizes.

## Box 5 | Frequentist tests that model genotype uncertainty

To fully account for the uncertainty in imputed genotypes, well-established statistical theory for missing data problems can be used. This theory partitions the data structure into observed data  $Y_o$  and missing data  $Y_m$  with  $Y_F = (Y_m, Y_o)$  denoting the full data. Then the observed data likelihood, for a parameter  $\theta$ , which is the log of the full data likelihood integrated over the missing data, is given by

$$l^*(\theta, Y_o) = \log P(Y_o|\theta) = \log \int P(Y_o, Y_m|\theta) dY_m \quad (9)$$

The score and information matrix of this likelihood<sup>53,54</sup> are given by

$$U^*(\theta) = \frac{dl^*(\theta)}{d\theta} = \mathbb{E}_{Y_m|Y_o, \theta} [U(\theta)] \quad (10)$$

and

$$I^*(\theta) = -\frac{d^2 l^*(\theta)}{d\theta^2} = \mathbb{E}_{Y_m|Y_o, \theta} [I(\theta)] - V_{Y_m|Y_o, \theta} [U(\theta)] \quad (11)$$

in which  $U(\theta)$  and  $I(\theta)$  are the full data score and information.

If we let  $\Phi_i$  denote the phenotype of the  $i$ th individual in a study of  $N$  samples with  $N_1$  cases and  $N_2$  controls, then the full data likelihood is

$$P(\Phi_i|G, H, \theta) = \prod_{i=1}^N \sum_{G_{ij}} P(\Phi_i|G_{ij}, \theta) P(G_{ij}|G, H) \quad (12)$$

and the conditional distribution of the missing data given the observed data and the parameters,  $P(Y_m|Y_o, \theta)$ , is given by

$$q_{ijk} = P(G_{ij} = k|\Phi, G, H, \theta) \propto P(\Phi_i|G_{ij} = k, \theta) p_{ijk} \quad (13)$$

For a binary phenotype, it is usual to use a logistic regression model of the form

$$P(\Phi_i|G_{ij}, \theta) = \prod_{i=1}^N p_i^{\Phi_i} (1-p_i)^{1-\Phi_i} \quad (14)$$

in which

$$\theta = (\mu, \gamma) \quad \log \frac{p_i}{1-p_i} = \mu + \gamma G_{ij} \quad p_i = \frac{e^{\mu + \gamma G_{ij}}}{1 + e^{\mu + \gamma G_{ij}}} \quad (15)$$

In this model,  $\mu$  is the baseline log-odds of disease for the 0 genotypes,  $\gamma$  specifies the increase in log-odds due to each copy of the allele coded 1 and  $p_i$  is the probability that individual  $i$  develops the disease. The odds ratios of disease for individuals with genotypes 1 and 2 (relative to individuals with the 0 genotype) are  $e^\gamma$  and  $e^{2\gamma}$ , respectively. This model is multiplicative on the odds scale and additive on the log-odds scale. The score and information matrix for this model are given by

$$U(\theta) = \sum_{i=1}^N (\Phi_i - p_i)(1G_{ij})^T \quad (16)$$

and

$$I(\theta) = \sum_{i=1}^N p_i(1-p_i)(1G_{ij})(1G_{ij})^T \quad (17)$$

Dominant, recessive, heterozygote and general two-parameter models of association can be dealt with similarly.

It may also be the case that the phenotype is quantitative. In this case, an option is to model the phenotype using a normal distribution,

$$\Phi_i|G_{ij}, \theta \sim N(\mu + \gamma G_{ij}, \sigma^2) \quad (18)$$

This model assumes normally distributed residuals, which may not be true. Transformation of the phenotype, which is a mixture of distributions, has been proposed to make this assumption more valid and should work well unless the effect size of a SNP is large<sup>9</sup>. Alternatively, Poisson and gamma regression models could be used as ways of handling discrete phenotypes bounded at 0 and phenotypes in which the error distribution is not symmetric.

One way of carrying out a test of association is to use a Score test, which needs calculations of the observed data score and information matrix only under the null hypothesis,  $H_0: \theta = \theta_0$ . For example, for a binary phenotype, if  $H_0: \gamma = 0$  then  $\theta_0 = (\hat{\mu}, 0)$  in which  $\hat{\mu}$  is the maximum likelihood estimate (MLE) of  $\mu$  with  $\gamma=0$  that is,  $\hat{\mu} = \log(N_1/N_2)$ . The score statistic is given by  $S = (U^*)^2/I^*$ , which has a  $\chi^2_1$  distribution under  $H_0$  ([Supplementary information S7](#) (box)). Alternatively, a Newton–Raphson algorithm can be used to maximize the likelihood directly. This has parameter updates

$$\theta^{t+1} = \theta^t + [I^*(\theta^t)]^{-1} U^*(\theta^t) \quad (19)$$

If this algorithm converges to  $\hat{\theta}$ , then a maximum likelihood ratio test (MLRT) statistic can be used, which has the form

$$S_{MLRT} = 2 \log \left( \frac{l^*(\hat{\theta}, Y_o)}{l^*(\theta_0, Y_o)} \right) \sim \chi^2_1 \text{ under } H_0 \quad (20)$$

Another option is a Wald test, which has the form

$$S_{Wald} = \frac{\hat{\theta}_\gamma^2}{I^*(\hat{\theta}_{TT})} \sim \chi^2_1 \text{ under } H_0 \quad (21)$$

assuming  $\hat{\theta} \sim N(\theta_0, I^*(\hat{\theta}))$  under  $H_0$ .

A further alternative is to use an expectation-maximization (EM) algorithm, which is guaranteed to converge to a local maximum of the likelihood, under reasonable regularity conditions for the likelihood ([Supplementary information S7](#)). All of these methods are implemented in SNPTEST v2.

Supplementary information S7 describes how this can also be achieved using a  $t$ -distribution prior and also discusses a method of setting priors for quantitative trait models. Currently, only SNPTEST can calculate Bayes factors for binary traits conditional on a set of covariates.

In the context of fine-mapping in which multiple SNPs in a gene or region may play a part in the underlying causal model, it is desirable to consider models that allow multiple SNPs. The BIMBAM approach<sup>9</sup> combines imputation with such an approach and can produce posterior probabilities of association for each SNP and also on the number of associated SNPs in the region.

**Bayes factors versus  $p$ -values.** At directly genotyped SNPs, Bayes factors and  $p$ -values can be made equivalent in the sense that they give the same ranking of SNPs<sup>34</sup>, but this occurs for a particular choice of prior in which the prior variance of the effect size increases as minor allele frequency decreases (or as the information at the SNP about the effect size parameter decreases). This prior assumes larger effects at rarer SNPs which may be a biologically reasonable assumption. At imputed SNPs, the level of uncertainty also influences the amount of information there is about the effect size parameter. To make Bayes factors give the same ranking of SNPs as  $p$ -values, we would need to allow the prior variance to increase as the amount of imputation uncertainty increases which makes no sense<sup>35</sup>. So even when adopting a prior that depends upon allele frequency, Bayes factors and  $p$ -values will not give the same ranking at imputed SNPs. In practice, studies have tended to filter out SNPs with low information so it seems unlikely that a reanalysis of studies using Bayes factors will result in very different outcomes but, as we probe ever rarer SNPs based on imputation from the 1000 Genomes Project data, it may become more important to take care of these details.

**Joint imputation and testing.** The two-stage imputation then testing approach may underestimate effect sizes as genotypes are effectively imputed under the null model. The SNPMSTAT<sup>17</sup> and UNPHASED approaches allow joint imputation and testing in a single model. A comparison of the joint approach (using SNPMSTAT) to a two-stage approach (using IMPUTE and then SNPTEST) on three different data sets suggested that the improved performance of imputation gained by using a method that uses as much flanking genotype data as possible (like

IMPUTE, MACH, fastPHASE and BEAGLE) outweighs the advantage of joint imputation and testing<sup>36</sup>.

## Perspectives and future directions

It seems likely that genotype imputation will continue to play an important part in the analysis of GWA studies over the next few years, as researchers apply the approach to an increasing set of diseases and traits and work together to combine cohorts through meta-analysis. The main factor that will influence the precise way in which imputation is used will be the increasing availability of next-generation sequencing data. Such data will allow researchers to assess many more SNPs as well as short indels and CNVs.

One public resource of such data will be the 1000 Genomes Project. Imputation is currently being used both within the data to reconstruct genotypes from the low-coverage sequencing reads and will also be used to impute from the data into other cohorts. Compared to HapMap2, the number of SNPs, the number of haplotypes and the number of populations will increase notably. This resource will include many more SNPs with a frequency of 1–5% that can be imputed. This may be key if rarer variation is an important part of the aetiology of a given trait. The availability of haplotypes on a larger set of populations should lead to an improvement in imputation in populations that are not well matched to the CEU, YRI and CHB + JPT haplotype sets in HapMap2.

The challenges for imputation methods will be in using the larger, more diverse set of haplotypes available for imputation. As any haplotype estimates produced from the 1000 Genomes Project data may have more inherent uncertainty than the HapMap2 haplotypes, owing to the low-coverage sequencing used and the larger number of rare SNPs, it may be important to take this into account when imputing from this data. Along these lines, both IMPUTE v2 and BEAGLE already offer the ability to accept genotypes estimated with uncertainty when carrying out imputation and phasing.

Care may also be needed when analysing rare imputed SNPs. It is well known that the asymptotic theory used by frequentist association tests breaks down at rarer SNPs, which means that  $p$ -values may not be well calibrated. In addition, the subtle effects of population structure when analysing rare variants will need to be handled carefully. The danger here is that a small number of extra rare alleles in cases or controls owing to population structure may lead to false positives.

1. Frazer, K., Ballinger, D., Cox, D., Hinds, D., Stuve, L. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
2. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genet.* **39**, 906–913 (2007).
3. Stephens, M. & Donnelly, P. Inference in molecular population genetics. *J. R. Statist. Soc. B* **62**, 605–635 (2000).
4. Fearnhead, P. & Donnelly, P. Estimating recombination rates from population genetic data. *Genetics* **159**, 1299–1318 (2001).
5. Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233 (2003).
6. Rabiner, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77**, 257–286 (1989).
7. 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). **This paper describes the IMPUTE v2 method and carries out a comprehensive evaluation of several methods. This reference should be read as the follow-on from Reference 2, which describes IMPUTE v1.**
8. 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).
9. Servin, B. & Stephens, M. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet.* **3**, e114 (2007). **The paper that describes the BIMBAM method for Bayesian multi-SNP and single SNP analysis using imputed data. Should be read together with Reference 8, which describes fastPHASE.**
10. Guan, Y. & Stephens, M. Practical issues in imputation-based association mapping. *PLoS Genet.* **4**, e1000279 (2008).



11. Kennedy, J., Mandoiu, I. & Pasaniuc, B. Genotype error detection using hidden Markov models of haplotype diversity. *J. Comput. Biol.* **15**, 1155–1171 (2008).
12. Browning, S. Multilocus association mapping using variable-length Markov chains. *Am. J. Hum. Genet.* **78**, 903–913 (2006).
13. Browning, S. & Browning, B. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
14. Browning, B. & Browning, S. 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).
15. Browning, S. Missing data imputation and haplotype phase inference for genome-wide association studies. *Hum. Genet.* **124**, 439–450 (2008).
- References 12–15 are a series of papers that describe the model underlying the BEAGLE method.**
16. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
17. Lin, D., Hu, Y. & Huang, B. Simple and efficient analysis of disease association with missing genotype data. *Am. J. Hum. Genet.* **82**, 444–452 (2008).
18. Nicolae, D. Testing untyped alleles (TUNA)-applications to genome-wide association studies. *Genet. Epidemiol.* **30**, 718–727 (2006).
19. Johnson, G. *et al.* Haplotype tagging for the identification of common disease genes. *Nature Genet.* **29**, 235–237 (2001).
20. Evans, D., Cardon, L. & Morris, A. Genotype prediction using a dense map of SNPs. *Genet. Epidemiol.* **27**, 375–384 (2004).
21. De Bakker, P. *et al.* Efficiency and power in genetic association studies. *Nature Genet.* **37**, 1217–1223 (2005).
22. Excoffier, L. & Slatkin, M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* **12**, 921–927 (1995).
23. Pastorino, R. *et al.* Association between protective and deleterious HLA alleles with multiple sclerosis in Central East Sardinia. *PLoS ONE* **4**, e6526 (2009).
24. Burdick, J., Chen, W., Abecasis, G. & Cheung, V. *In silico* method for inferring genotypes in pedigrees. *Nature Genet.* **38**, 1002–1004 (2006).
25. Kong, A. *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature Genet.* **40**, 1068–1075 (2008).
26. Spencer, C. C. A., Su, Z., Donnelly, P. & Marchini, J. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet.* **5**, e1000477 (2009).
27. Pei, Y., Li, J., Zhang, L., Papasian, C. & Deng, H. Analyses and comparison of accuracy of different genotype imputation methods. *PLoS ONE* **3**, e3551 (2008).
28. Hao, K., Chudin, E., McElwee, J. & Schadt, E. E. Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies. *BMC Genet.* **10**, 27 (2009).
29. Huang, L., Li, Y., Singleton, A., Hardy, J., Abecasis, G. *et al.* Genotype-imputation accuracy across worldwide human populations. *Am. J. Hum. Genet.* **84**, 235–250 (2009).
- A useful reference that illustrates the performance of imputation in a range worldwide human populations when using the HapMap 2 reference panels.**
30. Pasaniuc, B., Sankararaman, S., Kimmel, G. & Halperin, E. Inference of locus-specific ancestry in closely related populations. *Bioinformatics* **25**, 213–221 (2009).
31. Zeggini, E. *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature Genet.* **40**, 638–645 (2008).
32. Zeggini, E. *et al.* Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316**, 1336–1341 (2007).
- One of the earliest examples of the use of imputation in meta-analysis. This paper combined three GWA studies and was able to identify several novel associations.**
33. Lindgren, C. M. *et al.* Genome-wide association scan meta-analysis identifies three loci influencing adiposity and fat distribution. *PLoS Genet.* **5**, e1000508 (2009).
34. Wakefield, J. Bayes factors for genome-wide association studies: comparison with p-values. *Genet. Epidemiol.* **33**, 79–86 (2009).
35. Stephens, M. & Balding, D. Bayesian statistical methods for genetic association studies. *Nature Rev. Genet.* **10**, 681–690 (2009).
- An excellent Review on the subject of using Bayesian statistical methods in association studies with a particular focus on the calculation, choice of priors and the interpretation of single SNP Bayes factors.**
36. Marchini, J. & Howie, B. Comparing algorithms for genotype imputation. *Am. J. Hum. Genet.* **83**, 535–539 (2008).
37. Stephens, M., Smith, N. & Donnelly, P. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68**, 978–989 (2001).
38. Carlson, C. *et al.* Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.* **74**, 106–120 (2004).
39. Elston, R. & Stewart, J. A general model for the genetic analysis of pedigree data. *Hum. Hered.* **21**, 523–542 (1971).
40. Lander, E. & Green, P. Construction of multilocus genetic linkage maps in humans. *Proc. Natl. Acad. Sci. USA* **84**, 2363–2367 (1987).
41. Cooper, J. *et al.* Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nature Genet.* **40**, 1399–1401 (2008).
42. Houlston, R. *et al.* Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nature Genet.* **40**, 1426–1435 (2008).
43. De Jager, P. *et al.* Meta-analysis of genome scans and replication identify *CD6*, *IRF8* and *TNFRSF1A* as new multiple sclerosis susceptibility loci. *Nature Genet.* **41**, 776–82 (2009).
44. Loos, R. J. F. *et al.* Common variants near *MC4R* are associated with fat mass, weight and risk of obesity. *Nature Genet.* **40**, 768–775 (2008).
45. De Bakker, P. *et al.* Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum. Mol. Genet.* **17**, R122–R128 (2008).
46. Zollner, S. & Pritchard, J. Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics* **169**, 1071–1092 (2005).
47. Minichiello, M. & Durbin, R. Mapping trait loci by use of inferred ancestral recombination graphs. *Am. J. Hum. Genet.* **79**, 910–922 (2006).
48. Su, Z., Cardin, N., Wellcome Trust Case Control Consortium, Donnelly, P. & Marchini, J. A Bayesian method for detecting and characterizing allelic heterogeneity and boosting signals in genome-wide association studies. *Stat. Sci.* **24**, 430–450 (2009).
49. Browning, B. & Browning, S. Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. *Genet. Epidemiol.* **31**, 365–375 (2007).
50. Leslie, S., Donnelly, P. & McVean, G. A statistical method for predicting classical HLA alleles from SNP data. *Am. J. Hum. Genet.* **82**, 48–56 (2008).
51. Browning, B. L. & Yu, Z. Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *Am. J. Hum. Genet.* **85**, 847–861 (2009).
52. Marchini, J. *et al.* A comparison of phasing algorithms for trios and unrelated individuals. *Am. J. Hum. Genet.* **78**, 437–450 (2006).
53. Louis, T. A. Finding the observed information matrix when using the EM algorithm. *J. Royal Stat. Soc. B* **44**, 226–233.
54. Little, R. J. A. & Rubin, D. B. *Statistical Analysis with Missing Data* 2nd edn (Wiley, Hoboken, 2002).
55. Liu, J. Z. *et al.* (2010) Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nature Genet.* **42**, 436–440 (2010).

# Acknowledgements

B.N.H. was funded by a National Science Foundation Graduate Research Fellowship and the Overseas Research Students Awards Scheme. J.M. acknowledges support from the Medical Research Council.

# Competing interests statement.

The authors declare no competing financial interests.

# FURTHER INFORMATION

**1000 Genomes Project:** <http://www.1000genomes.org>  
**BEAGLE:** <http://www.stat.auckland.ac.nz/~bbrowning/beagle/beagle.html>  
**fastPHASE and BMBAM:** <http://stephenslab.uchicago.edu/software.html>  
**GEDI:** [http://dna.engr.uconn.edu/?page\\_id=81](http://dna.engr.uconn.edu/?page_id=81)  
**IMPUTE v1 and v2, SNPTTEST and HAPGEN:** <http://www.stats.ox.ac.uk/~marchini/#software>  
**MACH, MACH2DAT and MACH2QTL:** <http://www.sph.umich.edu/csg/abecasis/MACH>  
**PLINK:** <http://pngu.mgh.harvard.edu/~purcell/plink>  
**ProbABEL:** <http://mga.bionet.nsc.ru/~yurii/ABEL>  
**SNPMSTAT:** <http://www.bios.unc.edu/~lin/software/SNPMStat>  
**TUNA:** <http://www.stat.uchicago.edu/~wen/tuna>  
**UNPHASED:** <http://homepages.lshrm.ac.uk/frankdudbridge/software/unphased>

# SUPPLEMENTARY INFORMATION

See online article: S1 (box) | S2 (table) | S3 (box) | S4 (figure) | S5 (figure) | S6 (figure) | S7 (box) | S8 (figure) | S9 (figure) | S10 (figure) | S11 (table)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF