



## **A Hierarchical Statistical Model based on Latent Genotypes for Genome-Wide Association Studies**

Luis Carvalho\*  
BU, U.S.A. – [lecarval@math.bu.edu](mailto:lecarval@math.bu.edu)

Ian Johnston  
BU, U.S.A. – [ianj@bu.edu](mailto:ianj@bu.edu)

Genome-wide association studies (GWAS) attempt to determine which genomic markers are predictors of genetic traits, most commonly human diseases. In practice, despite the extreme imbalance of having millions of markers recorded for only a few thousand individuals, it is of great interest to glean as much information as possible from this type of data. To this end, we propose a novel statistical model that exploits a hierarchical structure between markers and genes to leverage information between levels and alleviate the "large-p smalln" regimen while still attaining a reasonably complex and realistic model. In addition, we propose a new latent block genotype procedure that aims to correct genotypical correlations. Fitting the model is challenging due to the high number of variables to select, so we discuss efficient computational approaches that we explored to estimate the parameters. Finally, we illustrate the proposed model and estimation procedures on simulated data and on a real-world data set from the Wellcome Trust Consortium.