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. ©2008 by The American Society of Human Genetics. All rights reserved.

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

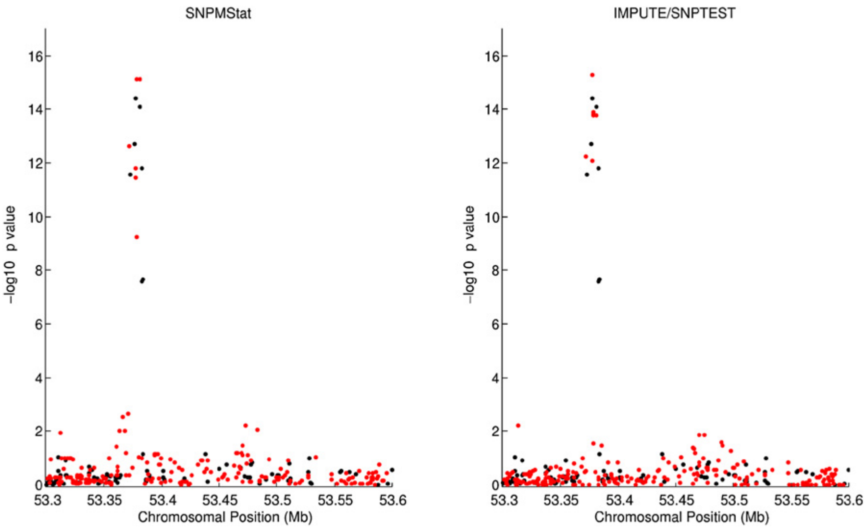

Figure 1. Results of Running SNPMStat and IMPUTE/SNPTEST on the Simulated Rheumatoid Arthritis Study Data when the Reference Panel Contains *all* of the HapMap SNPs

The −log10 p values under the additive model for the genotyped and untyped SNPs are shown in black and red dots, respectively.

study shown in Figure 1 of MH were based on a subset of the HapMap SNPs that was originally posted on our website for the users to test our software. As mentioned by MH, we recently updated the reference panel to include all of the HapMap SNPs. With this more realistic reference panel, the results of SNPMStat and IMPUTE/SNPTEST are very similar; see our Figure 1. For this example, SNPMStat was ten times faster than IMPUTE/SNPTEST. It is unclear how representative the two examples shown in Figures 2 and 3 of MH are or how robust the results of IMPUTE/SNPTEST are to the choices of parameters used in the population-genetics model. It does not seem possible for an imputation method with correct type I error to always produce p values at untyped SNPs that are much smaller than those at typed SNPs. The comparisons on the p value scale might exaggerate the differences between competing methods, because a small difference in the test statistic at the extreme tail(s) of the distribution translates into a substantial difference in the p value. As noted by MH, it would be preferable to compare the ML and two-stage approaches through extensive simulation studies with realistic SNP landscapes and disease effect sizes.

D.Y. Lin[1,*] and Y. Hu[1]
[1]Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, USA
*Correspondence: lin@bios.unc.edu

DOI 10.1016/j.ajhg.2008.09.008.