Predicting complex diseases: performance and robustness

Florian Privé [1,*], Hugues Aschard [2,3] and Michael G.B. Blum [1,*]

[1] Université Grenoble Alpes, CNRS, Laboratoire TIMC-IMAG, UMR 5525, France,

[2] Centre de Bioinformatique, Biostatistique et Biologie Intégrative (C3BI), Institut Pasteur, Paris, France

[3] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA.

[*] To whom correspondence should be addressed.

Abstract

Motivation:

Results:

Availability: **Contact:** florian.prive@univ-grenoble-alpes.fr & michael.blum@univ-grenoble-alpes.fr

Supplementary information:

1 Introduction

Polygenic Risk Scores (PRSs) combine the information contained in many single-nucleotide polymorphisms (SNPs) in a score that should reflect the risk of developing diseases. PRSs have proven to be useful for different applications, such as .... Personalized medicine is one of the major applications of PRSs. Personalized medicine will use PRSs in screening campains in order to identify people at higher risk for a given disease. ...This could lead to prevention for these identified people... Personalized medicine with also use PRSs for helping in the diagnostic of disease, either confirming a diagnostic or even early diagnostic. Yet, diagnostic based on PRSs won't be possible unless these PRSs shows a high discrimitative power between cases and controls [REF]. Many methods have been developed in order to maximize the predictive power of genotypes for diseases, or more generally phenotypes. A commonly used technique, called P+T (which stands for "Pruning + Thresholding") or genetic profiling, is used to derive a PRS from results of Genome-Wide Association Study (GWAS). This technique only use summary statistics which make it very ..suitable.. and also very fast. Linear Mixed-Models are also widely-used in fields such as plant and animal breeding or for predicting highly heritable quantitative human phenotypes such as height [REF]. Yet, these models are not ..constructed.. for predicting a binary trait such as a disease and have proven to ..fail.. at such in another comparative study [REF GAD]. Moreover, these methods and their derivations are often computationally demanding and won't be usable for largest cohorts to data [REFS]. Statistical learning methods such as logistic regression, Support Vector Machine (SVM) or random forests were also used to derive PRSs for complex human disease.

We recently developed two R packages, bigstatsr and bigsnpr, for efficient management and analysis of large-scale genome-wide data. This include efficient functions for computing penalized linear and logistic regressions on huge datasets. In this paper, we present a comprehensive analysis of the P+T method, our penalized logistic regression and the T-Trees algorithm, which is a derivation of random forests and has given exceptionally good results in its corresponding paper [REF BOTTA]. Note that SVM is expected to give similar results to logistic regression [REF GAD], and therefore isn't added to the comparison. For the P+T model, we compare different th

The aim of PRSsis to predict well enough the sfor predicting risk of disease based on genotypes First, they are useful in order to determine if combining genotypes can be used Personalized medicine may not be here yet, but we're not far. The scientific community need to continue to search for improving predicting performance based on genotypes, and maybe other factors.

2 Methods

3 Results

Acknowledgements

Authors acknowledge Grenoble Alpes Data Institute, supported by the French National Research Agency under the “Investissements d?avenir” program (ANR-15-IDEX-02) and the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01).

Supplementary Data