Pre-phasing: Fast and accurate genotype imputation in genome-wide association studies

Christian Fuchsberger[1], Bryan Howie[2], Matthew Stephens[2,3],

Gonçalo R. Abecasis[1], and Jonathan Marchini[4,5]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, US; [2]Department of Human Genetics, University of Chicago, Chicago, US; [3]Department of Statistics, University of Chicago, Chicago, US; [4]Department of Statistics, University of Oxford, Oxford, UK; [5]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK

Sequencing efforts, such as the 1,000 Genomes Project, are producing large collections of haplotypes that can be used for genotype imputation in genome-wide association studies. Imputing from these reference panels can help identify new risk alleles, but the use of large panels with existing methods imposes a high computational burden. We describe and evaluate an efficient approach for addressing this problem: we first estimate haplotypes in a GWAS dataset using standard phasing algorithms (pre-phasing), then we use the estimated haplotypes directly for imputation. While it would take an imputation method like IMPUTE v1 or MaCH >1 Million CPU hours to impute the GAIN Psoriasis dataset, pre-phasing requires <1850 CPU hours. Despite its much lower computational cost, the pre-phasing approach provides comparable accuracy to state-of-the-art imputation methods, even for rare variants and populations with reduced linkage disequilibrium. Furthermore, pre-phasing imputation accuracy can be improved by multiple imputations and by increasing the quality of the GWAS haplotypes. Our methods are implemented in C++, run on Windows, Mac and Linux, and are available at

http://genome.sph.umich.edu/wiki/minimac (minimac) and

http://mathgen.stats.ox.ac.uk/impute/impute_v2.html (IMPUTE 2.0).