Using genetic data to predict disease status based on statistical learning

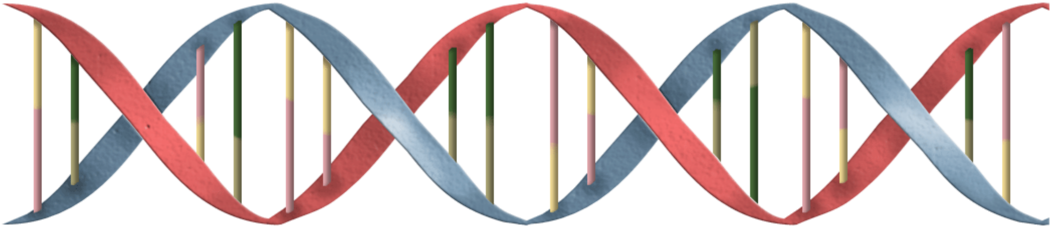
Florian Privé (TIMC-IMAG - BCM)

FADEX IA & Health - June 27, 2018

Introduction

The data I work with: very large genotype matrices

- Each variable (column): number of mutations for **one position of the genome** (generally between 100,000 to several millions) -> **ultra-high dimensional** data

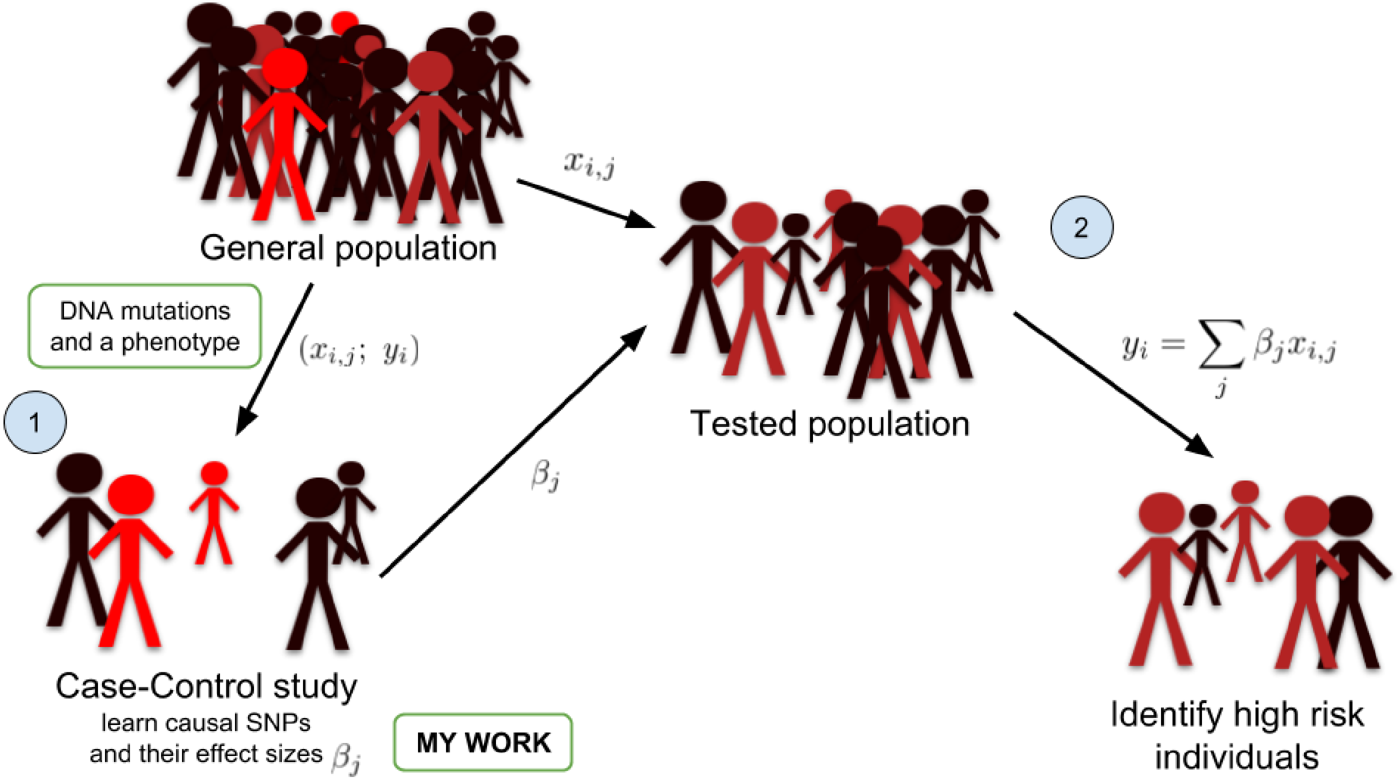


- Each observation (row): one individual (generally between 1000 and 1M)

Example of a dataset I previously worked with: 15K x 280K, celiac disease (~30GB)

Polygenic Risk Scores (PRS) for predictive medicine

Application: to identify high risk individuals


General population
$x_{i,j}$
DNA mutations and a phenotype
$(x_{i,j};\ y_i)$
1
Tested population
2
$y_i = \sum_j \beta_j x_{i,j}$
β_j
Case-Control study
learn causal SNPs and their effect sizes β_j
MY WORK
Identify high risk individuals

Interest in prediction: polygenic risk scores (PRS)

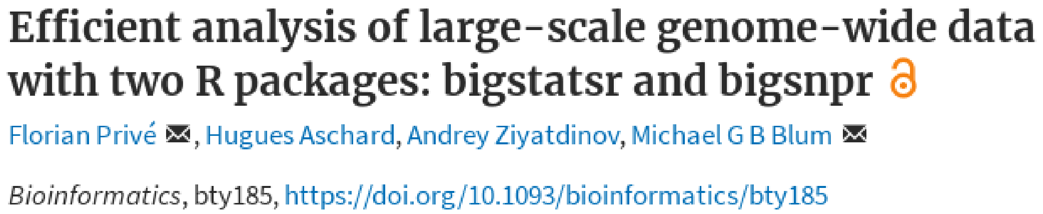
- Wray, Naomi R., Michael E. Goddard, and Peter M. Visscher. "**Prediction of individual genetic risk** to disease from genome-wide association studies." Genome research 17.10 (**2007**): 1520-1528.
- Wray, Naomi R., et al. "Pitfalls of **predicting complex traits** from SNPs." Nature Reviews Genetics 14.7 (**2013**): 507.
- Dudbridge, Frank. "Power and **predictive accuracy of polygenic risk scores**." PLoS genetics 9.3 (**2013**): e1003348.
- Chatterjee, Nilanjan, Jianxin Shi, and Montserrat García-Closas. "Developing and evaluating **polygenic risk prediction** models for stratified disease prevention." Nature Reviews Genetics 17.7 (**2016**): 392.
- Martin, Alicia R., et al. "Human demographic history impacts **genetic risk prediction** across diverse populations." The American Journal of Human Genetics 100.4 (**2017**): 635-649.

Still a gap between current predictions and clinical utility.
Need more optimal predictions + larger sample sizes.

How to analyze large genomic data?

Our two R packages: bigstatsr and bigsnpr

Statistical tools with big matrices stored on disk



Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr

Florian Privé, Hugues Aschard, Andrey Ziyatdinov, Michael G B Blum

Bioinformatics, bty185, https://doi.org/10.1093/bioinformatics/bty185

- {bigstatsr} for many types of matrix, to be used by any field of research
- {bigsnpr} for functions that are specific to the analysis of genetic data

Package {bigstatsr} provides fast PCA, association and predictive models, etc.

How to predict disease status based on genotypes?

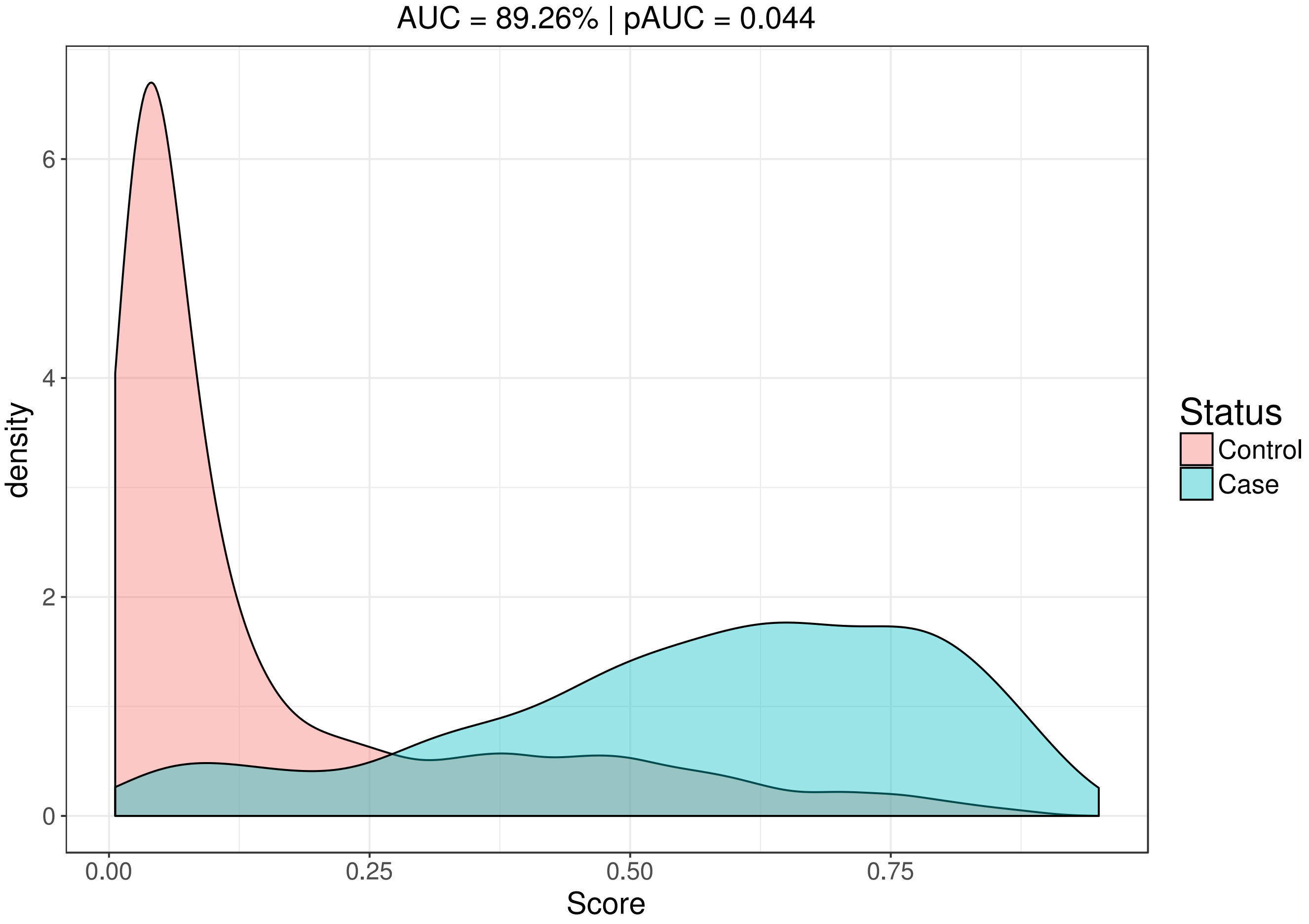
Penalized logistic regression

We are developing an **efficient implementation** for this problem:

$$\operatorname*{argmin}_{\beta_0,\, \beta}(\lambda, \alpha) \left\{ \underbrace{-\sum_{i=1}^{n} \left(y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \right)}_{\text{Loss function}} + \underbrace{\lambda \left((1 - \alpha) \frac{1}{2} \|\beta\|_2^2 + \alpha \|\beta\|_1 \right)}_{\text{Penalization}} \right\}$$

- x is denoting the genotypes and covariables (e.g. principal components),
- y is the disease status we want to predict,
- λ is a regularization parameter that needs to be determined and
- α determines relative parts of the regularization $0 \le \alpha \le 1$.

Predict Celiac disease



Thanks!

Presentation available at

https://privefl.github.io/thesis-docs/FADEX.html

privefl   privefl   F. Privé

Slides created via the R package **xaringan**.