Comparing Algorithms for Genotype Imputation

To the Editor: When the data from a genome-wide association study is analyzed, a key question is how to extract the strongest "signal" of association. Over the last few years, a class of genotype imputation methods[1–3] has become increasingly popular for boosting the signal above that obtained by standard single-SNP analyses.

Here, we define genotype imputation as the prediction of genotypes at SNPs that have not been assayed in an association study. This is typically accomplished by modeling allelic correlations among SNPs (many of which will not have been typed in a given study) in a panel of known haplotypes, such as the HapMap,[4] and extrapolating these correlations to a sample of interest through information from SNPs that have been typed in that sample.

Sophisticated imputation methods have been shown to be more powerful than tagging approaches that test only single SNPs or small haplotypes of SNPs on a genotyping chip,[3] to provide clearer pictures of associated regions that aid design of replication and fine-mapping studies,[5] and to facilitate meta-analysis projects[6] by allowing data sets collected with different genotyping chips to be combined for increased power.

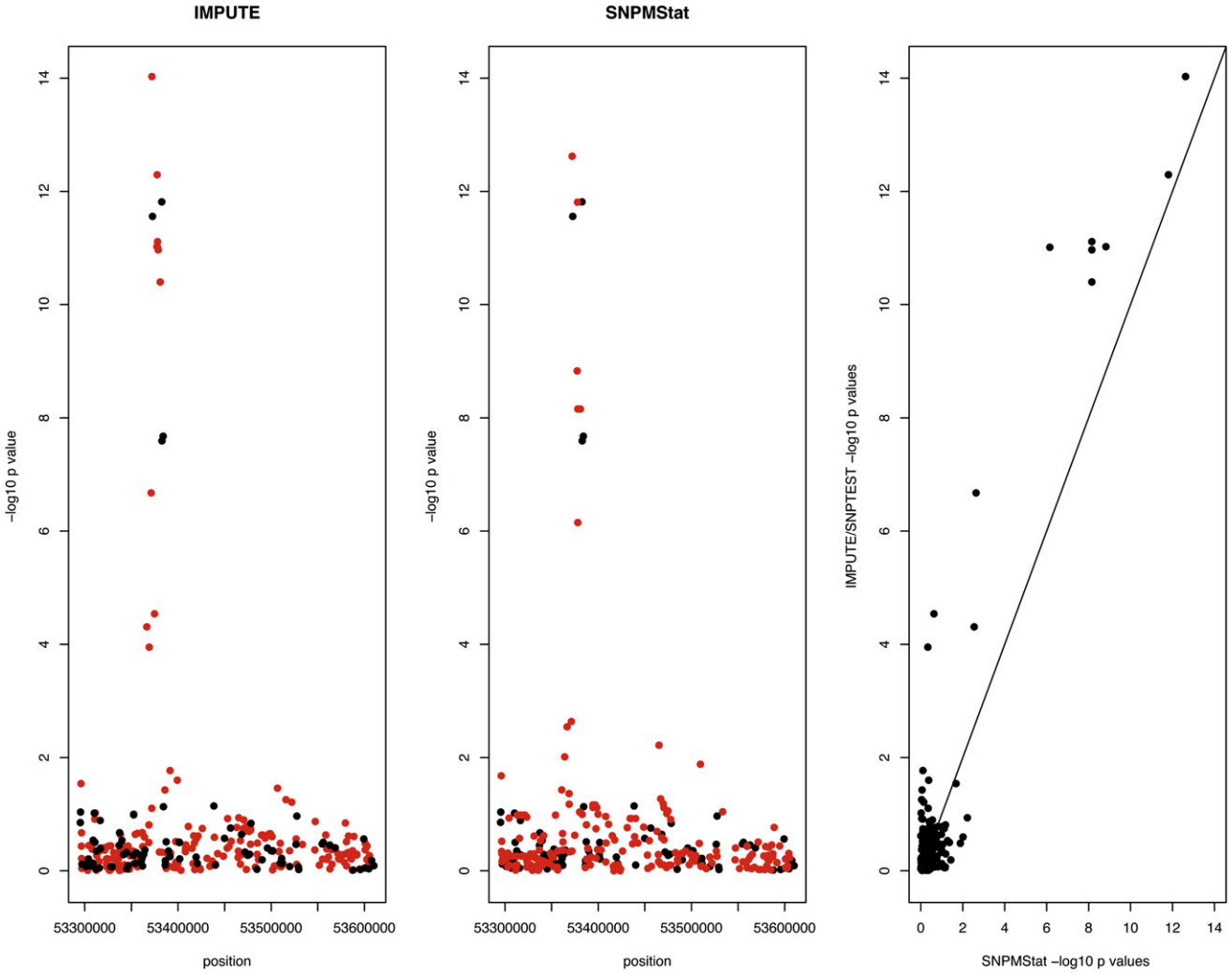

Figure 1. Results of Running IMPUTE/SNPTEST and SNPMStat on the Simulated Data Set Provided with the SNPMStat Software
This data set was designed to mimic a real Rheumatoid Arthritis study[7] and consists of data at 100 SNPs in 1000 cases and 1000 controls. Black dots represent tests at genotyped SNPs, and red dots represent tests at imputed SNPs. Test statistics are plotted on the −log10 p value scale. Both programs were run under the assumption of an additive model of association. The right-hand plot shows the test statistics of both methods plotted against each other at all genotyped and imputed SNPs.

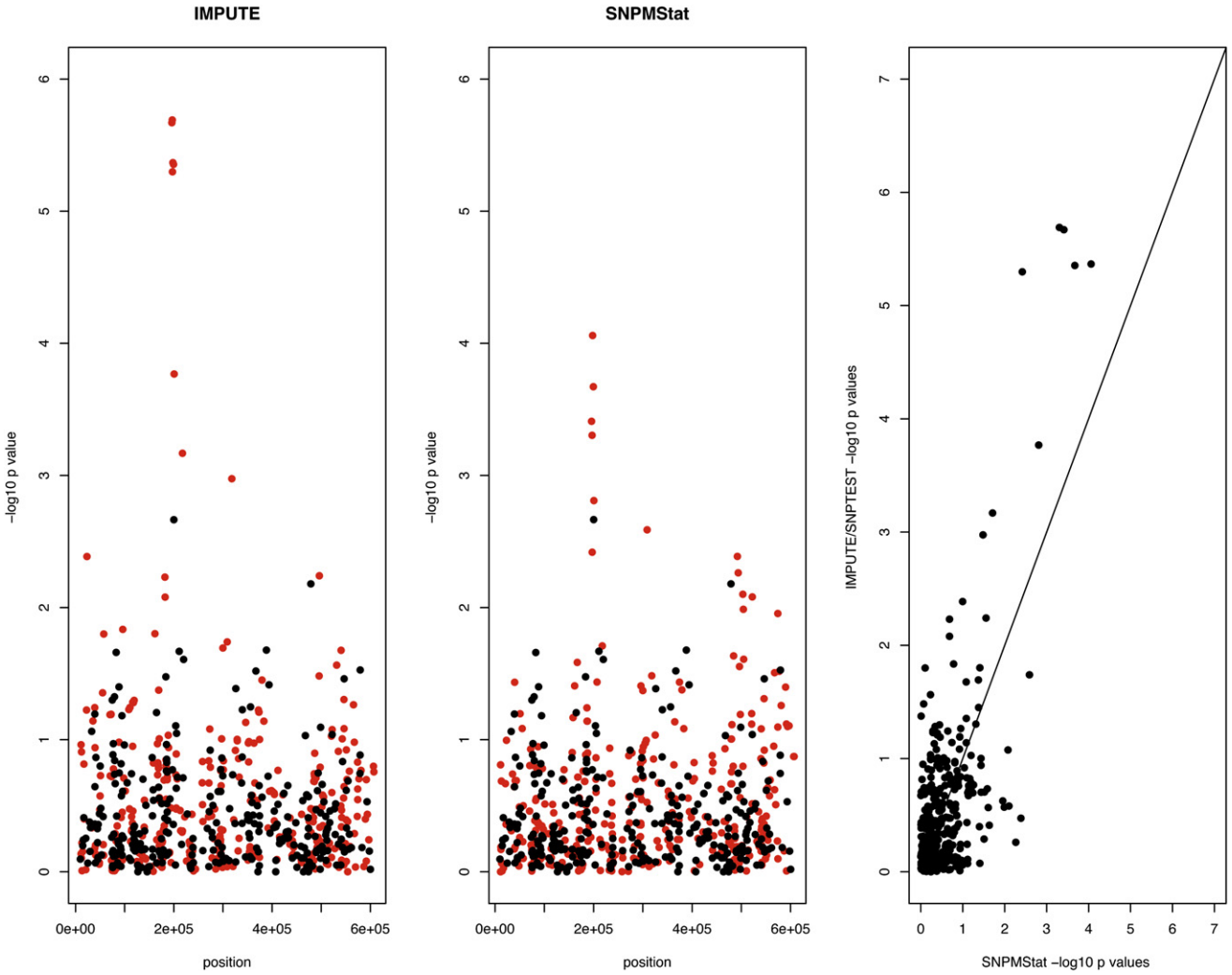

Figure 2. Results of Running IMPUTE/SNPTEST and SNPMStat on a Data Set Simulated with the HAPGEN Program
The data set consists of 500 cases and 500 controls at 300 SNPs on the Affymetrix 500k chip from a region on chromosome 20. Black dots represent tests at genotyped SNPs, and red dots represent tests at imputed SNPs. Test statistics are plotted on the −log10 p value scale. Both programs were run under the assumption of an additive model of association. The right-hand plot shows the test statistics of both methods plotted against each other at all genotyped and imputed SNPs.

In the February 2008 issue of the *Journal*, Lin et al.[7] proposed a new method of genotype imputation, called SNPMStat. The main strength of the method is that it simultaneously fits a model of association and imputes missing genotypes. Competing methods that impute genotypes without acknowledging phenotypic status formally assume that all of the individuals in a study are no more related to one another than would be a set of people sampled at random from a "population." Near a disease locus, however, cases are more closely related to each other than this assumption would imply; consequently, explicitly modeling each individual's disease status could lead to more accurate imputation and measures of association strength.

A limiting feature of the proposed method is that it uses only a small number of SNPs (four at most) to impute each untyped SNP; this constraint arises from the computational challenges of fitting a joint model of genotype and phenotype, and it diminishes the capacity to model complex correlations between SNPs. An additional limitation is the use of a simple, parametric, multinomial model of haplotype frequencies: the method's likelihood-maximization process involves an implicit phasing of the SNP genotype data, and previous comparisons[8] have shown that this model performs much worse at phasing than do the models underlying competing imputation approaches.[2,3]

By contrast, our own method[3] (called IMPUTE) and other methods[2] impute genotypes without reference to phenotype but use more of the flanking SNPs and more sophisticated population-genetics models to predict unobserved genotypes. These genotypes can then be included in tests of association (e.g., with our program SNPTEST) that account for the uncertainty in the predictions and at the same time condition upon observed covariates

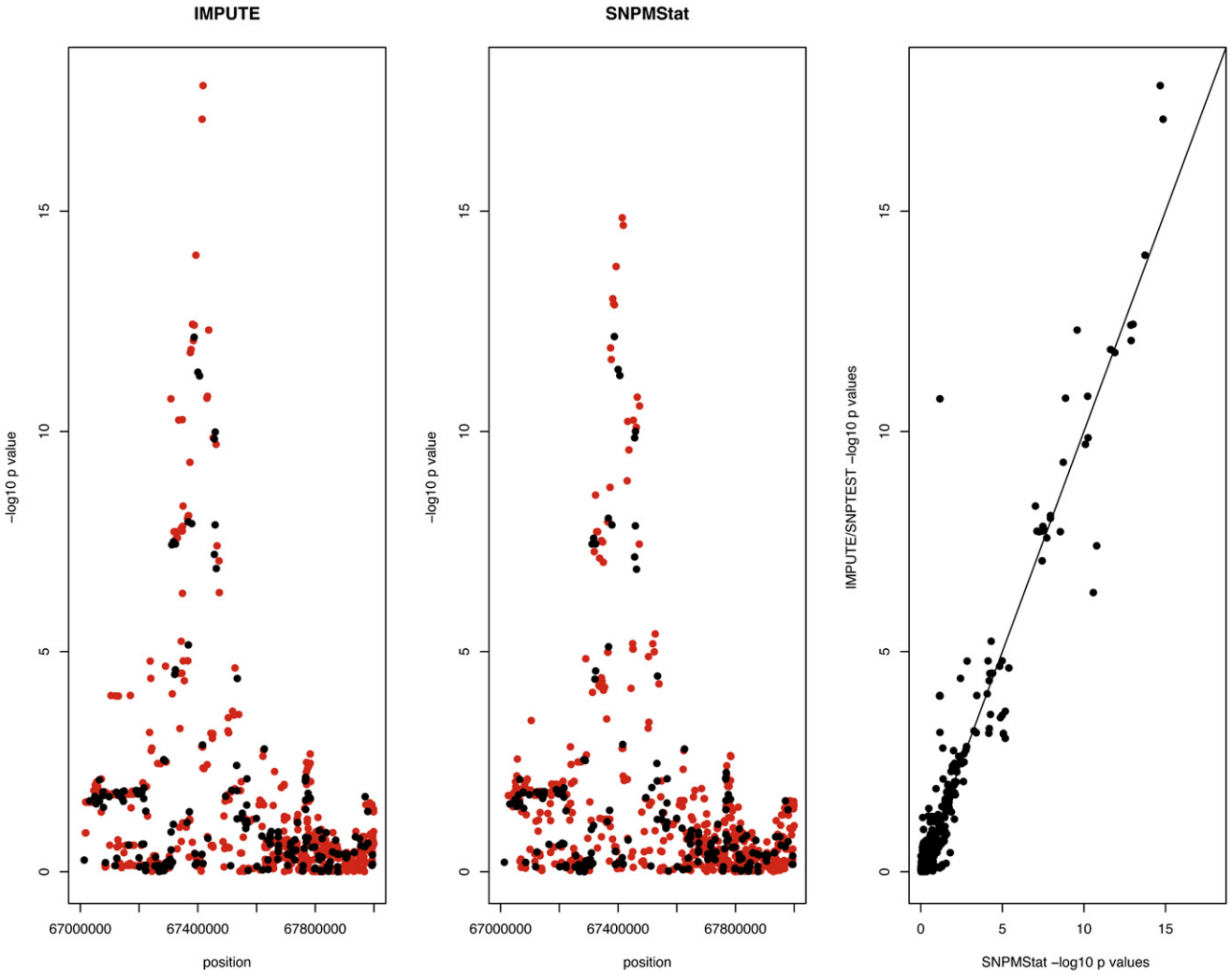

Figure 3. Results of Running IMPUTE/SNPTEST and SNPMStat on a Real Data Set in a Region Shown to be Associated with Crohn's Disease in the WTCCC Study[5]

This data set consists of 2938 controls and 1758 cases at 181 of the SNPs on the Affymetrix 500k chip in a 1 Mb region of chromosome 1. Black dots represent tests at genotyped SNPs, and red dots represent tests at imputed SNPs. Test statistics are plotted on the −log10 p value scale. Both programs were run under the assumption of an additive model of association. The right-hand plot shows the test statistics of both methods plotted against each other at all genotyped and imputed SNPs.

(completing what Lin et al. call a "two-stage" procedure, which contrasts with their "joint" approach).

Many researchers working in this field would like to know which method is most powerful. With regard to the approaches discussed above, this boils down to whether a joint model fitted to a small amount of data is more accurate than a two-stage strategy that makes fuller use of the genotype data when carrying out imputation.

Intuitively, we can predict that a joint model will prove most useful when there are large differences between cases and controls (i.e., risk alleles with strong effects) and that a two-stage model will fare better if cases and controls look more similar (this claim is based on the reasonable assumption that the models underlying the two-stage approaches would impute genotypes more accurately in a controls-only data set). The conditions under which the cases and controls are "different enough" for the joint model to gain an advantage remain unclear, so researchers are currently looking to simulation studies (and some real ones) for guidance.

Lin et al. present simulations that they claim extensively examine the problem of untyped SNPs, and they use these to suggest that their method is more powerful than existing two-stage approaches. Unfortunately, their simulations do not apply the competing approaches in realistic settings, so it is difficult to justify these claims. Specifically, all of their simulations involve data sets of five consecutive SNPs from chromosome 18 of the CEU HapMap, where four of the SNPs are used to impute the fifth (whose genotypes in a simulated case-control set are hidden from the imputation methods). None of the referenced papers describe or recommend their use on such small data sets,

nor are such data sets typical of modern association studies, so the comparisons might well be biased in favor of Lin et al.'s joint model.

To investigate this issue, we applied the SNPMStat method and our own two-stage approach, using the programs IMPUTE and SNPTEST, to two simulated data sets and one real data set from the Wellcome Trust Case Control Consortium (WTCCC) (Figures 1–3). The first simulated data set was originally supplied with the SNPMStat software. This data set was designed to mimic a real Rheumatoid Arthritis study[7] and consists of data at 100 SNPs in 1000 cases and 1000 controls. In the process of replying to this letter, the authors of SNPMStat found bugs in SNPMStat, changed the file formats, and added new SNPs to the data set. We removed these new SNPs from the newly formatted files and reran our analysis with the new version of the software. For the second data set, we used the HAPGEN program to simulate 500 cases and 500 controls at 300 SNPs on the Affymetrix 500k chip from a region on chromosome 20. The third data set is a real one from a region shown to be associated with Crohn's Disease (CD [MIM 266600]) in the WTCCC study.[5] This data set consists of 2938 controls and 1758 cases at 181 of the SNPs on the Affymetrix 500k chip in a 1 Mb region of chromosome 1.

Figure 1 depicts a region with a clear signal of association. Both methods are able to identify a SNP that is more strongly associated with disease status than is any genotyped SNP (illustrating the capacity of imputation methods to inform subsequent fine-mapping studies), but IMPUTE/SNPTEST picks out this SNP much more clearly. Figure 2 shows a region with a much weaker signal of association: the smallest p value at the genotyped SNPs does not even reach 10^{-3}. Here, we see again that both imputation methods boost the signal, but with a disparity between the smallest p values (nearly 10^{-6} for IMPUTE/SNPTEST, as compared to 10^{-4} for SNPMStat) that could easily mean the difference between carrying this region forward for further scrutiny and losing it among the genomic noise. The comparison in Figure 3 uses real data from the WTCCC study. As before, although it is clear that either imputation method can enhance our understanding of this associated region, our two-stage method amplifies the signal to a much greater extent—IMPUTE/SNPTEST achieves a p value over 100 times smaller than the smallest p value generated by SNPMStat, and our claim about the strength of this association is supported by external data.[9]

These three figures do not amount to a systematic power comparison, but they are highly suggestive, as shown further by the consistently (and appropriately) stronger signals extracted by IMPUTE/SNPTEST at highly associated, untyped SNPs throughout these regions (rightmost panels in Figures 1–3). Thus, on data sets with realistic SNP landscapes and disease effect sizes, it appears that much more is gained by the use of advanced population-genetics models than is lost by failure to model the differences between cases and controls.

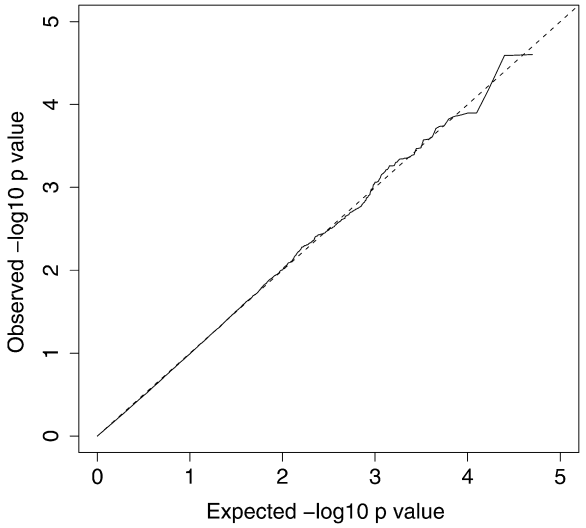

Figure 4. An Evaluation of Type I Error of the IMPUTE/SNPTEST Approach

HAPGEN was used to simulate a case-control data set consisting of 1000 cases and 1000 controls at all of the SNPs on the Affymetrix 500K chip on chromosome 1 under the null hypothesis of no association. The plot shows the observed −log10 p-values at a random subset of 50,000 of the imputed SNPs versus their expected values under the null.

As a technical point, we note that power can be compared between two methods only if they both control their type I error. To assess this for our own method, we used HAPGEN to simulate a case-control data set consisting of 1000 cases and 1000 controls at all of the SNPs on the Affymetrix 500K chip on chromosome 1 under the null hypothesis of no association. We then applied the IMPUTE/SNPTEST approach to this data set. Figure 4 shows the PP plot based on 50,000 of these SNPs, which indicates that type I error is controlled very well (as might be expected from a method that imputes genotypes "under the null").

Finally, it is important in this context to distinguish between two forms of type I error. The first, which we address here and Lin et al.[7] mention briefly, pertains to the detection of novel regions of association in the genome via imputation. The second, which is the focus of Figure 1 in Lin et al., reflects the ability of a method to separate causal from noncausal associations in the neighborhood of a true risk variant. Both are important issues, but our methods development has focused primarily on the first of these questions; others have found that analysis techniques grounded in imputation models similar to ours show great promise for fine-mapping applications.[2]

In summary, the large and consistent differences shown in these early results on realistic data sets suggest that methods that use as much of the available genotype data as possible might be more powerful than those that fit a phenotype model using only a subset of the data and

that the simulation studies presented by Lin et al. should, therefore, be interpreted with caution.

Jonathan Marchini[1,*] and Bryan Howie[1]
[1]Department of Statistics, University of Oxford, Oxford OX1 3TG, UK
*Correspondence: marchini@stats.ox.ac.uk

Acknowledgments

This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113.

Web Resources

The URLs for data presented herein are as follows:

HAPGEN, IMPUTE and SNPTEST programs, http://www.stats.ox.ac.uk/~marchini/software/gwas/gwas.html
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/OMIM/

References

1. Nicolae, D.L. (2006). Testing untyped alleles (TUNA)-applications to genome-wide association studies. Genet. Epidemiol. *30*, 718–727.
2. Servin, B., and Stephens, M. (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet *3*, e114.
3. Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet. *39*, 906–913.
4. Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.
5. WTCCC T. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661–678.
6. Zeggini, E., Scott, L.J., Saxena, R., Voight, B.F., Marchini, J.L., Hu, T., de Bakker, P.I., Abecasis, G.R., Almgren, P., Andersen, G., et al. (2008). Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat. Genet. *40*, 638–645.
7. Lin, D.Y., Hu, Y., and Huang, B.E. (2008). Simple and efficient analysis of disease association with missing genotype data. Am. J. Hum. Genet. *82*, 444–452.
8. Marchini, J., Cutler, D., Patterson, N., Stephens, M., Eskin, E., Halperin, E., Lin, S., Qin, Z.S., Munro, H.M., Abecasis, G.R., and Donnelly, P. (2006). A comparison of phasing algorithms for trios and unrelated individuals. Am. J. Hum. Genet. *78*, 437–450.
9. Duerr, R.H., Taylor, K.D., Brant, S.R., Rioux, J.D., Silverberg, M.S., Daly, M.J., Steinhart, A.H., Abraham, C., Regueiro, M., Griffiths, A., et al. (2006). A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science *314*, 1461–1463.

DOI 10.1016/j.ajhg.2008.09.007.

Reply to Marchini and Howie

To the Editor: As noted by Marchini and Howie (MH), an advantage of our maximum likelihood (ML) approach is that the genotypes of untyped SNPs are inferred from proper posterior distributions. The two-stage approach, which ignores the phenotype information in the imputation of genotypes, can yield biased estimates of genetic effects near disease loci and consequently reduce power, especially when the genetic effects are strong. It is difficult to fully account for the uncertainties of the imputed genotypes in the two-stage approach, especially if environmental covariates are involved.

From a frequentist point of view, it is impossible to do better than the ML approach, which has the highest statistical efficiency among all valid methods (that use the same data and make the same assumptions). The two-stage approach might produce more accurate results than the ML approach in certain situations because it allows the use of sophisticated population-genetics models in the first stage. The ML approach is more robust, in that it estimates the joint distribution between the untyped SNP and the flanking markers nonparametrically. Although we use a small number of flanking markers, we search over all subsets of flanking markers around the untyped SNP and select the subset that provides the best prediction of genotypes at the untyped SNP. By searching over all possible subsets of four SNPs among the 20 SNPs closest to each untyped HapMap SNP, we can typically obtain Rs^2 of 1 for more than 50% of untyped SNPs and Rs^2 of > 0.9 for 80% of untyped SNPs. It is unclear how much improvement sophisticated population-genetics models can bring.

MH are absolutely right that our simulation studies did not evaluate the role of sophisticated population-genetics models. Indeed, we stated this fact in the Discussion of our article. Our simulation studies were designed to compare the ML and two-stage approaches when the same set of flanking markers is used. The results showed the efficiency gain of the ML approach due to the use of the phenotype information when inferring unobserved genotypes and the use of retrospective likelihood for reflecting case-control sampling. When applying the ML method to real data, we always search over a large region around each untyped SNP to find a set of flanking markers that provides the best prediction of genotypes for the untyped SNP.

We are intrigued by the comparisons between SNPMStat and IMPUTE/SNPTEST reported by MH. However, it is difficult to draw any firm conclusion from a small number of selective data sets. The results for the Rheumatoid Arthritis