In Silico Genotyping for Genome-Wide Association Studies

Y. Li, C.J. Willer, J. Ding, P. Scheet, G.R. Abecasis

Center for Statistical Genetics, Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI.

Large scale genome-wide association (GWA) studies hold the promise of detecting the small genetic effects that underlie genetic susceptibility to complex diseases but pose a range of analytical and computational challenges. We propose a method that can rapidly impute several million SNPs genotyped by the HapMap consortium using genome-wide SNP genotyping data such as that provided by commercial genotyping platforms by Illumina, Affymetrix or Perlegen. The method uses a hidden Markov model to assemble mosaics of haplotypes observed in the appropriate HapMap reference population that match each of the sampled individuals. We illustrate our approach with real data sets studying age-related macular degeneration and type 2 diabetes. We demonstrate the capability of our method (1) to generate highly accurate genotypes along with correspondent measures of imputation uncertainty, (2) to improve coverage and gain power in association mapping, and (3) to facilitate meta-analysis across studies that use different commercial panels for genotyping. Our method is computationally efficient. For example, in our study of type 2 diabetes, we imputed several billion genotypes using genotypes from the Illumina 317K panel as input. The computation took less than two days for the largest chromosome and multiple chromosomes were conveniently run in parallel. The allelic concordance between imputed and true genotypes is ~98.5%, which can be further improved to over 99% by excluding the 5% of SNPs that are estimated to have lower quality imputed genotypes. Our method is implemented in C++, runs on Windows, Mac and Linux and is available at www.sph.umich.edu/csg/abecasis/mach/