AN EFFICIENT GENOME-WIDE MULTILOCUS EPISTASIS SEARCH

MIKKO J. SILLANPÄÄ

High-throughput laboratory techniques are producing vast amount of genomic marker data – discrete predictors to association studies. Linear regression model is often considered to link study phenotypes and these marker measurements to each other. Number of predictors in multi-marker regression models can easily be much larger than number of observations. Therefore, one needs application of variable selection to find small subset of important predictors out of large number of candidates. Such models can occasionally include also all pairwise locus-by-locus (epistasis) interactions which increases dimensionality of the model very rapidly. We consider variable selection problem of linear model containing large amount of predictors and all of their pairwise interactions in the model jointly. Our suggested approach (Kärkkäinen et al. 2015) use sure-independence-screening to first drop dimension of the problem by considering marginal importance of each interaction term within the huge loop. Subsequent estimation step then consider Bayesian variable selection approach (Extended Bayesian LASSO – Mutshinda and Sillanpää 2010). We also show that it is important to separate search of main and interaction effects in the algorithm to control number of false positives. Examples illustrates superior performance of our method over PLINK in terms of computation time and empirical power. Our successful examples consider even problem of originally of order of 280,000,000 interactions within a reasonable time frame.

REFERENCES Mutshinda CM, Sillanpää MJ (2010) Extended Bayesian LASSO for multiple quantitative trait loci mapping and unobserved phenotype prediction. Genetics 186: 1067-1075. Kärkkäinen HP, Li Z, Sillanpää MJ (2015) An efficient genome-wide multilocus epistasis search. Genetics 201: 865-870.