An Efficient Algorithm to Perform Multiple Testing in Epistasis Screening

Francois Van Lishout*, Tom Cattaert, Jestinah M. Mahachie John, Elena Gusareva,
Victor Urrea, Isabelle Cleynen, Emilie Theatre, Benoît Charloteaux, Alex Kvasz,
M. Luz Calle, Louis Wehenkel, Kristel Van Steen

Systems and Modeling Unit, Montefiore Institute, University of Liege, 4000 Liege, Belgium
AND Bioinformatics and Modeling, GIGA-R, University of Liege, 4000 Liege, Belgium

Contact: f.vanlishout@ulg.ac.be

Background: Research in epistasis or gene-gene interaction detection for human complex traits has grown exponentially over the last few years. It has been marked by promising methodological developments, improved translation efforts of statistical epistasis to biological epistasis and attempts to integrate different omics information sources into the epistasis screening to enhance power. The quest for gene-gene interactions poses severe multiple-testing problems. In this context, the maxT algorithm is one technique to control the false-positive rate. However, the memory needed by this algorithm rises linearly with the amount of hypothesis tests. In main- effects detection, this is not a problem since the memory required is thus proportional to the number of SNPs. In contrast, gene-gene interaction studies will require a memory proportional to the squared amount of SNPs. A genome wide epistasis would therefore require terabytes of memory. Hence, cache problems are likely to occur, increasing the computation time.

Methods: In this work we present a new version of maxT, requiring an amount of memory independent from the number of genetic effects to be investigated. This algorithm was implemented in C++ in our epistasis screening software MB-MDR- 2.6.2 and compared to MB-MDR's first implementation as an R-package (Calle et al., Bioinformatics 2010). We evaluate the new implementation in terms of memory efficiency and speed using simulated data. The software is illustrated on real-life data for Crohn's disease.

Results: The sequential version of MBMDR-2.6.2 is approximately 5,500 times faster than its R counterparts. The parallel version (tested on a cluster composed of 14 blades, containing each 4 quad-cores Intel Xeon CPU E5520@2.27 GHz) is approximately 900,000 times faster than the latter, for results of the same quality on the simulated data. It analyses all gene-gene interactions of a dataset of 100,000 SNPs typed on 1000 individuals within 4 days. Our program found 14 SNP-SNP interactions with a p-value less than 0.05 on the real-life Crohn_s disease data.

Conclusions: Our software is able to solve large-scale SNP-SNP interactions problems within a few days, without using much memory. A new implementation to reach genome wide epistasis screening is under construction. In the context of Crohn's disease, MBMDR-2.6.2 found signal in regions well known in the field and our results could be explained from a biological point of view. This demonstrates the power of our software to find relevant phenotype-genotype associations.