Bayesian imputation-based methods for genome-wide association studies

Jonathan Marchini[1], Zhan Su[1], Bryan Howie[1], Peter Donnelly[1,2]

[1] Department of Statistics, University of Oxford
[2] The Wellcome Trust Centre for Human Genetics, University of Oxford

Genome-wide association studies are a relatively new way for scientists to identify which genes are involved in human disease. This study design searches the genome for small variations (called SNPs) that differ in frequency between cases and controls. Each study will collect data at up to 1 million SNPs in thousands of cases and controls. The simplest method of analysis involves testing each SNP one at a time but it is widely recognized that this will not be the most powerful strategy. We have developed two related methods that attempt to detect a signal of association using the data at all the SNPs in a given region. The first method attempts to impute or predict the data at SNPs that are known to exist but have not been sampled in the study. We use a hidden markov model to combine our study data with a much denser set of SNP data (such as the HapMap) and then calculate a Bayes Factor to test for association at the imputed SNPs. Our second method attempts to infer or predict data at SNPs that do not exist in dense SNP panels such as HapMap. At each position in the genome we approximate the genealogy of the sample of individuals and then calculate a Bayes Factor by averaging over all possible disease SNPs that could have occured within that genealogy. This allows us to pick up signals that involve combinations of mutations at different SNPs. We also allow for more than one disease SNP mutation so that we can carry out inference on the number of likely disease mutations in a region. We illustrate both methods using real data from the 7 genome-wide association studies carried out by the Wellcome Trust Case Control Consortium.