High-dimensional data: a different kind of big data

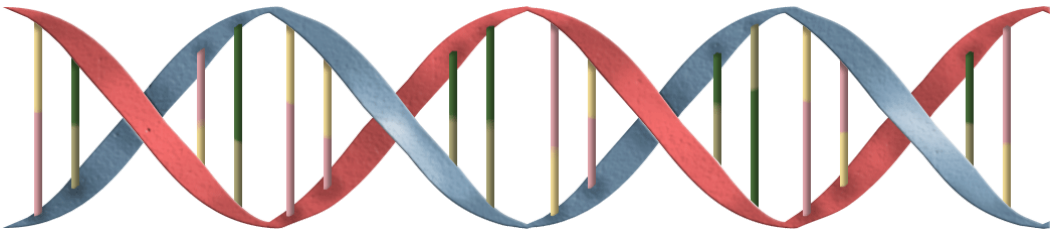
Florian Privé

Data Club - 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



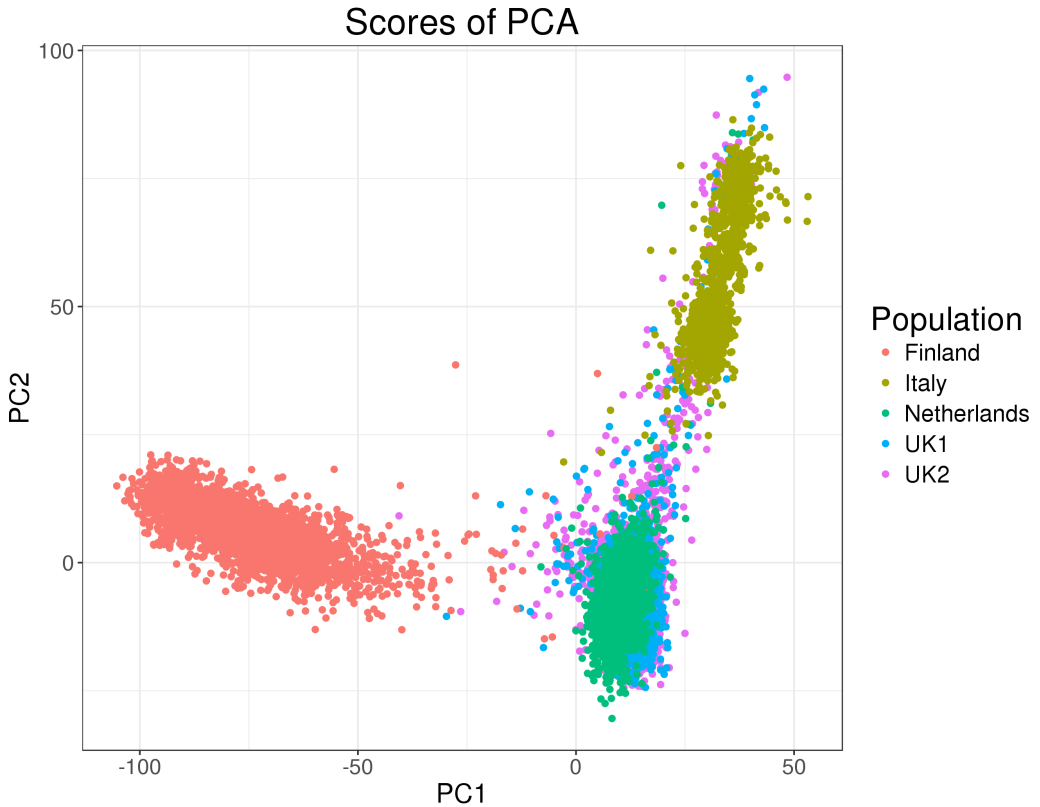
- Each observation (row): one individual (generally between 1000 and 1M)

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

What types of analysis we do?

(and how?)

Principal Component Analysis (PCA) captures population structure



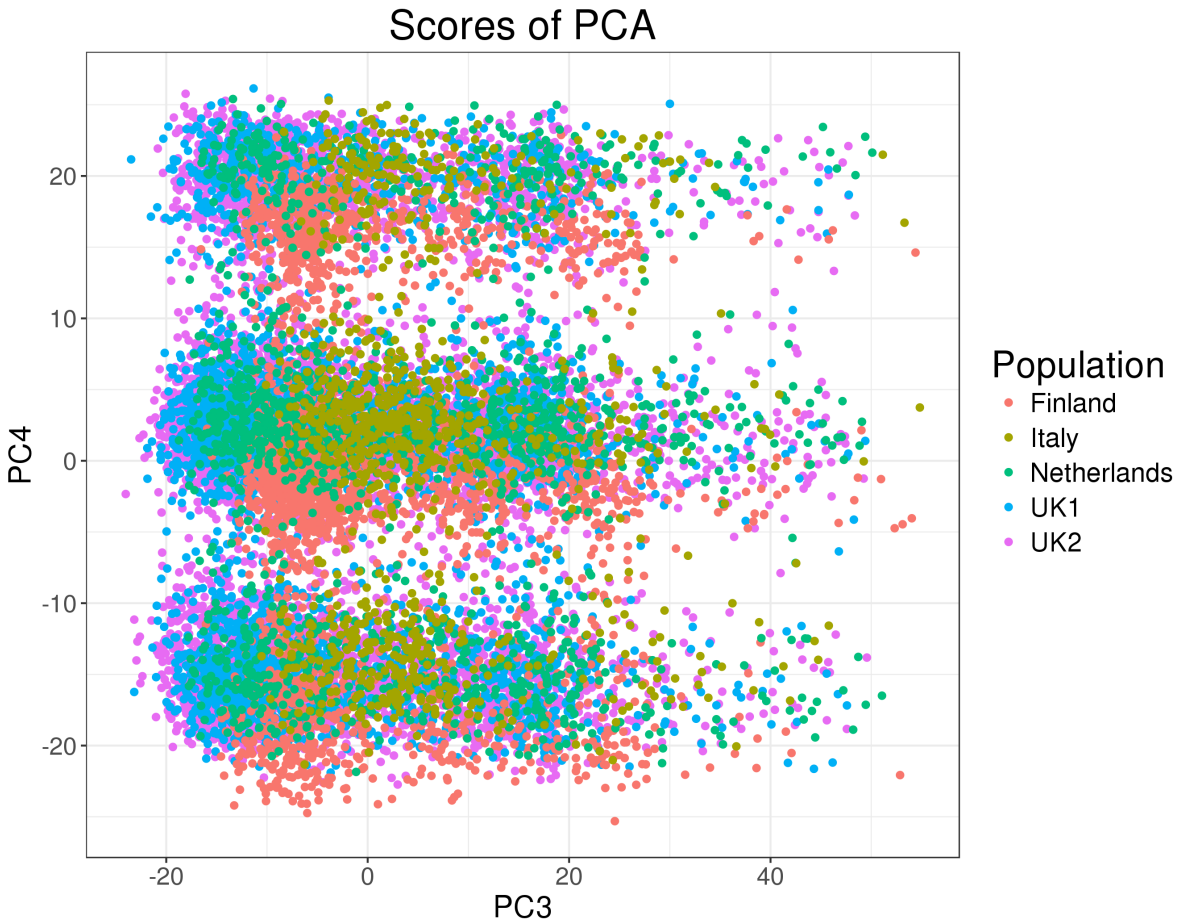
Partial PCA algorithm

You have a matrix X with n observations (rows) and p variables (columns).

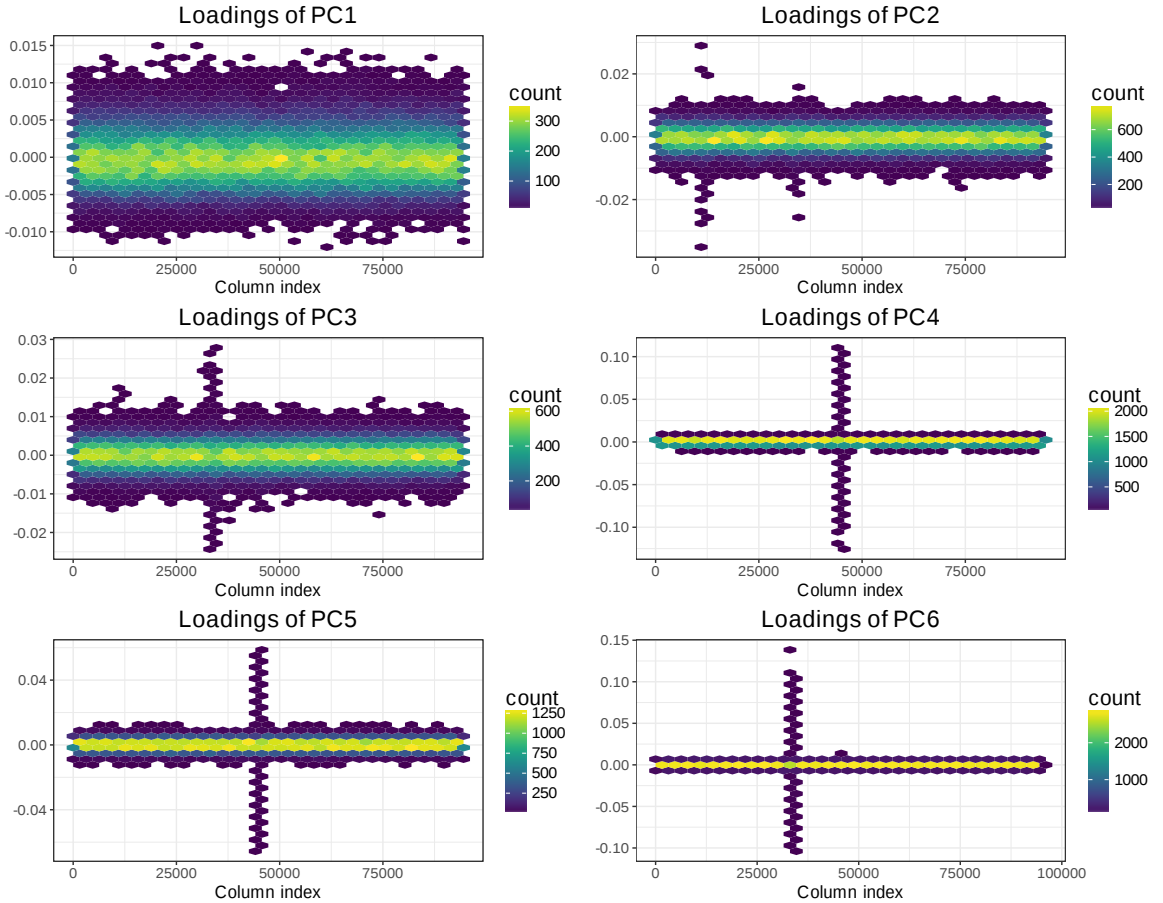
- Usually, p is small and you can get the SVD from the eigen decomposition of $X^T X$. In bioinformatics, p is usually too large, so you can't use this standard algorithm.
- If n is not to large (say $n < 10,000$), you can use the eigen decomposition of XX^T instead (an $n \times n$ matrix).
- **Now, n is also large** (both dimensions of the matrix are large), so we use algorithms based on random projections to get first PCs (usually we are interested in only first 10-20 PCs).

With my implementation, you can get first 10 PCs of a 15K x 100K matrix in one minute only.

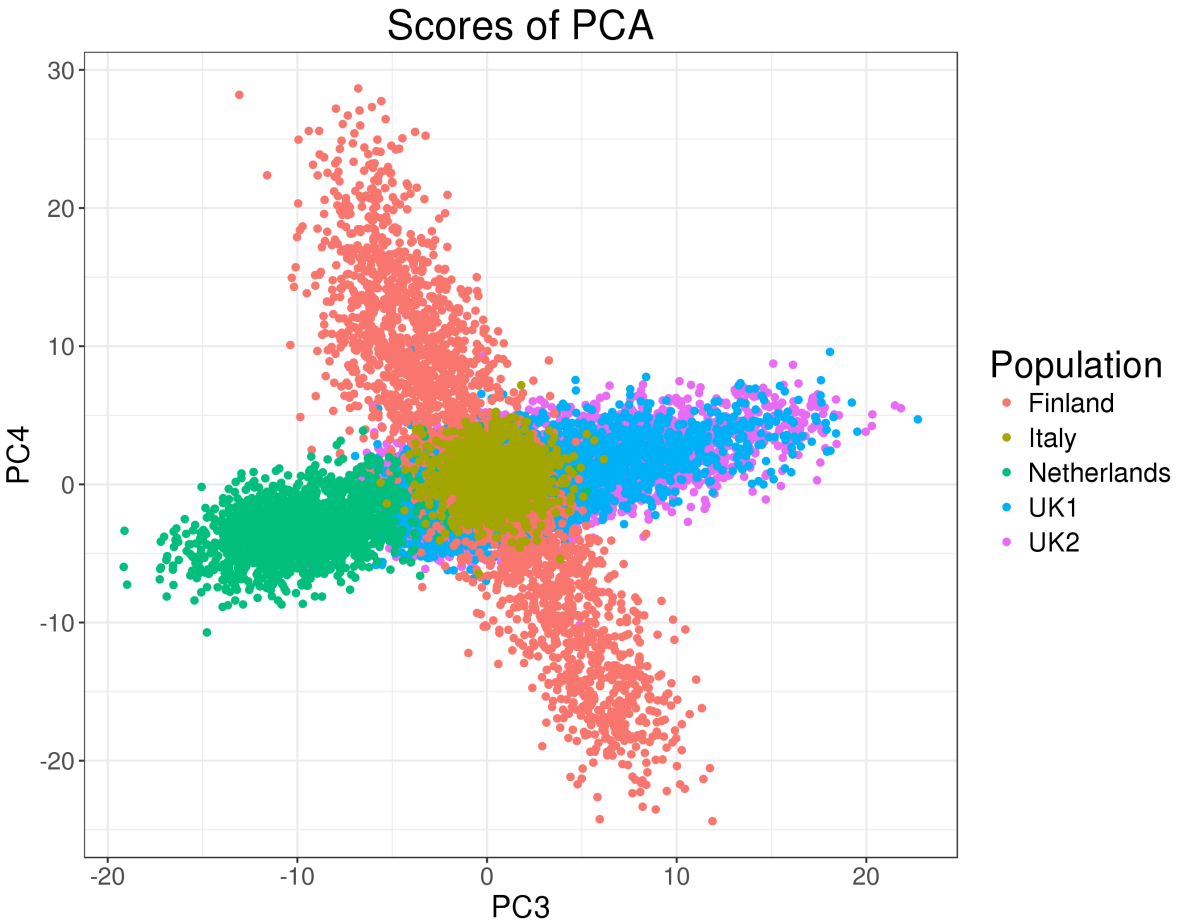
Still, we can't do PCA naively



Cause of the problem



After some filtering



Genome-wide association studies

For linear regression, a t-test is performed **for each variable** j on $\beta^{(j)}$ where

$$\hat{y} = \alpha^{(j)} + \beta^{(j)} X^{(j)} + \gamma_1^{(j)} PC_1 + \cdots + \gamma_K^{(j)} PC_K \\ + \delta_1^{(j)} COV_1 + \cdots + \delta_K^{(j)} COV_L \, ,$$

and K is the number of principal components and L is the number of other covariates (such as age and gender).

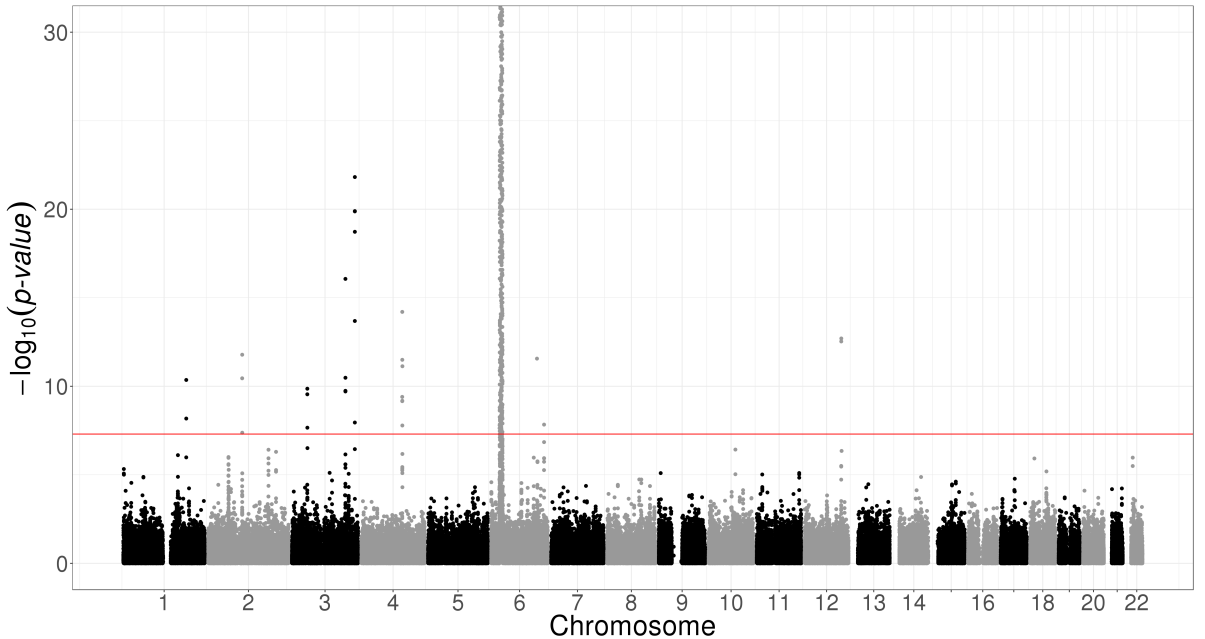
Similarly, for logistic regression, a Z-test is performed for each variable j on $\beta^{(j)}$ where

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \alpha^{(j)} + \beta^{(j)} X^{(j)} + \gamma_1^{(j)} PC_1 + \cdots + \gamma_K^{(j)} PC_K \\ + \delta_1^{(j)} COV_1 + \cdots + \delta_K^{(j)} COV_L \, ,$$

and $\hat{p} = \mathbb{P}(Y = 1)$ and Y denotes the binary phenotype.

Genome-wide association studies

Which genes are associated with the disease?



Here, you do ~1M tests, so beware **multiple testing**!

Prediction

Can you fit a statistical learning model when you have more variables than observations ($n > p$)?

Quiz

How can you fit a prediction model when you have too many variables?

Regularization / Penalization

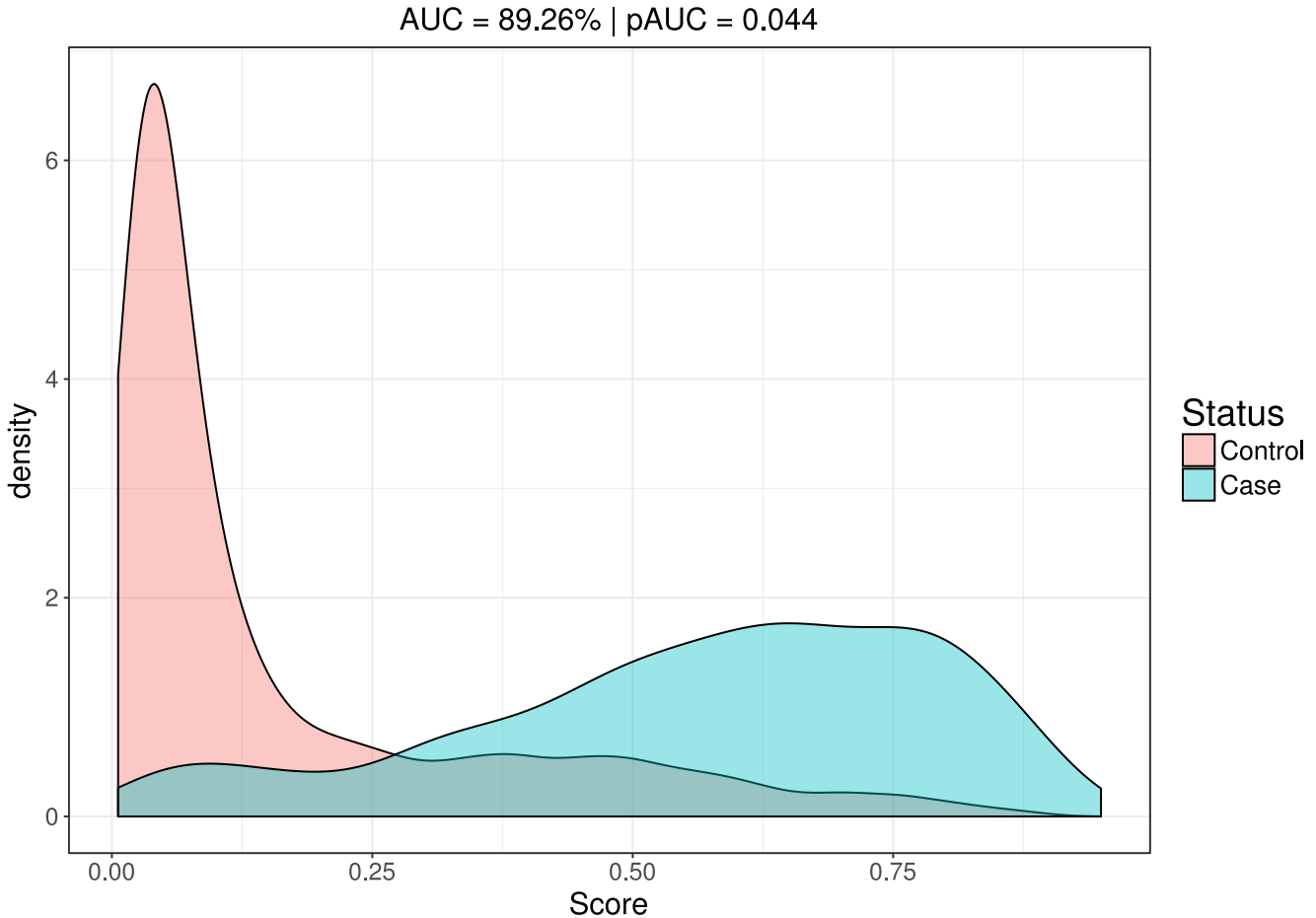
Minimize

$$F(\lambda, \alpha) = \text{Loss function} + \underbrace{\lambda \left((1-\alpha) \frac{1}{2} \|\beta\|_2^2 + \alpha \|\beta\|_1 \right)}_{\text{Penalization}}$$

Different regularizations can be used to make the problem solvable and to prevent overfitting:

- the L2-regularization ("ridge") shrinks coefficients and is ideal if there are many predictors drawn from a Gaussian distribution (corresponds to $\alpha = 0$ in the previous equation)
- the L1-regularization ("lasso") forces some of the coefficients to be equal to zero and can be used as a means of variable selection, leading to sparse models (corresponds to $\alpha = 1$)
- the L1- and L2-regularization ("elastic-net") is a compromise between the two previous penalties and is particularly useful in the $p \gg n$ situation, or any situation involving many correlated predictors (corresponds to $0 < \alpha < 1$).

Predict Celiac disease based on penalized logistic regression



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

High-dimensional data
come with their own problems

Data are becoming larger and larger

Will we all need skills in computer science?

Thanks!

Presentation available at

https://privefl.github.io/thesis-docs/data-club.html

privefl privefl F. Privé

Slides created via R package **xaringan**.