nature genetics

A new multipoint method for genome-wide association studies by imputation of genotypes

Jonathan Marchini[1,2], Bryan Howie[1,2], Simon Myers[1], Gil McVean[1] & Peter Donnelly[1]

Genome-wide association studies are set to become the method of choice for uncovering the genetic basis of human diseases. A central challenge in this area is the development of powerful multipoint methods that can detect causal variants that have not been directly genotyped. We propose a coherent analysis framework that treats the problem as one involving missing or uncertain genotypes. Central to our approach is a model-based imputation method for inferring genotypes at observed or unobserved SNPs, leading to improved power over existing methods for multipoint association mapping. Using real genome-wide association study data, we show that our approach (i) is accurate and well calibrated, (ii) provides detailed views of associated regions that facilitate follow-up studies and (iii) can be used to validate and correct data at genotyped markers. A notable future use of our method will be to boost power by combining data from genome-wide scans that use different SNP sets.

It has been known for over 10 years that genome-wide association studies may be a powerful alternative to more traditional family-based linkage studies for mapping the genetic variants that underlie common human diseases[1]. It has taken the Human Genome Project, comprehensive SNP databases, substantial catalogs of human haplotype variation[2], extensive case series collections and technological advances in genotyping for these studies to become a reality.

Current genome-wide association studies assay a very dense set of markers (>100,000) across the genome in individuals affected and unaffected by a disease using one of the commercially available genotyping chips. The simplest analysis strategy involves carrying out a test of association at each assayed SNP. As the set of SNPs on the chip is unlikely to include the true causal variant, we can think of this approach as using the markers on the chip as predictors of possible untyped disease variants. It is widely accepted that this approach will not be the most powerful for studies of this sort[3]. A major challenge in this field is gaining added value for the analysis by combining information across markers and using existing catalogs of variation such as HapMap. Various so-called 'multipoint approaches' have been suggested in the literature to address the first of these concerns, but there is currently no consensus on the best approach.

In this paper, we suggest a coherent framework for thinking about this problem and then illustrate this with a number of different applications. The main idea behind our approach is to think of the problem as one involving a combination of observed data and missing data, where the core aim is to predict (or 'impute') the missing data based upon the observed data. All multipoint methods can be thought of in terms of this prediction aspect, but many are not routinely described as such. Informally, we use (i) data from the SNPs genotyped in our study, (ii) the HapMap data and estimates of

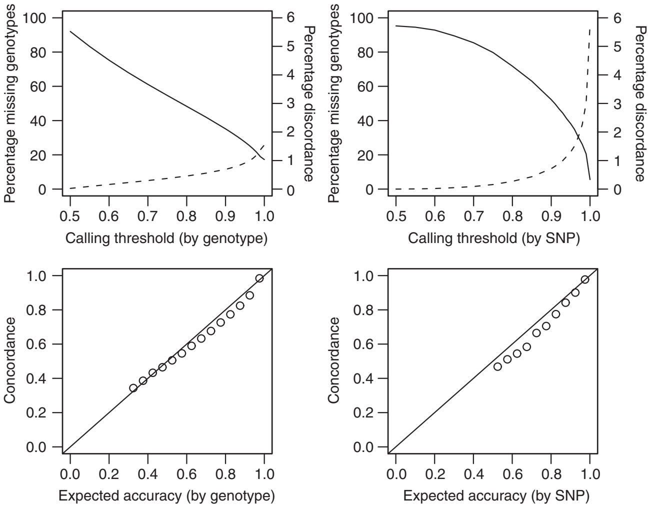

Figure 1 Accuracy and calibration of imputed genotypes. The upper left panel shows the discordance (solid line) and missing data rate (dashed line) for different calling thresholds applied to the imputed genotypes one at a time (see text for details). The lower left plot shows how well the probabilities estimated by the method are calibrated. The plot shows the predicted accuracy of the genotype calls versus an estimate of the actual accuracy measured as concordance with the Illumina calls. The upper right and lower right panels show the discordance, missing rate and calibration plots, where the calling threshold is applied on a per-SNP basis to the average maximum posterior genotype call probability.

[1]Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK. [2]These authors contributed equally to this work. Correspondence should be addressed to P.D. (donnelly@stats.ox.ac.uk).

Received 30 March; accepted 30 May; published online 17 June 2007; doi:10.1038/ng2088

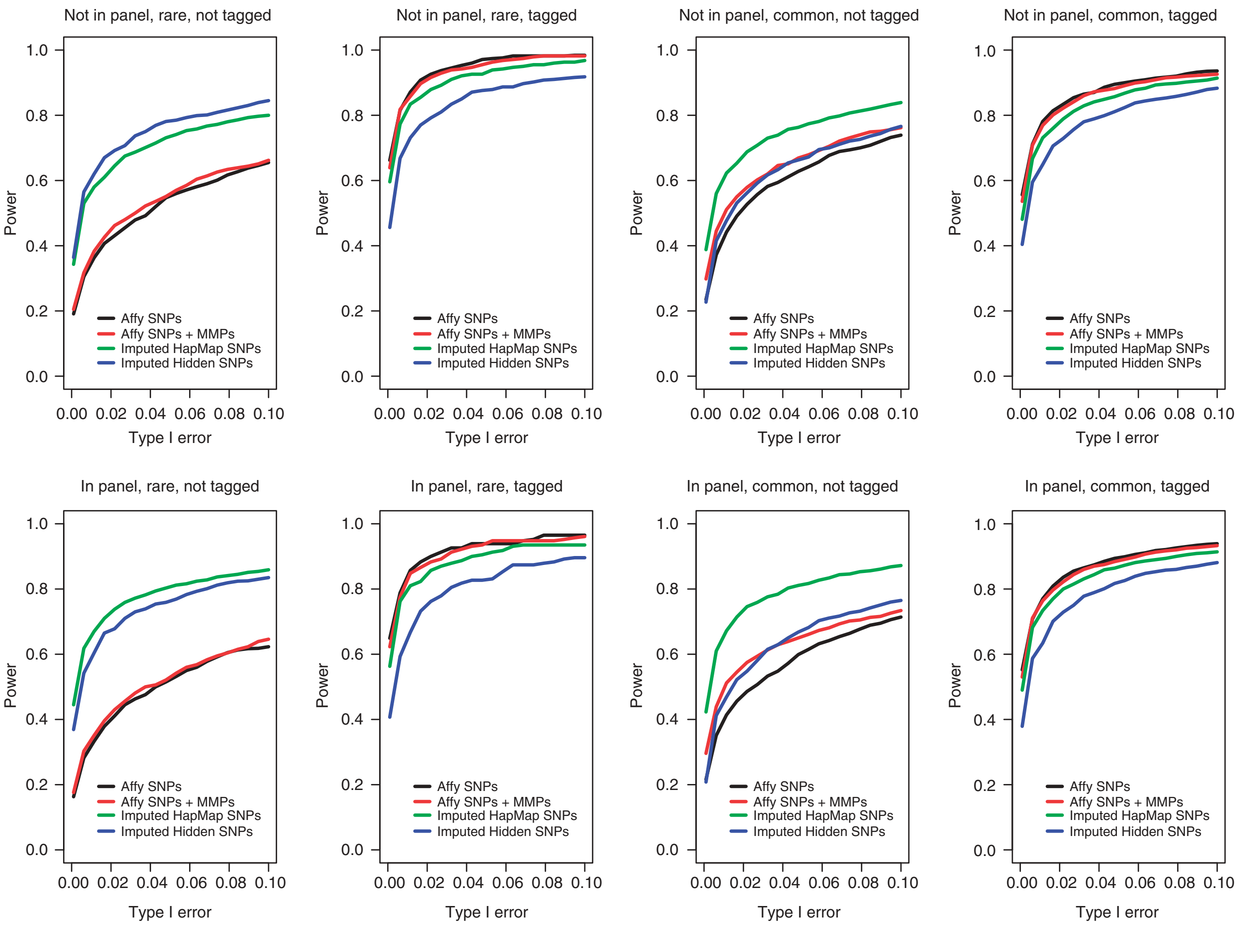

Figure 2 Power versus region-wide type I error for the mapping methods described in the main text, based on simulating case-control data sets conditional upon the haplotype data in the ten ENCODE regions. To measure the power of each method, we compared the distribution of the maximum Bayes factor across each simulated case-control data set with the distribution using the set of null data sets. This takes into account the different numbers of Bayes factors used in each approach and hence corrects for multiple comparisons. The eight plots break down the power according to the properties of the causal SNP. For example, the upper left plot shows the power for causal SNPs that are not in the pseudo-HapMap panel, have a low allele frequency (MAF < 5%) and are not tagged by a multi-marker prediction (MMP) using the Affymetrix ('Affy') SNPs in the ENCODE regions. The properties of the causal SNPs are indicated by the title of each plot. As described in the main text, 'hidden' SNPs are those for which no data are available in the HapMap or in the disease study.

the fine-scale recombination map across the genome and (iii) a population genetics model to simulate or impute genotypes at SNPs not assayed in our study (see Methods). These '*in silico*' genotypes can then be used as if the SNPs involved were directly genotyped. For example, association with disease can be tested at a much finer grid of locations across the genome by directly testing this much larger set of SNPs, and data at imputed SNPs can be combined directly with data from other studies that use different genotyping chips.

A key feature of our approach is our use of an approximate population genetics model that gives more weight to genotypes that are consistent with the local patterns of linkage disequilibrium (LD). This approach has advantages over other multipoint methods: the population genetics approach uses information from all markers in LD with an untyped SNP, but in a way that decreases with genetic distance from the SNP being imputed, thus avoiding the decisions faced in some other prediction methods, such as to how many markers to use, how to use them or over what physical distance to define haplotypes for haplotype analyses[4,5].

RESULTS

Imputation accuracy

We validated our approach using control data from the Wellcome Trust Case Control Consortium (WTCCC)[6]. Specifically, we assessed the accuracy of imputation using individuals from the 1958 British Birth Cohort, who were typed at approximately 500,000 SNPs on the Affymetrix GeneChip Mapping Array Set (the 'Affymetrix 500K chip') and separately typed on a custom Illumina chip at approximately 15,000 largely nonsynonymous SNPs. We used a filtered data set of 1,444 individuals[6] and considered 10,180 autosomal SNPs typed on the Illumina chip that were also polymorphic in the CEU HapMap sample. We used the called Affymetrix genotypes at SNPs passing WTCCC quality control filters and having minor allele frequencies (MAFs) >1% for these individuals and imputed genotypes, as described below, for each of the 10,180 Illumina SNPs. (For Illumina SNPs that were also on the Affymetrix chip, we ignored the Affymetrix data at that SNP when doing the imputation.) Our imputation method outputs probabilities associated with each possible genotype call for each individual. The comparisons showed

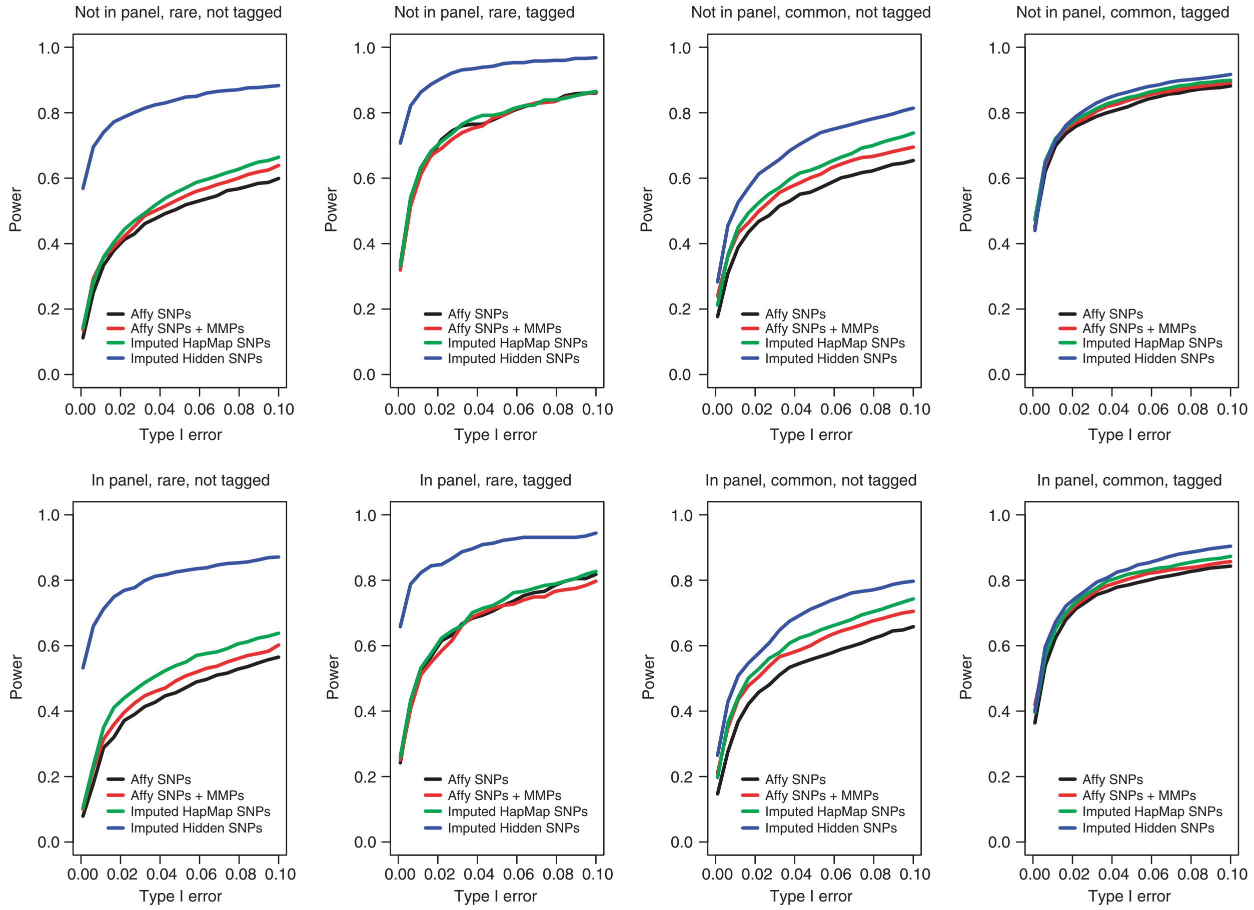

Figure 3 Power versus region-wide type I error for mapping methods described in the main text, based on simulating case-control data sets conditional upon the haplotype data in the ten ENCODE regions. To measure the power of each method, we compared the distribution of the region Bayes factor across each simulated case-control data set with the distribution using the set of null data sets. This takes into account the different numbers of Bayes factors used in each approach and hence corrects for multiple comparisons. The eight plots break down the power according to the properties of the causal SNP. For example, the upper left plot shows the power for causal SNPs that are not in the pseudo-HapMap panel, have a low allele frequency (MAF $<$ 5%) and are not tagged by a multi-marker prediction (MMP) using the Affymetrix ('Affy') SNPs in the ENCODE regions. The properties of the causal SNPs are indicated by the title of each plot. As described in text, 'hidden' SNPs are those for which no data are available in the HapMap or in the disease study.

the imputed genotypes to be accurate. For example, if all genotypes with a maximum posterior genotype probability of greater than 0.95 are considered, the agreement with the Illumina genotypes is 98.4%. In addition, our method is well calibrated in that its estimates of uncertainty are also accurate (**Fig. 1**).

The power of imputation-based association tests

Imputed genotypes may be used in many different downstream analyses. A notable application is in testing for association in genome-wide studies. A key question is what gain we achieve using imputed genotypes over single-locus approaches in detecting associations. Similar imputation approaches[7,8] have been shown to be useful in other contexts, but none has attempted to quantify the potential gains in power in association studies. To answer this question, we compared the power of our approach with other mapping methods through a previously suggested strategy[5] of creating case-control panels using empirical genotype data from the ten ENCODE regions analyzed as part of the HapMap project. We took each of the 9,842 segregating SNPs in the ENCODE regions of CEU data in turn and simulated a case-control data set using that SNP as the causal SNP (see Methods). The effect size of the causal SNP was set to achieve a power of 95% at a nominal *P* value of 0.01. These data sets were then thinned to include only those SNPs found on the Affymetrix 500K chip in the ENCODE regions in order to provide an assessment of the power of the different approaches in a realistic situation. As in other studies[2], the ENCODE data were also thinned to produce pseudo-HapMap panels of data for use in the various multipoint approaches. (The ENCODE regions have a higher SNP density in HapMap than the rest of the genome, and this thinning is required to allow extrapolation to the remainder of the genome.) An additional set of 'null' data sets was simulated under a model of no association to allow empirical assessment of Type I error. Note that in using a single statistic for each region, our approach automatically handles multiple comparisons.

We compared four different strategies for detecting the causal variants: (i) a simple single-SNP approach that tests only the SNPs on the Affymetrix chip, (ii) an established multipoint approach in which single-SNP tests are augmented with multimarker prediction

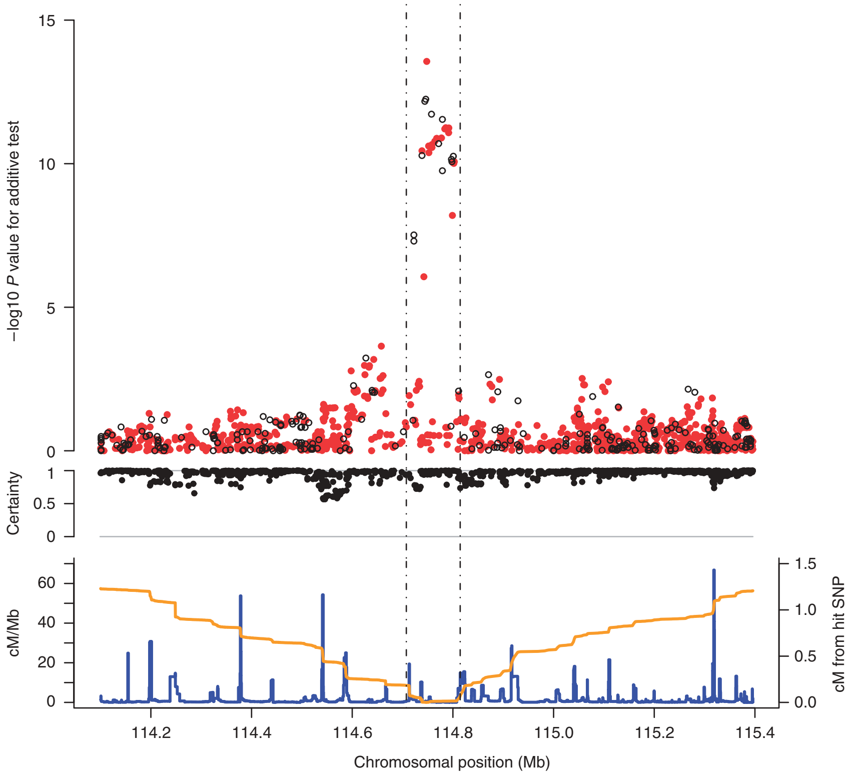

Figure 4 Results of imputing SNPs in the region of the *TCF7L2* gene from the WTCCC data. The upper part of plot shows the $-\log_{10}$ *P* value for the additive model versus a model of no association. The *P* values were calculated using called genotypes (black circles) and imputed genotypes (red circles), called at a threshold of 0.9. The middle panel shows a measure of certainty for each SNP, which is defined as the average maximum posterior genotype call probability. The lower panel shows the fine-scale recombination rate across the region (blue) and the cumulative recombination rate measured away from the most highly associated genotyped SNP (orange). The vertical dashed lines on the plot delineate the main region of association. 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.

(MMP) tests derived from the pseudo-HapMap panel[5], (iii) single-SNP tests at the Affymetrix SNPs plus tests at imputed pseudo-HapMap panel SNPs and (iv) tests on imputed genotypes at a grid of 100 hypothesized SNPs completely unobserved in the ENCODE data set. The inferred genotypes are sometimes (appropriately) uncertain, so we found it necessary to develop tests of association (both frequentist and bayesian) that take account of the uncertainty in the genotypes we impute (see Methods). For each of the above approaches, we calculated Bayes factors at each tested SNP (see Methods).

To assess the power of each method, we considered two different summaries of the evidence of association in a given region. The first summary is the maximum of the Bayes factors in each region[3,5], and the second is a region Bayes factor (see Methods). We compared the distribution of each of these statistics across the simulated case-control data sets with their distributions from the set of null simulations we carried out.

To gain a detailed understanding of how the performance of the various methods differ, we cross-classified causal SNPs according to (i) whether they were common (MAF $\geq$ 5%) or rare (MAF $<$ 5%), (ii) whether they were in the pseudo-HapMap panel and (iii) whether the SNP was tagged by a chip SNP or an MMP (that is, whether it had $r^2 \geq 0.8$ with either a SNP on the chip or an MMP). **Figures 2** and **3** show the results for the maximum and region Bayes factor summary statistics in each of the eight resulting categories. Both of our methods provided a clear boost in power for causal SNPs that were not tagged. The effect was more pronounced for rare SNPs, which in general are harder to tag using a small number of surrogate SNPs and are better predicted by the extended haplotypes upon which the SNP mutation resides. As our methods have no arbitrary window size or tag set size but rather make use of all the surrounding SNP data modulated by the local recombination rate, we are able to predict these SNPs better. For tagged variants, the difference between methods is generally small, but overall the new imputation-based methods provide an increase in power over both the single-SNP tests and the MMP tests (**Supplementary Figs. 1** and **2** online).

We were interested to find that the relative performance of our two methods depends on which test statistic is used. Our method of imputing completely unobserved SNPs tends to produce elevated Bayes factors in an extended region around the causal SNP, whereas our method of imputing HapMap SNPs tends to produce larger signals at a smaller set of locations. The region Bayes factor reduces to the average of the Bayes factors across a region, so it tends to produce a large signal when there are many reasonably elevated signals and thus works well in combination with our method of imputing unobserved SNPs. In contrast, the maximum Bayes factor will perform well when the signal is large and localized rather than small and extended and thus is more suited to the method of directly imputing known SNPs. This difference is clear for rare, untagged causal SNPs, which tend to produce elevated signals over the length of the extended haplotype upon which the causal SNP resides. Most genome-wide association studies (GWAs) currently underway will not be well powered for rare causal SNPs (even with the improvements in power offered by our approach). For such studies, our recommendation would be[6] to impute HapMap SNPs (and to favor analyses using maximum Bayes factors over appropriate regions).

Application to real genome-wide data

We have applied this method to all seven GWAs carried out as part of the WTCCC study, with full results and detailed discussion reported elsewhere[6]. As an illustration, we applied our approach to the WTCCC data in the region of the known type 2 diabetes gene *TCF7L2* (ref. 9) (**Fig. 4**). The imputation was based on 1,924 type 2 diabetes cases and 2,938 control subjects at all genotyped SNPs that passed WTCCC quality control filters and that had MAF $>$ 1%.

We imputed data at all Phase II HapMap SNPs in the region. The figure shows test statistics at both imputed SNPs (red circles) and genotyped SNPs (black circles). The imputation provides a much

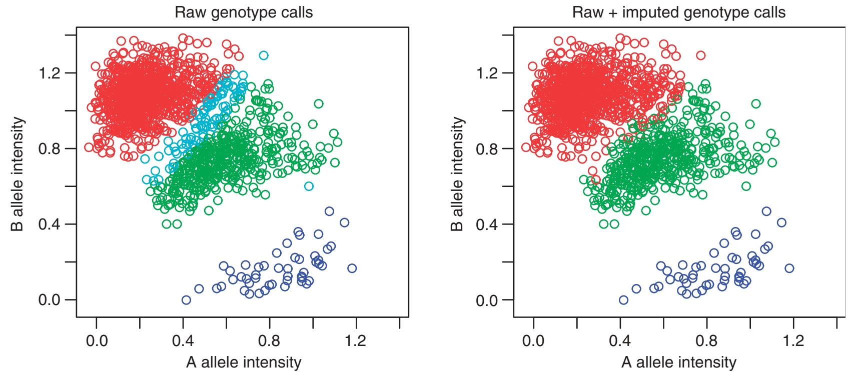

Figure 5 Imputing missing data at genotyped SNPs. The left panel shows the normalized intensity data for a SNP genotyped using both the Affymetrix chip and the custom Illumina chip in the 1958 birth cohort of the WTCCC study. The *x* and *y* axes on this plot denote the intensity measurements for the two alleles (A and B, respectively) at the SNP. Each point represents the measurement for a single individual. The points in the left panel are colored according to the genotype calls made by the algorithm used by the WTCCC project (CHIAMO) using a calling threshold of 0.9 on the posterior probability of genotype calls (blue denotes AA, green denotes AB, red denotes BB and light blue denotes missing). The right panel shows the non-missing CHIAMO genotype calls plus the imputed missing calls. The imputed data agreed 100% with Illumina calls at both missing and non-missing genotypes.

more detailed view of the associated region. The results from imputed SNPs are useful in (i) assessing strength of signal within the region; (ii) providing a wider range of SNPs for follow up and (iii) indicating possible locations for causal variants. For example, there is a substantially stronger signal from an imputed SNP (rs7903146) in the region than for any of the typed SNPs. This predicted pattern is confirmed by direct genotyping of the SNPs in question[10]. The use of a bayesian measure of association leads to a similar picture (**Supplementary Fig. 3** online).

In addition, we observed strong correlation between the extent and decay of the signal of association and the underlying recombination rate. Furthermore, estimates of the certainty of the imputation at each SNP indicate that the underlying model shows high confidence for our imputations. These observations are not restricted to this region. For example, in the 15 regions in the WTCCC study with strong signals of association at genotyped SNPs for the trend test, there were nine in which the *P* value for the best imputed SNP reduced the *P* value of the best genotyped SNP passing quality control by a factor of at least 1.5, and there were four where the change was by more than an order of magnitude[6].

Validation and missing data imputation at genotyped SNPs

Another application of our imputation engine is in validation of called genotypes and imputation of missing data at SNPs that are actually genotyped in a study. Genotypes can still be imputed at such SNPs (excluding the genotypes at the SNP from the information used for imputation) to provide independent estimates of genotypes at any typed SNP. To illustrate this, we show the normalized intensity data from which genotypes are called for a SNP genotyped in the 1958 birth cohort of the WTCCC study on the Affymetrix chip (**Fig. 5**). The SNP was also typed on a custom Illumina chip. There is a significant amount of overlap between clusters; this is one of the common reasons for genotype-calling problems[6] and can lead to elevated false positive rates if cohort effects are not taken into account in the calling[11]. CHIAMO, the genotype calling algorithm used[6] here, calls 7% of the genotypes as missing, as it cannot be confident of accurate calls in the region of overlap.

When we used our approach to impute the genotypes at this SNP, we found that the average maximum posterior genotype call probability was 0.998, suggesting that the method is very confident of its imputed calls. These imputed calls had a 2.3% discordance with the cluster-based calls among the 1,444 individuals typed on both the Affymetrix and Illumina platforms. The calls made by Illumina at this SNP agreed perfectly with the imputed calls, leading us to conclude that the imputed calls substantially improve data quality at this SNP.

Although we chose this SNP as a particularly good example of the benefits of imputation, we have found that our method offers systematic improvements across a range of SNPs with less marked calling difficulties and, for example, that across the unfiltered set of SNPs in the WTCCC study, using imputed rather than actual genotype data can noticeably reduce false positive rates (data not shown). The optimal way to combine called genotypes with imputed data is not clear and is likely to vary from study to study and to depend on the downstream analysis methods used. For this reason, it is not straightforward to quantify the gain in power from this particular use of imputation, but the issue seems worthy of further attention and empirical study.

DISCUSSION

All multipoint methods for testing association in genome-wide studies can be thought of as predicting missing data at untyped variants, and it is becoming clear that a missing data approach has great utility for many problems in genetics[7,12–14]. In this paper, we have described an imputation engine for genotypes with a primary focus on association studies, but we also emphasize the broader implications of this work in pointing toward a unifying framework for genetic studies. For example, we note that our approach for testing for association is directly comparable to that used in parametric linkage analysis, in which the genotypes of an untyped variant are imputed and averaged over to test for correlation between genetic type and disease status. Both methodologies use a genetic map across the genome to weight the contribution made by the markers surrounding each putative causal locus, and both methods use a likelihood of the same form that involves summation over untyped variation. A significant difference is that in parametric linkage, precise familial relationships are used together with a model of haplotype inheritance to impute untyped variation, whereas in case-control studies, known familial structure is replaced by an unknown population genealogy. Imputation of variants based on an unknown genealogy is a more challenging problem that is facilitated by the use of a population genetics model.

The imputation methods at the core of our approach are analogous to the Elston-Stewart[15] and Lander-Green[16] algorithms used in linkage. A similar relationship exists between our methods and those used in model organisms[17] and suggests that there exists a general statistical framework that may unify ideas across disciplines and stimulate the development of methods for detecting risk factors in a general class of genetic studies. For example, most linkage studies use nonparametric linkage methods. These methods search for regions of elevated allele sharing between affected individuals and are known to be robust to allelic heterogeneity. Along these lines, we are currently developing analogs of allele sharing approaches for genome-wide association mapping.

Genome-wide association studies will be used extensively in the coming years to uncover disease genes. It is becoming clear[6] that most such disease variants will have small effects (odds ratios of 1.2 or smaller). Even today's large studies are underpowered to detect most of these effects, and combination of data across studies will be essential. Thus, a major use of imputation is likely to be in combining data from studies that use different genotyping chips to facilitate these meta-analytic approaches. Other extensions of our approach include (i) the imputation of other types of genetic variants that may show substantial association with phenotypes, such as CNVs[18], microsatellites and HLA loci, (ii) imputation of genotypes for different study designs such as trio designs for association mapping and mapping by admixture LD (MALD) studies and (iii) development of statistical information measures for untyped variants.

We have seen that assessment of imputation methods requires care. For example, imputation accuracy will depend on SNP density as well as the similarity of LD patterns between the data used and the HapMap populations, making it impossible simply to compare headline measures of accuracy across studies.

Our approach makes various modeling assumptions that will not be true in practice. There is growing general evidence that the population genetics model underlying our approach[19], captures many features of human variation data[20]. A particular concern in our context is population structure, for two different reasons. First, differences in LD patterns between the study sample and the HapMap sample used are likely to reduce the accuracy of imputation, and second, whether imputed or genotype data are used, population structure within a study sample can result in false-positive associations. In the analysis presented here, we used the CEU HapMap haplotypes to impute genotypes in a UK sample and saw that accuracy was high, but there is no reason why the other HapMap panels may not be used as well for other studies. Previous work[21] suggests that by conditioning on these panels and including a model of ancestry[22] into our approach, the accuracy of imputation will extend to other less homogeneous studies, and data on the extent to which HapMap data captures LD in other populations are also encouraging here[23]. Obviously, there will be limits, and we would advise caution when applying this approach to data from severely isolated populations. (In a similar way, our model, which assumes a uniform mutation rate, no indels and a recombination rate estimated from HapMap, will not be a cause for concern if the study samples are not too dissimilar from the HapMap populations.) If there is population structure within a study sample, imputation may still work well (as informally, what matters is that haplotypes like those in the study sample occur in the HapMap sample used), but the structure could nonetheless lead to false-positive associations. Approaches for dealing with population structure for genotyped SNPs (such as conditioning on inferred covariates that code for population stratification[24] or applying Genomic Control[25]) can also be applied to imputed SNPs and should work just as well.

The standard paradigm for association analysis is based on detecting disease variants based on their marginal effects, but this may not be the most powerful strategy if significant allelic heterogeneity or gene-gene interactions have a role in the genetic architecture of complex traits[26]. The use of imputed genotypes is not restricted to marginal analysis, and we envisage the development of more complex models of association that allow for these effects. In our current framework, our method of imputing completely unobserved SNPs has some flexibility to accommodate possible allelic heterogeneity at a locus that may occur owing to an underlying multilocus or haplotype model of disease risk.

METHODS

Imputation of missing genotypes. We use $H = \{H_1, \ldots, H_N\}$ to denote a set of N known haplotypes, where $H_i = \{H_{i1}, \ldots, H_{iL}\}$ is a single haplotype, $H_{ij} \in \{0, 1\}$ and L is the number of SNP loci. For all the analyses in this paper, we have set H to be the 120 CEU haplotypes estimated as part of the HapMap project[2]. We let $G = \{G_1, \ldots, G_K\}$ denote the genotype data on the K individuals in a new study, where $G_i = \{G_{i1}, \ldots, G_{iL}\}$ and $G_{ij} \in \{0, 1, 2, \text{missing}\}$. We partition G into an observed and missing component $G = \{G_O, G_M\}$. To impute the missing genotypes, we require the joint distribution of observed and missing genotype data, and we make the modeling assumption that each individual's genotype vector can be considered independently of the others. That is,

$$\Pr(G_M|G_O, H) \propto \Pr(G_M, G_O|H) = \Pr(G|H) = \prod_{i=1}^{K} \Pr(G_i|H)$$

Our model for each individual's genotype vector, $\Pr(G_i|H)$, is a Hidden Markov Model in which the hidden states are a sequence of pairs of the N known haplotypes in the set H. That is,

$$\Pr(G_i|H) = \sum_{Z_i^{(1)}, Z_i^{(2)}} \Pr(G_i|Z_i^{(1)}, Z_i^{(2)}, H)\,\Pr(Z_i^{(1)}, Z_i^{(2)}|H)$$

where $Z_i^{(1)} = \{Z_{i1}^{(1)}, \ldots, Z_{iL}^{(1)}\}$ and $Z_i^{(2)} = \{Z_{i1}^{(2)}, \ldots, Z_{iL}^{(2)}\}$ are the two sequences of hidden states at the L sites and $Z_{il}^{(j)} \in \{1, \ldots, N\}$. These hidden states can be thought of as the pair of haplotypes in the set H that are being copied to form the genotype vector G_i. The term $\Pr(Z_i^{(1)}, Z_i^{(2)}|H)$ defines our prior probability on how sequences of hidden states change along the sequence, and $\Pr(G_i|Z_i^{(1)}, Z_i^{(2)}, H)$ models how the observed genotypes will be close to but not exactly the same as the haplotypes being copied. This model extends a related haplotype model[19] to genotype data. The model allows for recurrent mutation at each SNP but assumes a uniform mutation rate across the genome. The fine-scale recombination map (in units of cM/Mb) estimated from the phase II HapMap is used as a fixed set of parameters in the model and is scaled by a (user-defined) estimate of the effective population size to obtain the population scaled recombination map (ρ) across the region. The precise forms of these terms are described in **Supplementary Methods** online. Our approach can also deal with the situation in which the set G consists of haplotypes in which case a haplotype model[19] is used directly to impute unobserved alleles in these haplotypes.

Under our model, the imputation at a particular SNP can theoretically combine information from all typed SNPs on the same chromosome, although the influence of these SNPs decreases with increasing genetic distance from the locus of interest. Ideally, we would like to consider only the typed SNPs that could plausibly influence the imputation of a given untyped SNP. This can be accomplished in practice by including all SNPs within some 'large' window in the analysis. For the analysis of the WTCCC project, we analyzed regions in 10-Mb windows, padded in each direction with 500-kb buffers to avoid edge effects in the prediction.

We use this model to calculate the (marginal) probability of each possible genotype (0, 1, 2) for each of the missing or unknown genotypes in the study. We also provide a probability distribution for each called genotype to facilitate correction of genotyping errors. These probabilities can be used to carry out association tests at all typed and untyped SNPs.

Imputation of completely missing SNPs. We have also developed methodology for imputing genetic variation at SNPs that are completely unobserved (so-called 'hidden' SNPs) in both the set of haplotypes H and the set of sampled genotypes G. Our method proceeds by simulating M realizations of each such SNP in the N observed haplotypes in the set H. Variation in the set G at the SNP is then simulated by conditioning upon this sample. An approximation to a population genetics model is used to carry out this simulation and is described in more detail in **Supplementary Methods**. Currently, this approach is implemented only in the case in which the haplotype phase of the set G is known, so in practice this would require a preprocessing step of haplotype estimation. A version of this approach that can handle genotype data is currently in development.

Testing association at a SNP or within a region. Once genotypes have been imputed, we can carry out a test of association at a much larger set of SNPs than we had originally typed. By testing each SNP in turn, we assume that disease variants will be detectable based on their marginal effects. As our imputation engine generates probability distributions of untyped genotypes, we found it useful to develop single-SNP tests of association that take this uncertainty into account. A detailed description of the statistical theory and methods we have developed and their relationship to standard tests of association is given in **Supplementary Methods**. A main part of this development involves bayesian measures of single-SNP association known as Bayes factors. Bayes factors are somewhat analogous to frequentist P values, and their use is beginning to emerge in the literature as a more powerful and interpretable alternative to classical tests of association[27].

There are several advantages of using Bayes factors over frequentist test statistics or P values. Proper interpretation of P values requires knowledge of the power of the tests used[6]. Informally, a small P value may arise by chance under the null or from a true association. Assessing which of these might be the case is difficult without knowledge of the power of the study for likely effect sizes (for example, for an underpowered study, we would expect most significant P values to arise by chance). Calculation of Bayes factors, like power calculations, requires assumptions about effect sizes, but Bayes factors have a natural interpretation in their own right as the factor by which our prior odds of association are changed in light of the data. Bayes factors can be naturally combined across different models of association at a given SNP. For example, we can average the Bayes factor across additive, dominant, recessive and general models to avoid having to specify a single model to use at a locus. A similar idea can be used to combine Bayes factors across SNPs within a region. Following recent evidence about the gains in power from bayesian approaches[27], we focused on test statistics based on Bayes factors and compared methods within the two sets of test statistics used in order to focus the results on the ability of each method to predict the causal variants rather than focusing on differences in the power of different test statistics. We used the non-conjugate and conjugate priors (**Supplementary Methods**) for the analysis of the type 2 diabetes data set and the simulated data sets, respectively, to reflect our belief about the genetic effect sizes that are appropriate for these data sets.

Simulated case-control data sets. We simulate case and control individuals conditional upon a set of known haplotype data and an estimate of the fine-scale recombination rate across a region. This approach allows the specification of a SNP in the set of known haplotypes as the causal SNP together with the disease model parameters in terms of relative risks. Genotypes at the causal SNP are simulated under the disease model, and data at flanking SNPs are simulated conditional upon the known haplotypes using a Hidden Markov Model approximation to a population genetics model[19]. This approach is preferable to a direct resampling approach[5], which will tend to produce a set of new haplotypes that are too similar to the HapMap haplotypes. We used the phased data from parents of CEU trios in the ten ENCODE 500-kb regions generated as part of the HapMap project. These haplotypes are expected to be very accurate as a consequence of the trio design of the CEU HapMap panel[2]. We used the fine-scale recombination rates estimated by the PHASE program across these regions (in preference to other estimates of fine-scale recombination rates, because the model underlying PHASE is close to the one on which our imputation is based). As in similar studies[3], we reduced the considerable computational burden of the simulation studies by boosting effect sizes so that variants are detectable in smaller studies. The data sets simulated in the paper thus consisted of haplotypes for 100 case and 100 control individuals, but we would expect conclusions based on comparisons of methods to extend to the larger sample sizes and smaller effect sizes typical of GWAs for common human diseases[3].

We created a 'pseudo' HapMap panel by thinning the ENCODE data to match the SNP density and MAF distribution of the phase II HapMap data, with the added restriction that this panel contain the SNPs on the Affymetrix 500K mapping chip that lie within a given region. Multi-marker predictors (MMPs) were designed using the Affymetrix SNP set in each region based on LD patterns within the pseudo-HapMap panels. We searched for the best MMPs of sizes 2 and 3 that predict SNPs in the panel with a sample $r^2 \geq 0.8$, with the constraint that all pairs of predictor and predicted SNPs in each MMP rule lie within 200 kb of each other.

Software implementation. Three programs were written to carry out the analysis described in this paper. A program called IMPUTE was written to determine the probability distribution of missing genotypes conditional upon a set of known haplotypes and an estimated fine-scale recombination map. A program called SNPTEST was written that implements all of the frequentist and bayesian tests used in this paper and described in the **Supplementary Methods**. A program called HAPGEN was written to simulate case-control data sets conditional upon a set of observed haplotypes. These programs are available on the website http://www.stats.ox.ac.uk/~marchini/#software.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank M. McCarthy, E. Zeggini, A. Hattersley and the WTCCC for allowing us to use the *TCF7L2* data. We thank M. Stephens, B. Servin, P. De Bakker, P. Fearnhead, J. Barrett, Z. Su, C. Spencer, D. Vukcevic and N. Cardin for discussions. We acknowledge support from The Wellcome Trust, the US National Institutes of Health, the SNP Consortium, the Wolfson Foundation, the Nuffield Trust and the Engineering and Physical Sciences Research Council. B.H. is supported by a National Science Foundation Graduate Research Fellowship and the Overseas Research Students Award Scheme.

COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
2. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
3. Zollner, S. & Pritchard, J.K. Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics* **169**, 1071–1092 (2005).
4. Morris, A.P. Direct analysis of unphased SNP genotype data in population based association studies via Bayesian partition modelling of haplotypes. *Genet. Epidemiol.* **29**, 91–107 (2005).
5. de Bakker, P.I.W. *et al.* Efficiency and power in genetic association studies. *Nat. Genet.* **37**, 1217–1223 (2005).
6. The 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).
7. 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).
8. Eyheramendy, S., Marchini, J., McVean, G., Myers, S. & Donnelly, P.A. Model-based approach to capture genetic variation for future association studies. *Genome Res.* **17**, 88–95 (2007).
9. Grant, S.F.A. *et al.* Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat. Genet.* **38**, 320–323 (2006).
10. Groves, C.J. *et al.* Association analysis of 6,736 U.K. subjects provides replication and confirms tcf7l2 as a type 2 diabetes susceptibility gene with a substantial effect on individual risk. *Diabetes* **55**, 2640–2644 (2006).
11. Clayton, D.G. *et al.* Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat. Genet.* **37**, 1243–1246 (2005).
12. Stephens, M., Smith, N.J. & Donnelly, P. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68**, 978–989 (2001).
13. Kraft, P., Cox, D.G., Paynter, R.A., Hunter, D. & De Vivo, I. Accounting for haplotype uncertainty in matched association studies: a comparison of simple and flexible techniques. *Genet. Epidemiol.* **28**, 261–272 (2005).
14. Cordell, H.J. Estimation and testing of genotype and haplotype effects in case-control studies: comparison of weighted regression and multiple imputation procedures. *Genet. Epidemiol.* **30**, 259–275 (2006).
15. Elston, R.C. & Stewart, J. A general model for the genetic analysis of pedigree data. *Hum. Hered.* **21**, 523–542 (1971).
16. Lander, E.S. & Green, P. Construction of multi-locus genetic linkage maps in humans. *Proc. Natl. Acad. Sci. USA* **84**, 2363–2367 (1987).
17. Sen, S. & Churchill, G.A. A statistical framework for quantitative trait mapping. *Genetics* **159**, 371–387 (2001).

18. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
19. Li, N. & Stephens, M. Modelling linkage disequilibrium, and identifying recombination hotspots using SNP data. *Genetics* **165**, 2213–2233 (2003).
20. Crawford, D.C. *et al.* Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nat. Genet.* **36**, 700–706 (2004).
21. de Bakker, P.I.W. *et al.* Transferability of tag SNPs in genetic association studies in multiple populations. *Nat. Genet.* **38**, 1298–1303 (2006).
22. Falush, D., Stephens, M. & Pritchard, J.K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587 (2003).
23. Conrad, D.F. *et al.* A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat. Genet.* **38**, 1251–1260 (2006).
24. Pritchard, J.K., Stephens, M., Rosenberg, N.A. & Donnelly, P. Association mapping in structured populations. *Am. J. Hum. Genet.* **67**, 170–181 (2000).
25. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
26. Marchini, J., Donnelly, P. & Cardon, L.R. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.* **37**, 413–417 (2005).
27. Balding, D.J. A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.* **7**, 781–791 (2006).