Simultaneous Analysis of all SNPs in a GenomeWide association study

John Whittaker[1], Claudio Verzilli[1], Clive Hoggart[2], Maria De Iorio[2] and David Balding[2]

[1] London School of Hygiene & Tropical Medicine, UK
[2] Imperial College London, UK

We describe two methods to allow simultaneous analysis of the entire set of SNPs from a genomewide study. The first is based on the use of discrete graphical models as a data-mining tool, searching for single- or multilocus patterns of association around a causative site. The approach is fully Bayesian, allowing us to incorporate prior knowledge on the spatial dependencies around each marker due to linkage disequilibrium, which reduces considerably the number of possible graphical structures. A Markov chainMonte Carlo scheme is developed that yields samples from the posterior distribution of graphs conditional on the data from which probabilistic statements about the strength of any genotype-phenotype association can be made. The second employs a Bayesian inspired penalised maximum likelihood approach in which every SNP can be considered for additive, dominant and recessive contributions to disease risk. Posterior mode estimates are obtained for regression coefficients that are each assigned a prior with a sharp mode at zero. A nonzero coefficient estimate is interpreted as corresponding to a significant SNP. We also derive an explicit approximation for type I error that avoids the need to employ permutation procedures. Our method is fast and can handle very large numbers of SNPs, making it suitable for analyses of datasets arising from imputation or resequencing.The approaches are illustrated using simulated and real data.

Keywords: Genome wide association, Bayesian analysis