STATISTICAL METHODS TO ANALYZE HIGH DIMENSIONAL GENOMIC DATA: A SIMULATION STUDY

J.R. González[†1], M. Gratacòs[1], L. Armengol[1], X. Estivill[1]

[1] *Barcelona Node - National Genotyping Center (CeGen), Center for Genomic Regulation, and University Pompeu Fabra, Barcelona, Spain*

[†] E-mail: *juanramon.gonzalez@crg.es*

Recent advances in high-throughput genotyping techniques have increased the possibilities to carry out a large number of studies to associate single-nucleotide polymorphism (SNPs) to clinical outcomes. Understanding how these SNPs relate to complex diseases helps us understand the genetic contribution to those disorders. Biomedical researchers are mainly interested in determining whether interactions between SNPs are associated with the outcome. Several methods have been proposed to analyze SNP contribution to disease, such as Neural Networks, Combinatorial Partitioning methods, or Logistic Regression with Stepwise model selection among others. Maybe the most promising tool is a new adaptive regression methodology, called Logic Regression. Nonetheless, this methodology has still some drawbacks such as how to deal with missing data obtained from genotyping platforms. So far, a more efficient way to analyze this data remains to be established. We propose an alternative methodology based on Bagging and tree-based Boosting methods. In such methods a simulation study is carried out to assess the usefulness of the new proposed method and to compare its performance with respect to other techniques, in particular with Logic Regression. Finally, these approaches are compared using real data arising from a case-control study of patients with psychiatric disorders, where information about 1,300 SNPs from several genes has been obtained.