Journée des doctorants

Florian Privé

BCM team, Laboratoire TIMC-IMAG

supervised by Michael Blum (BCM) and Hugues Aschard (Institut Pasteur)

March 20, 2018

Outline

1. Main objective of the thesis

2. R packages

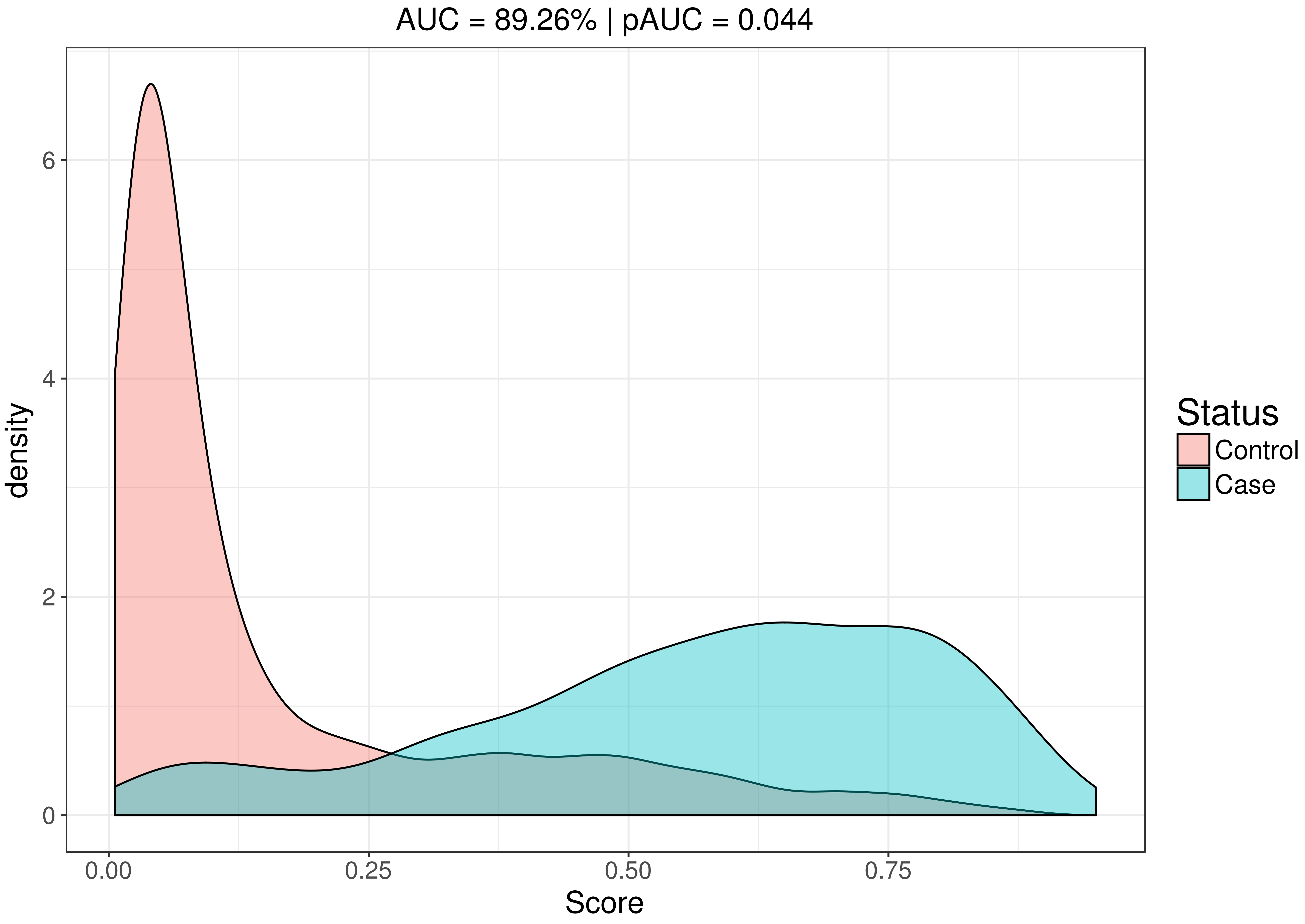
3. Ongoing paper

4. Future work

Main objective

Compute Polygenic Risk Scores (PRS)

in order to differentiate a healthy person from a diseased person



4 main difficulties

- Size of the data (dozens to hundreds of GB)
- Hundreds of thousands of correlated variables (variables with overlapping information)
- Generalization of models on different populations
- Integration of non-genetic data in the models

Big Data

Simpler solutions are easier to implement

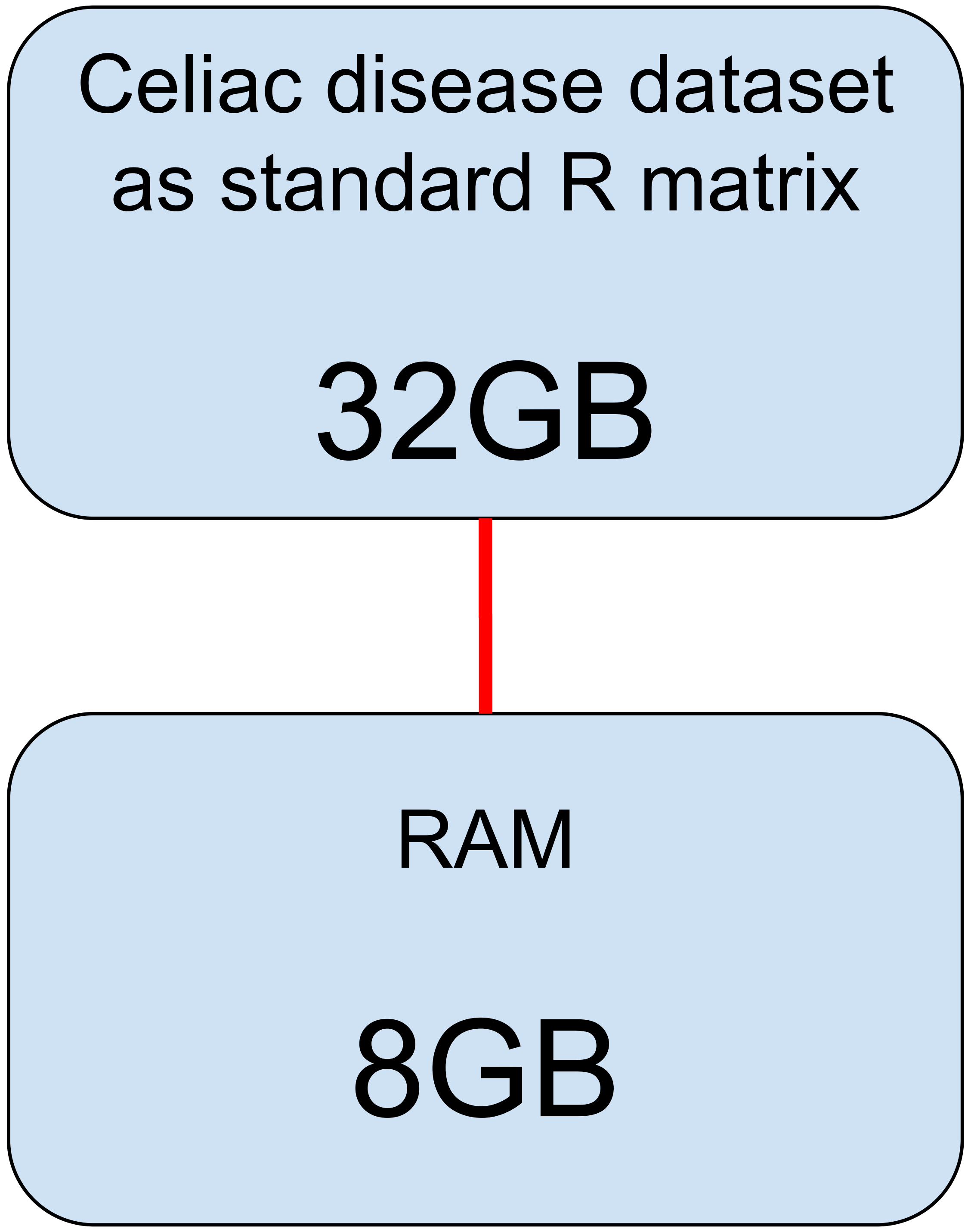
What I want to be able to do

Data analysis on large-scale genotype matrices!

- Be fast to test many ideas quickly
  - code should be fast
  - I shouldn't have to make many conversions
  - it should be easily to combine multiple functions
- Not be restricted in my analysis
  - Basically use all I already know in R
- Work on my computer
  - I have 64 GB of RAM and 12 cores
  - Working on a server is not as easy as on my computer

Smooth and fast analysis!

Memory problem when working in R



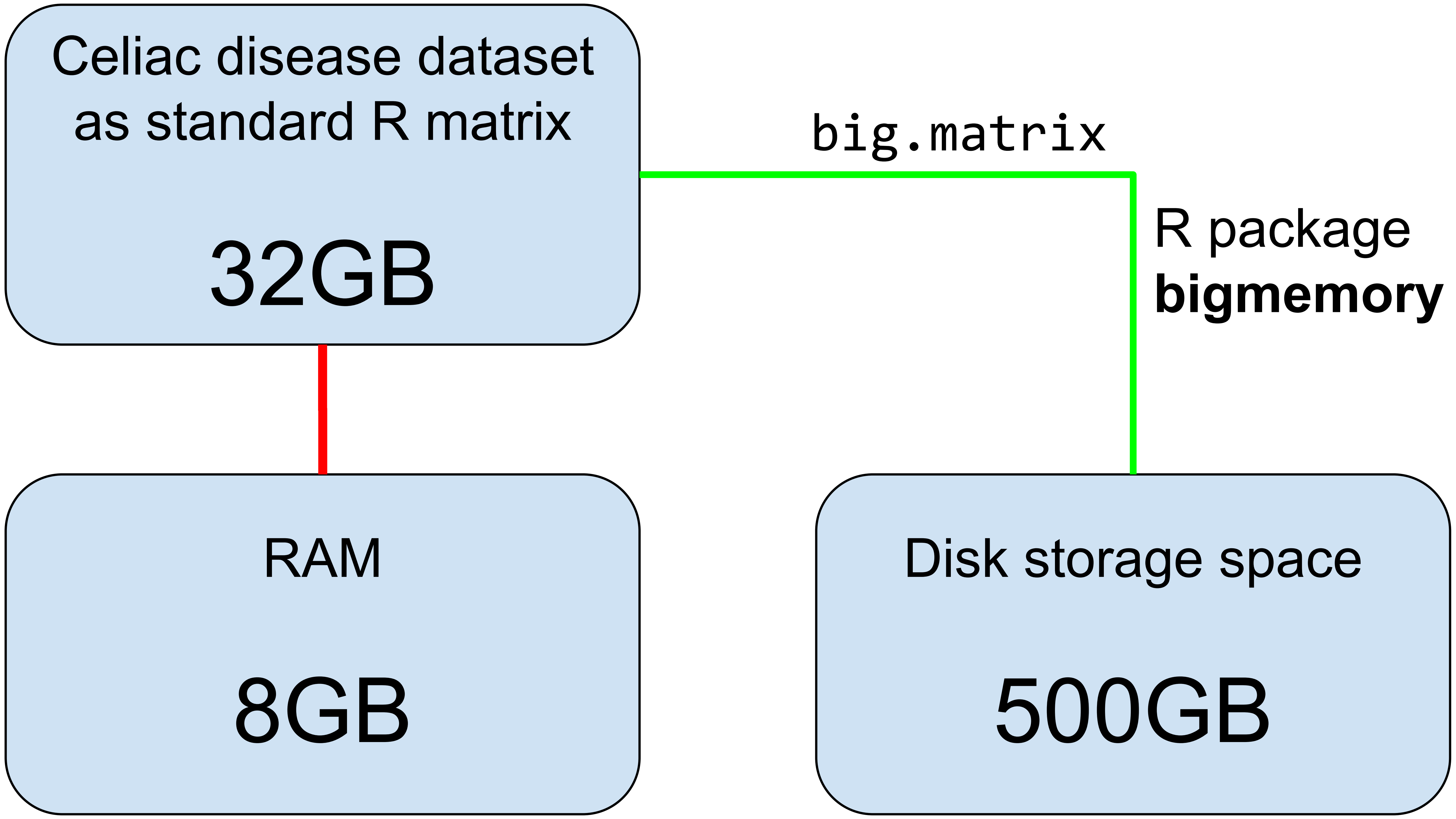
Celiac disease dataset as standard R matrix

32GB

RAM

8GB

Memory solution when working in R

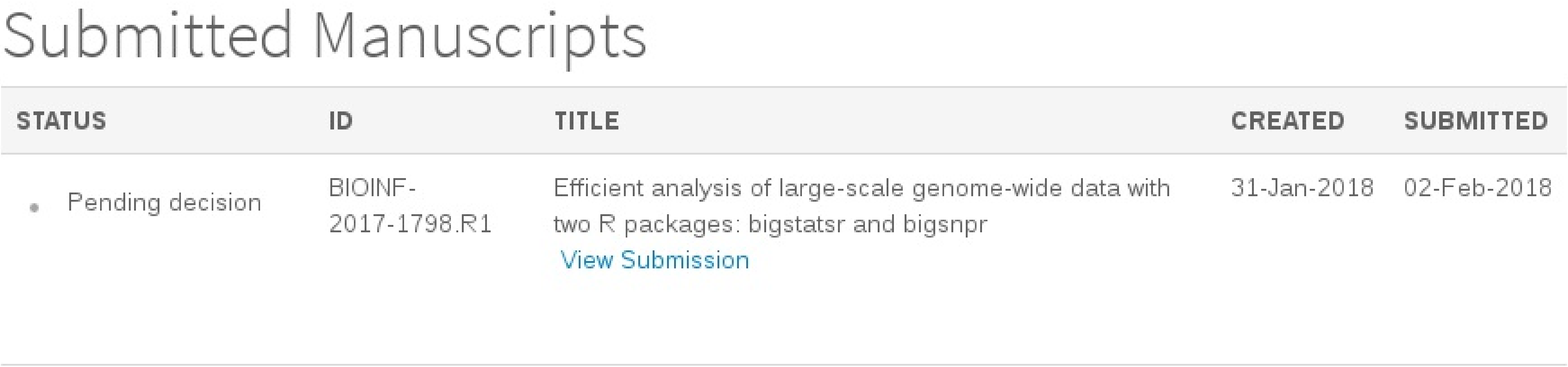


I don't use **bigmemory** anymore but still something very similar.

Two R packages: bigstatsr and bigsnpr

Statistical tools with big matrices stored on disk

- **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

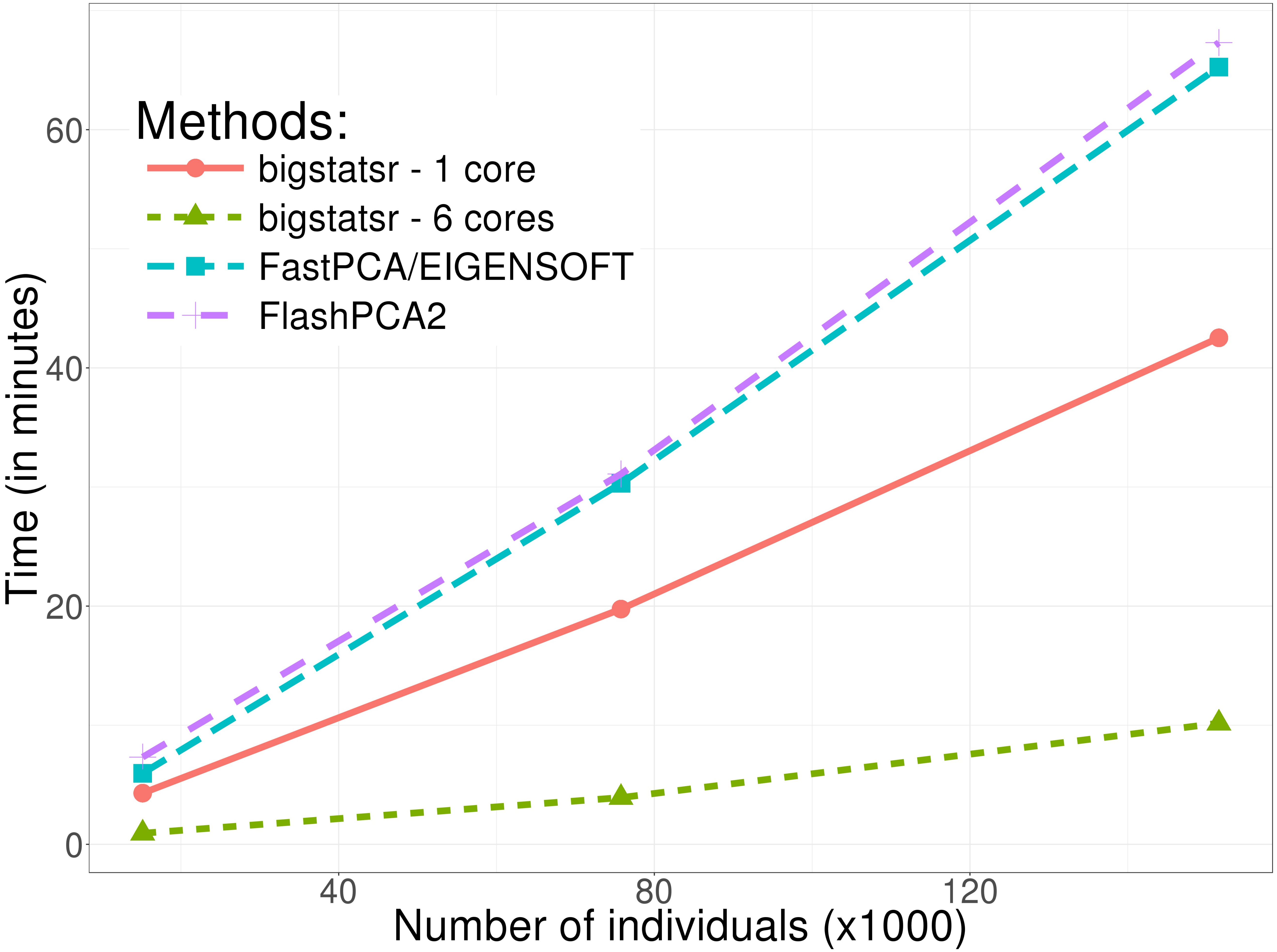


Submitted Manuscripts

STATUS	ID	TITLE	CREATED	SUBMITTED
• Pending decision	BIOINF-2017-1798.R1	Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr View Submission	31-Jan-2018	02-Feb-2018

Comparative performance

Computing partial SVD



Ongoing paper

Comparison of methods for computing PRS

Recall of what we want to achieve

Predict a phenotype: pitfalls of the widely-used model

- Weigths learned independently
- Correlation is taken care of heuristically
- Regularization is taken care of heuristically

A better solution?

We can use **statistical learning methods**.

For example, we can use logistic regression on all variables at once by using a clever implementation.

Future work

UK Biobank

UK Biobank

It is an extremely large dataset with

- genetic data
- clinical data
- environmental data

Prospects

- training in one population to improve training and prediction in another population
- assess how can we combine the information provided by genetic data with clinical and environmental data, possibly in a non-linear way

Thanks!

Presentation available at

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

privefl privefl F. Privé

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