DOI 10.18699/SBB-2018-31

Heterogeneity of linkage disequilibrium across human population and its influence on statistical properties of the conditional and joint models for genetic association testing

A.V. Severinov[1, 3]*, S.A. Slavskii[1, 3, 5]**, T.I. Shashkova[1, 2, 3], D.D. Gorev[1, 2, 3], D.G. Alexeev[1, 2],G.A. Bazykin[5, 6], Y.S. Aulchenko[2, 4]

[1] *LLC 'Knomics', Moscow, Russia*
[2] *Novosibirsk State University, Novosibirsk, Russia*
[3] *Moscow Institute of Physics and Technology, Moscow, Russia*
[4] *Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia*
[5] *Skolkovo Institute of Science and Technology, Skolkovo, Russia*
[6] *Institute for Information Transmission Problems* (*Kharkevich Institute*) *of the RAS, Moscow, Russia*
** e-mail: aleksandr.severinov@phystech.edu; ** slavsky@phystech.edu*

Key words: genome-wide association study, linkage disequilibrium, conditional and joint association testing

Motivation and Aim: Conditional and joint (COJO) tests of association between multiple single-nucleotide polymorphisms (SNPs) on the basis of summary statistics reported in single-SNP genome-wide association scans (GWAS) require knowledge of linkage disequilibrium (LD) in the region undergoing the testing. Because of different reasons, in most situations the LD structure from the sample(s) where the original GWAS was performed is not available. Therefore methods based on analysis of GWAS results often have to use LD computed from some other, reference sample. When doing so, it is (implicitly) assumed that the LD structure in the reference sample approximates the LD in original GWAS sample well. It is not quite clear how possible difference in the LD structure between the GWAS and reference populations influence the statistical properties of methods, such as COJO, that utilize summary statistics. In practice, it is also not known how large could be the LD difference between a typical GWAS and a reference populations. In this work, our aim was to estimate the distribution of difference in the LD structure between European populations and to relate the error in estimation of LD due to use of reference population to the statistical properties of multi-SNP conditional and joint (COJO) analysis of genetic associations.

Methods and Algorithms: For estimation of the variance of LD between different populations we developed and applied a model of genetic drift. Using effective population size and number of generations since divergence as input parameters along with allele frequencies we considered a problem of the genetic drift of haplotypes across generations. The theoretical results were compared to the distributions observed in real whole genome data from different European populations. To investigate the effects of the variance in LD between reference and GWAS population, we have developed and implemented an algorithm for simulating genotypes for two SNPs with certain LD coefficient and a quantitative phenotype. After obtaining simulated data we ran COJO analysis on it varying input parameters (particularly LD-coefficient between SNPs). Then, we estimated the rate of false positives and false negatives for both conditional and joint tests as a function of deviation of the LD-coefficient from its true value.

Results and Conclusions: Using *in silico* analysis we estimated possible differences in LD between European populations and confirm our results by analysis of real data. On the other hand, we estimated the effect of use of biased LD estimates onto the statistical properties of the COJO method. Lastly, we demonstrated what may be the practical consequences of use of LD computed in a reference population instead of true unknown LD in the context of COJO analysis.