Institut Pasteur
25 rue du Docteur Roux
75015 Paris

Septembre 2019

Advanced association mapping in massive multivariate genetic data

Context

Public health research on the genetic and environmental causes of common human diseases has fully embraced the era of Big Data. For example, the UK Biobank, the largest human genetic cohort to date, includes half a million individuals with genetic data for almost 100 million variants, and thousands of phenotypic (disease status and quantitative traits) and environmental variables. These data offer great opportunities to decipher part of the complex mechanisms underlying human diseases. However, the analysis of these data faces major methodological and computational challenges. Indeed, despite a strong multidimensional component, the predominant strategy remains univariate analysis –i.e. testing of a single predictor and a single dependent variable at a time. The main advantage of this approach is its robustness and simplicity, and genome-wide univariate association studies (GWAS) have been shown to be effective in identifying many genetic variants associated with quantitative traits and multifactorial diseases. Conversely, integrative and multivariate analysis methods (e.g. the joint analysis of several predictors or several dependent variables), although potentially more powerful, are applied only marginally because of their computational cost and potential challenging in interpreting the results.

Objectives

Our team has developed a new approach named *CMS*[1] that combine the advantages of univariate and multivariate approaches for association studies in multidimensional data sets. The superior performance of this method over the existing ones has been demonstrated on both simulated and real data[2]. The approach will now be implemented in the UK Biobank for the analysis of hundreds of outcomes. However, one of the significant limitations in systematically searching these data using our method is the heavy computational burden –potentially over several months on a large-scale cluster (e.g.> 2000 cores). To solve this issue, we started to develop a new implementation of the algorithm in C++, optimizing both the algorithm itself, and its implementation.

The main objective of the internship is to collaborate on this development and to lead the real data application in the UK Biobank. The main steps of the internship will be:

- Understanding the main steps of the algorithm
- Identifying potential improvement in the method and its implementation
- Updating the algorithm and validating the improvement through simulations
- Applying the developed approach for real data analysis in the UK Biobank.

Application

The internship is intended for Master2 students and 3rd year engineers school students. Interested candidates should send a CV and contact information of at least one referee to Dr. Hugues Aschard (hugues.aschard@pasteur.fr). Advance knowledge in statistics and applied mathematics are required. Good coding skills in C/C++ are also mandatory.

Information about the Institut Pasteur and Dr. Aschard team are available here http://www.pasteur.fr/en and here https://research.pasteur.fr/en/team/statistical-genetics/.

[1] *Covariate Selection for Association Screening in Multi-Phenotype Genetic studies. Aschard et al, 2017 Nature Genetics*
[2] *A comprehensive study of metabolite genetics reveals strong pleiotropy and heterogeneity across time and context. Gallois et al, in press in Nature Communication.*