

Efficient statistical tools for analyzing omics data

with a focus on polygenic risk prediction

Florian Privé

Aarhus - January 2019

Introduction & Motivation

About

I'm a PhD Student (2016-2019) in **Predictive Human Genetics** in Grenoble.

$$\text{Disease} \sim \text{DNA mutations} + \dots$$



Introduction

Data

Matrices of Single Nucleotide Polymorphisms (SNPs)

counting the number of alternative alleles (**0, 1, or 2**)

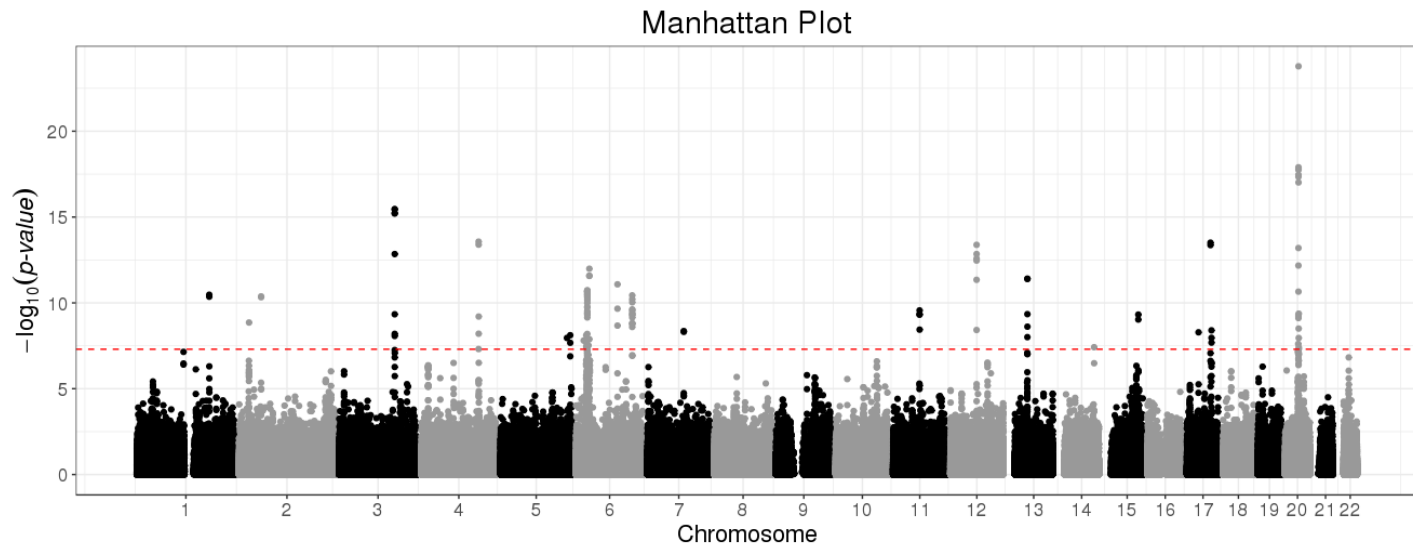
for each individual (row) and each genome position (column)

+ some phenotype(s) (e.g. disease status you want to predict)

+ other metadata

$$\boxed{\text{Disease} \sim \text{DNA mutations} + \dots}$$

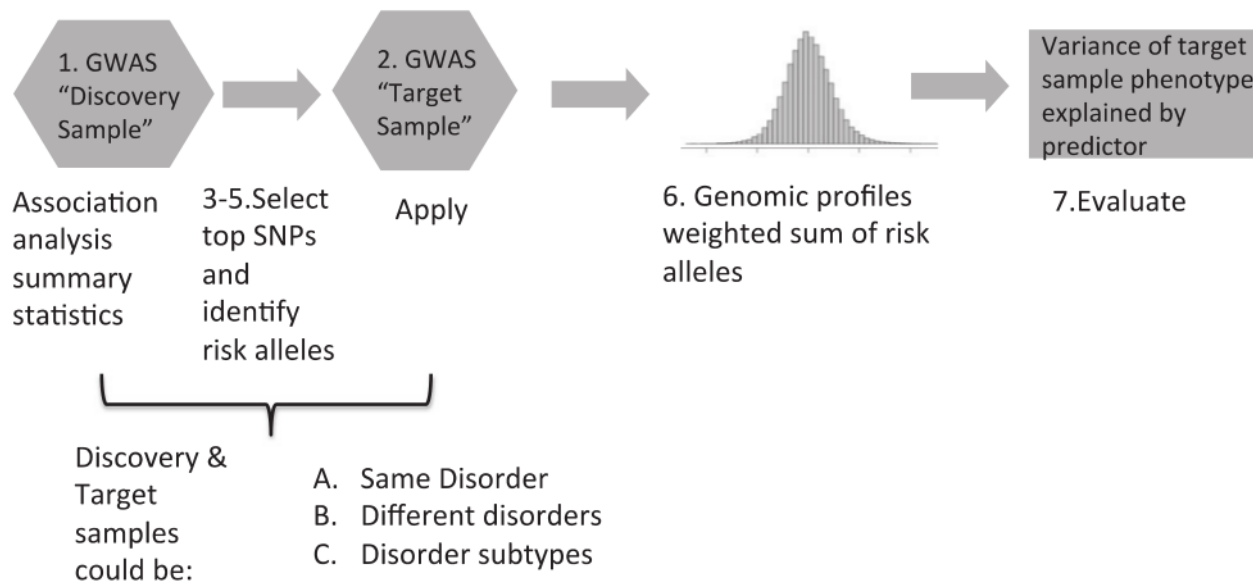
From genome-wide association studies (GWAS) to polygenic risk scores (PRS)



$$PRS_i = \sum_{\substack{j \in S \\ p_j < p_T}} \hat{\beta}_j \cdot G_{i,j}$$

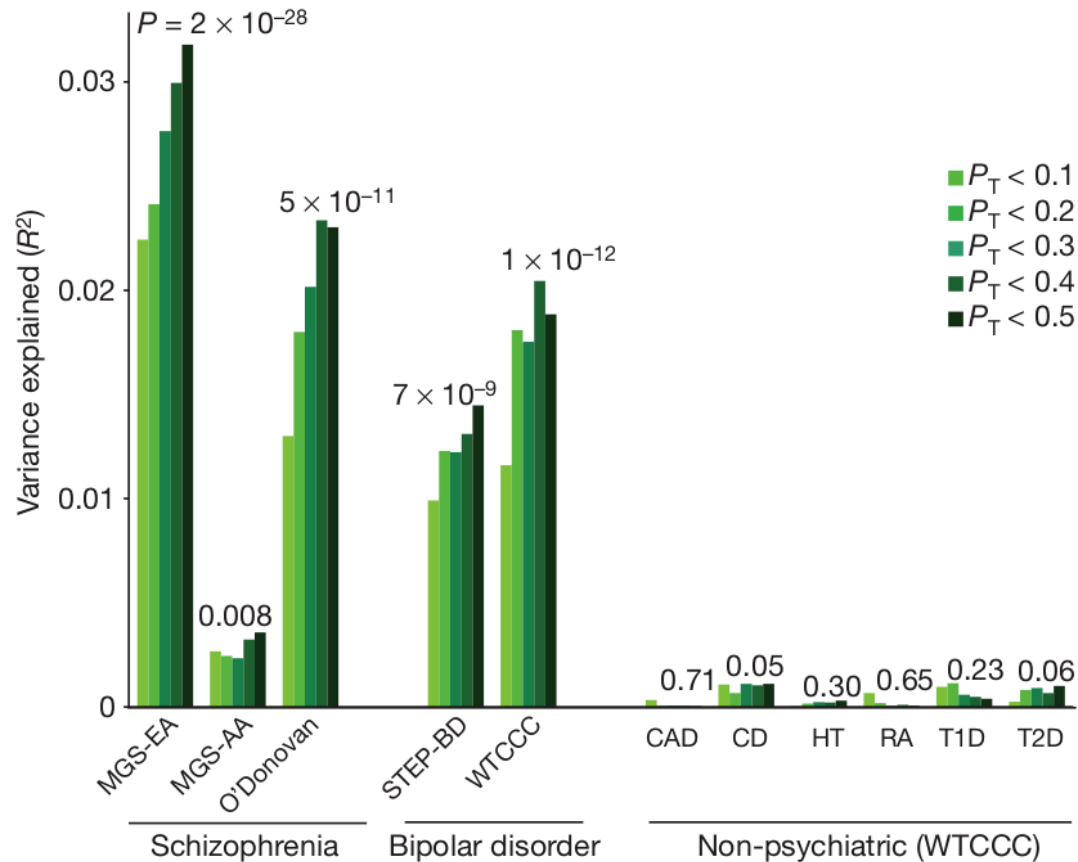
Polygenic Risk Scores (PRS) for epidemiology

One application: to provide evidence for a polygenic contribution to a trait or a shared polygenic relationship between traits.



Source: 10.1111/jcpp.12295

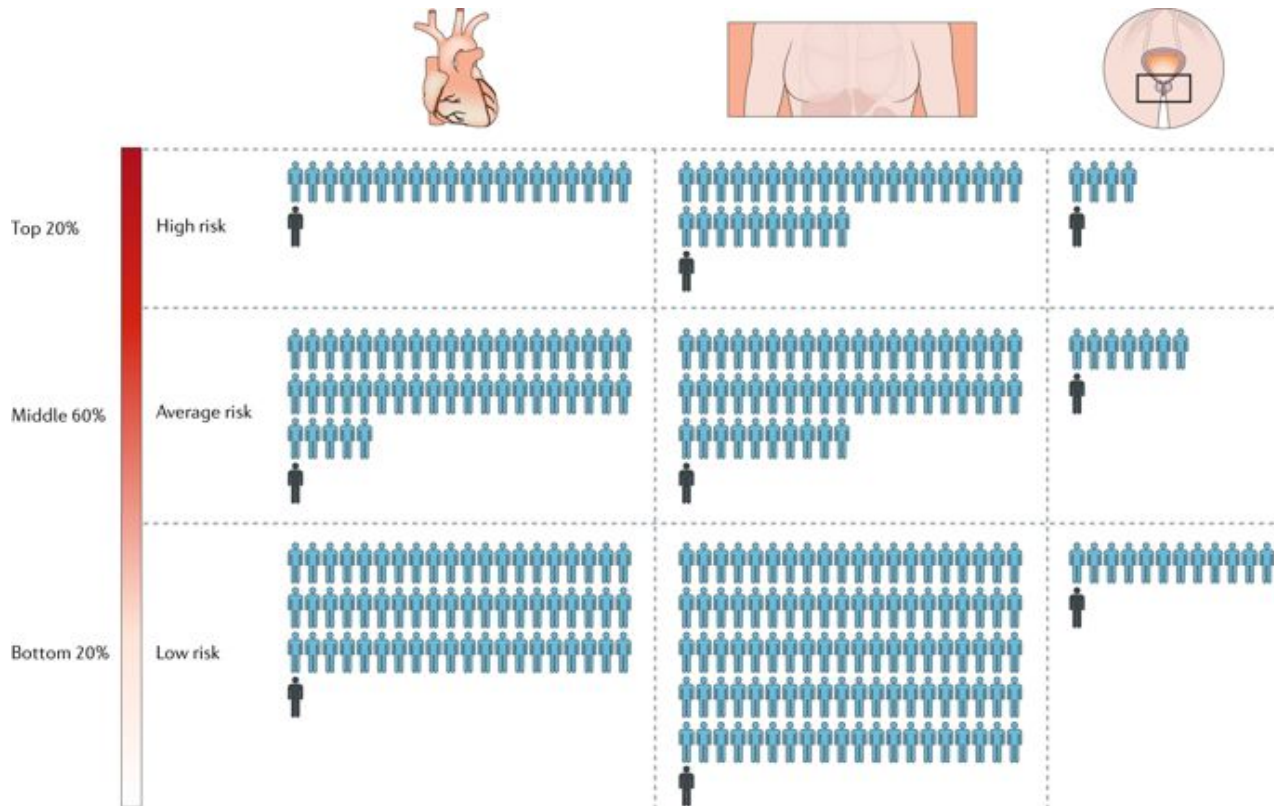
Polygenic Risk Scores (PRS) for epidemiology



Source: [10.1038/nature08185](https://doi.org/10.1038/nature08185)

Polygenic Risk Scores (PRS) for predictive medicine

Another application: to 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.

Very large genotype matrices

- previously: 15K x 280K, **celiac disease** (~30GB)
- currently: 500K x 500K, **UK Biobank** (~2TB)



But I still want to use .

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** 

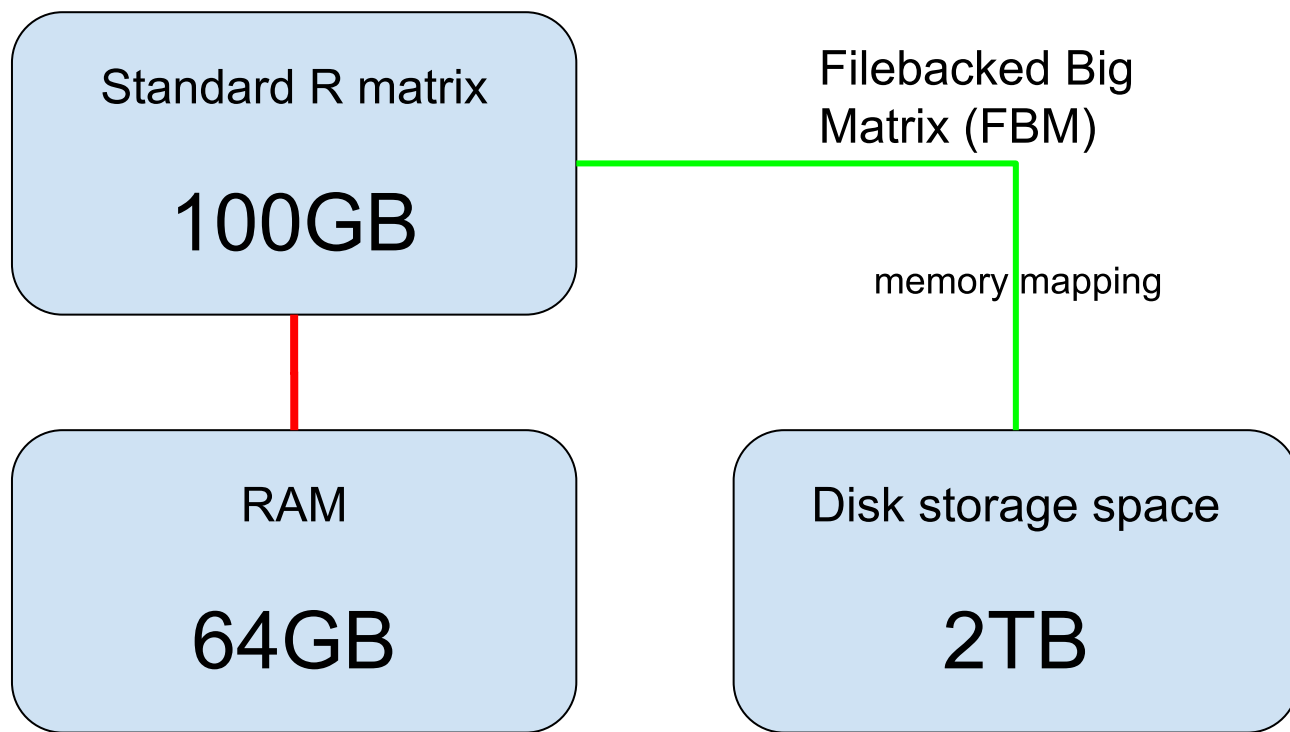
Florian Privé , Hugues Aschard, Andrey Ziyatdinov, Michael G B Blum 

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.

The solution I found



Format FBM is very similar to format `filebacked.big.matrix` from package `{bigmemory}` (details in [this vignette](#)).

Multiple association testing

The idea behind Genome-Wide Association Studies (GWAS) is simple: test each variant one by one for association with the phenotype of interest. For a continuous phenotype (e.g. height), linear regression is used and a t-test is performed for each SNP j on $\beta^{(j)}$ where

$$\hat{y} = \alpha^{(j)} + \beta^{(j)} SNP^{(j)} + \gamma_1^{(j)} COV_1 + \dots + \gamma_K^{(j)} COV_K ,$$

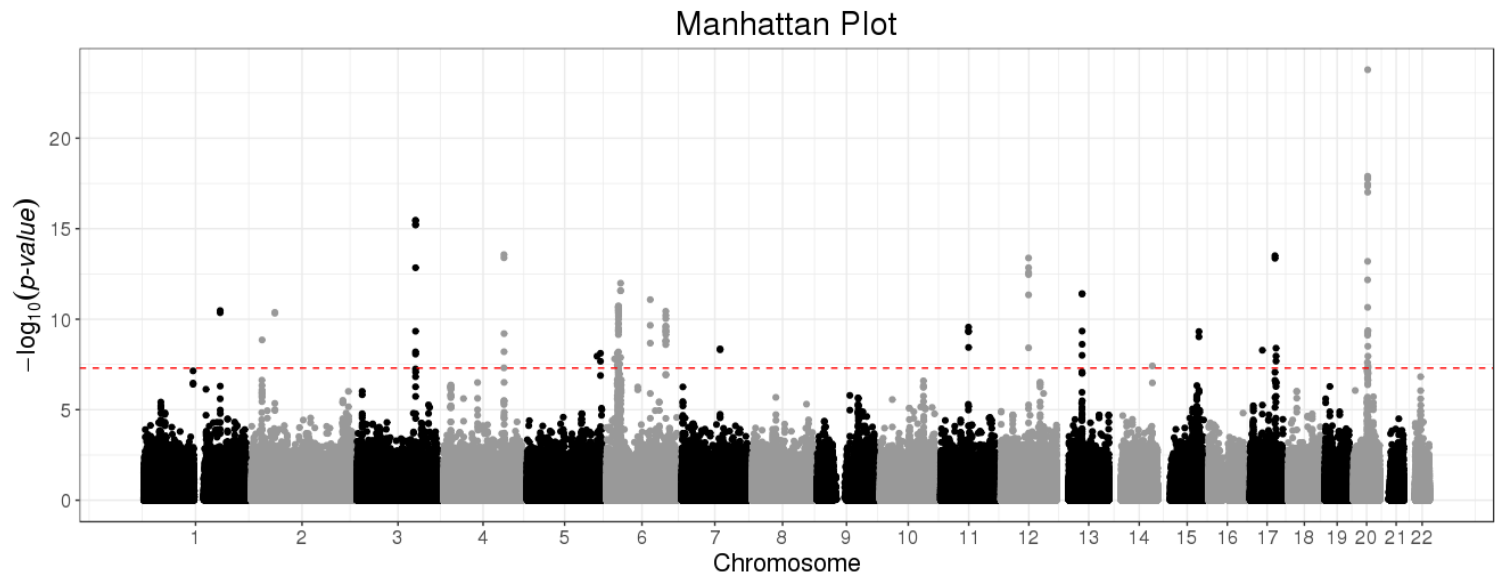
and K is the number of covariables, including **first principal components** and other covariates such as age and gender.

Similarly, for a binary phenotype (e.g. disease status), logistic regression is used and a Z-test is performed for each SNP j on $\beta^{(j)}$ where

$$\log \left(\frac{\hat{p}}{1 - \hat{p}} \right) = \alpha^{(j)} + \beta^{(j)} SNP^{(j)} + \gamma_1^{(j)} COV_1 + \dots + \gamma_K^{(j)} COV_K ,$$

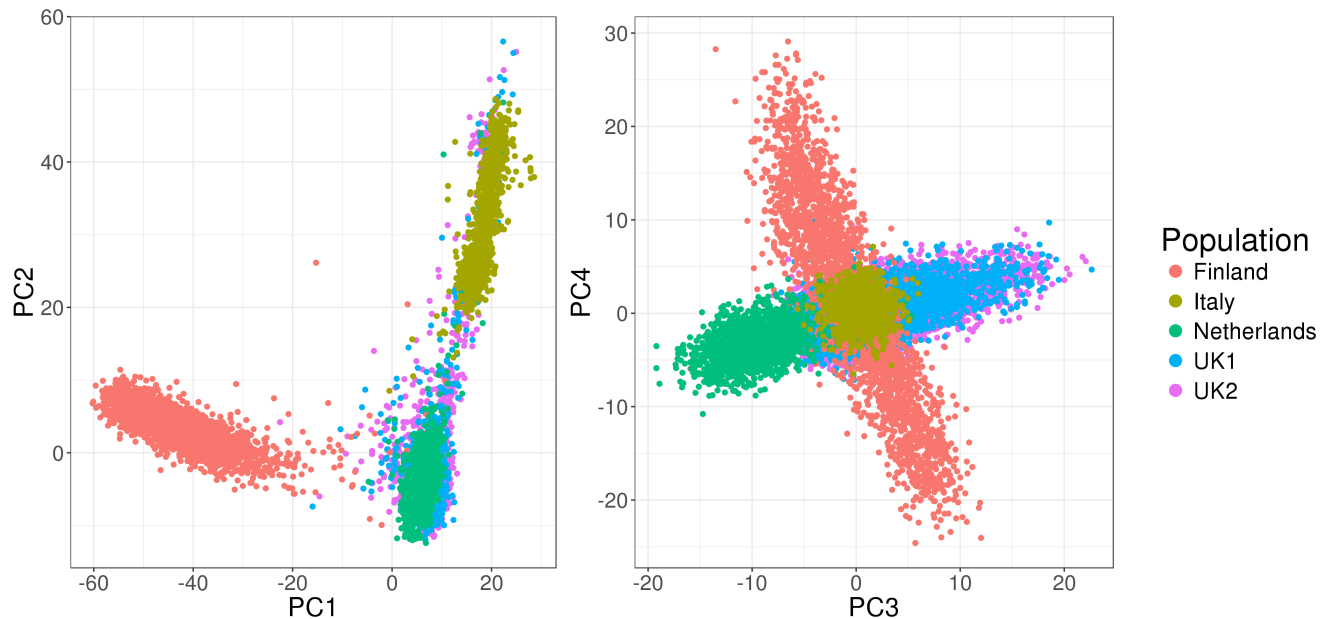
and $\hat{p} = \mathbb{P}(Y = 1)$ and Y denotes the binary phenotype.

Which DNA mutations are associated with one disease?



Partial Singular Value Decomposition

15K \times 100K -- 10 first PCs -- 6 cores -- **1 min** (vs 2h in base R)



Implemented in `big_randomSVD()`, powered by R packages `{RSpectra}` and `{Rcpp}`.

Benchmarks (GWAS)

Operation \ Software	Execution times (in seconds)		
	FastPCA PLINK 1.9	bigstatsr bigsnpr	SNPRelate GWASTools
Converting PLINK files	n/a	6 / 20	13 / 33
Pruning	4 / 4	14 / 52	33 / 32
Computing 10 PCs	305 / 314	58 / 183	323 / 535
GWAS (binary phenotype)	337 / 284	291 / 682	16220 / 17425
GWAS (continuous phenotype)	1348 / 1633	10 / 23	6115 / 7101
Total (binary)	646 / 602	369 / 937	16589 / 18025
Total (continuous)	1657 / 1951	88 / 278	6484 / 7701

Table 1. Execution times with bigstatsr and bigsnpr compared to PLINK 1.9 and FastPCA (EIGENSOFT) and also to R packages SNPRelate and GWASTools for making a GWAS for the Celiac dataset (15,155 individuals and 281,122 SNPs). The first execution time is with a desktop computer (6 cores used and 64GB of RAM) and the second one is with a laptop (2 cores used and 8GB of RAM).

Precision (approximate PCA)

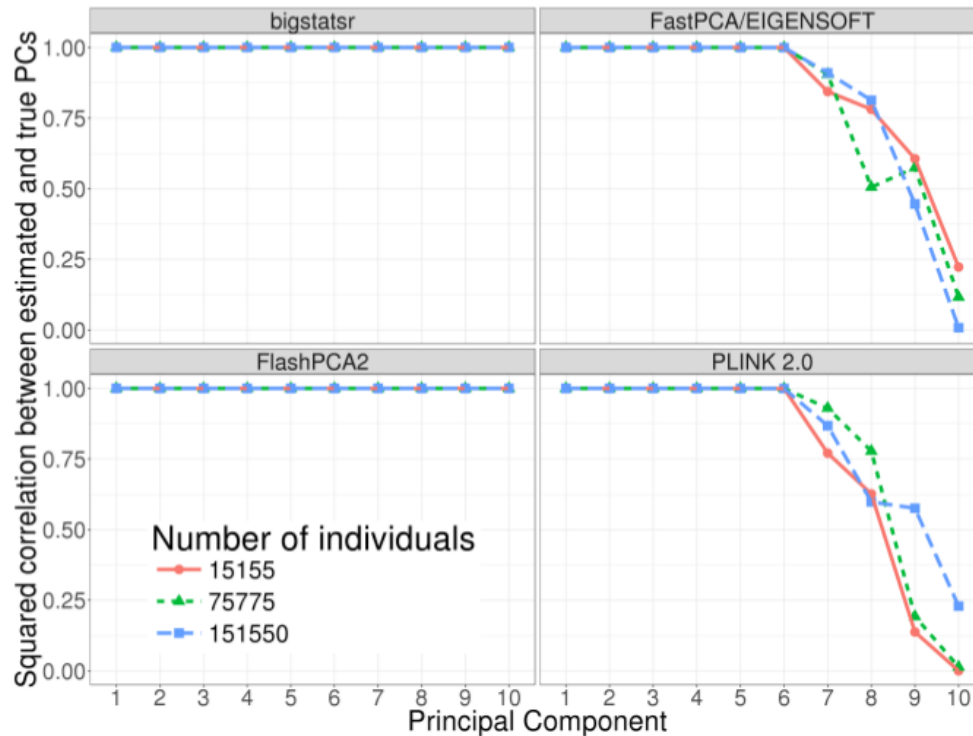


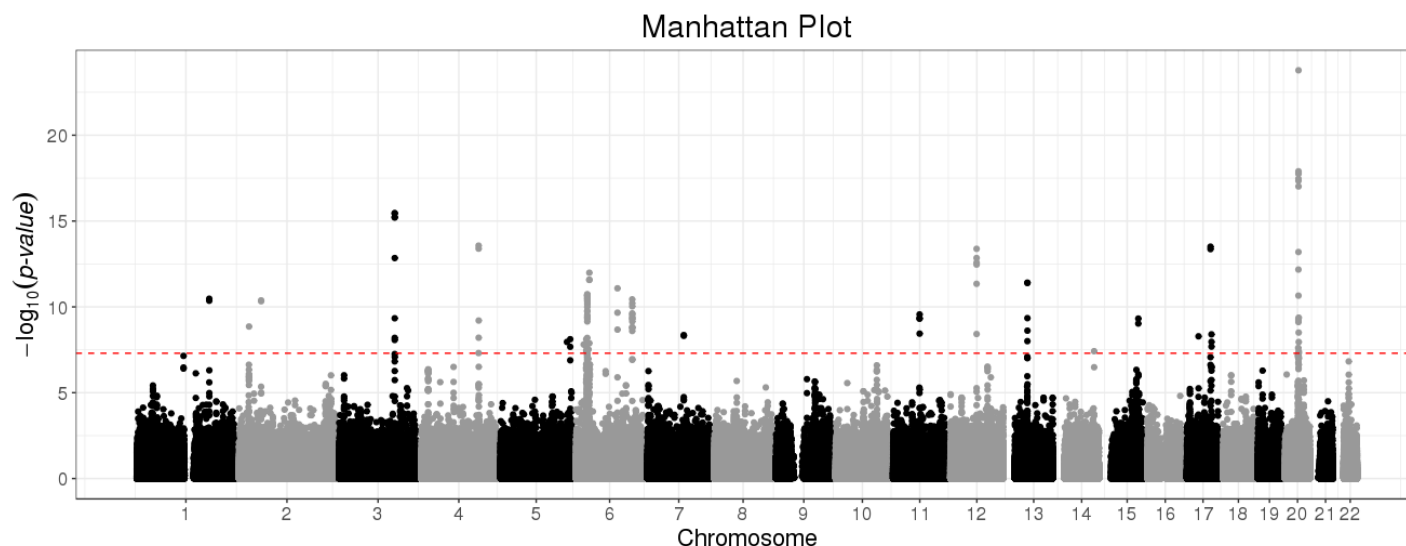
Fig. 5. Precision comparisons between randomized partial Singular Value Decomposition available in FlashPCA2, FastPCA (fast mode of SmartPCA/EIGENSOFT), PLINK 2.0 (approx mode) and package bigstatsr. It shows the squared correlation between approximated PCs and “true” PCs (produced by the exact mode of PLINK 2.0) of the Celiac dataset (whose individuals have been repeated 1, 5 and 10 times).

How to predict disease status
based on genotypes?

Standard PRS - part 1: estimating effects

Genome-wide association studies (GWAS)

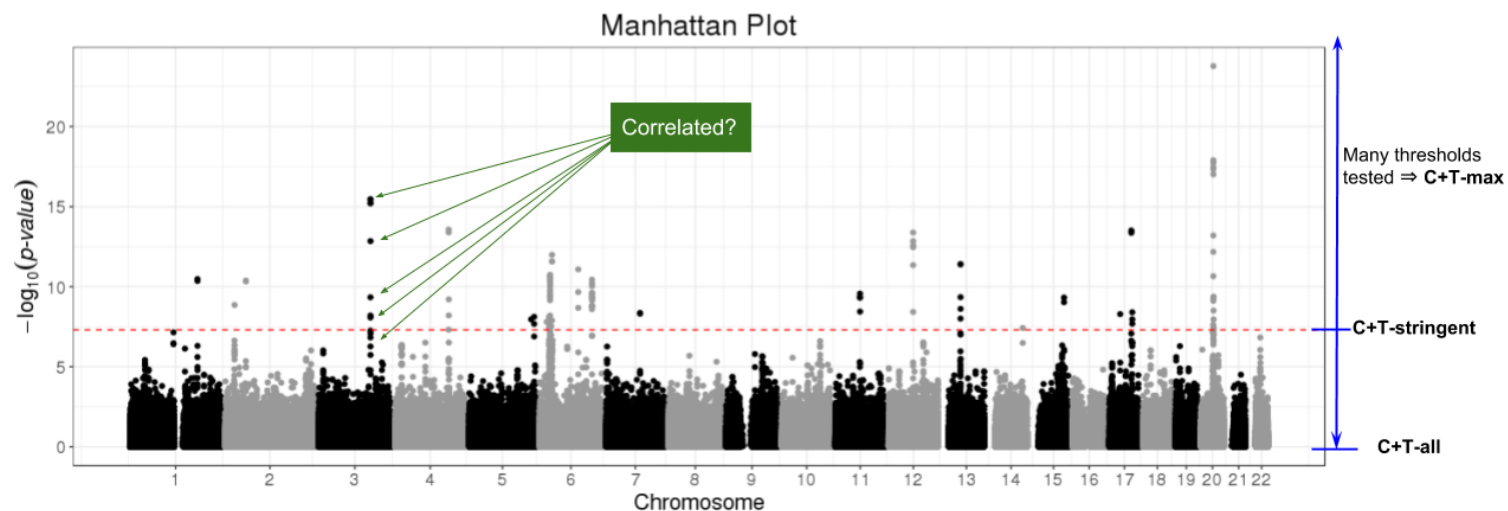
In a GWAS, each single-nucleotide polymorphism (SNP) is tested **independently**, resulting in one **effect size** $\hat{\beta}$ and one **p-value** p for each SNP.



Easy combining: $PRS_i = \sum_j \hat{\beta}_j \cdot G_{i,j}$

Standard PRS - part 2: restricting predictors

Clumping + Thresholding ("C+T" or just "PRS")



$$PRS_i = \sum_{\substack{j \in S_{\text{clumping}} \\ p_j < p_T}} \hat{\beta}_j \cdot G_{i,j}$$

A more optimal approach to computing PRS?

In C+T: weights learned independently and heuristics for correlation and regularization.

Statistical learning

- joint models of all SNPs at once
- use regularization to account for correlated and null effects
- already proved useful in the literature (Abraham et al. 2013; Okser et al. 2014; Spiliopoulou et al. 2015)

Our contribution

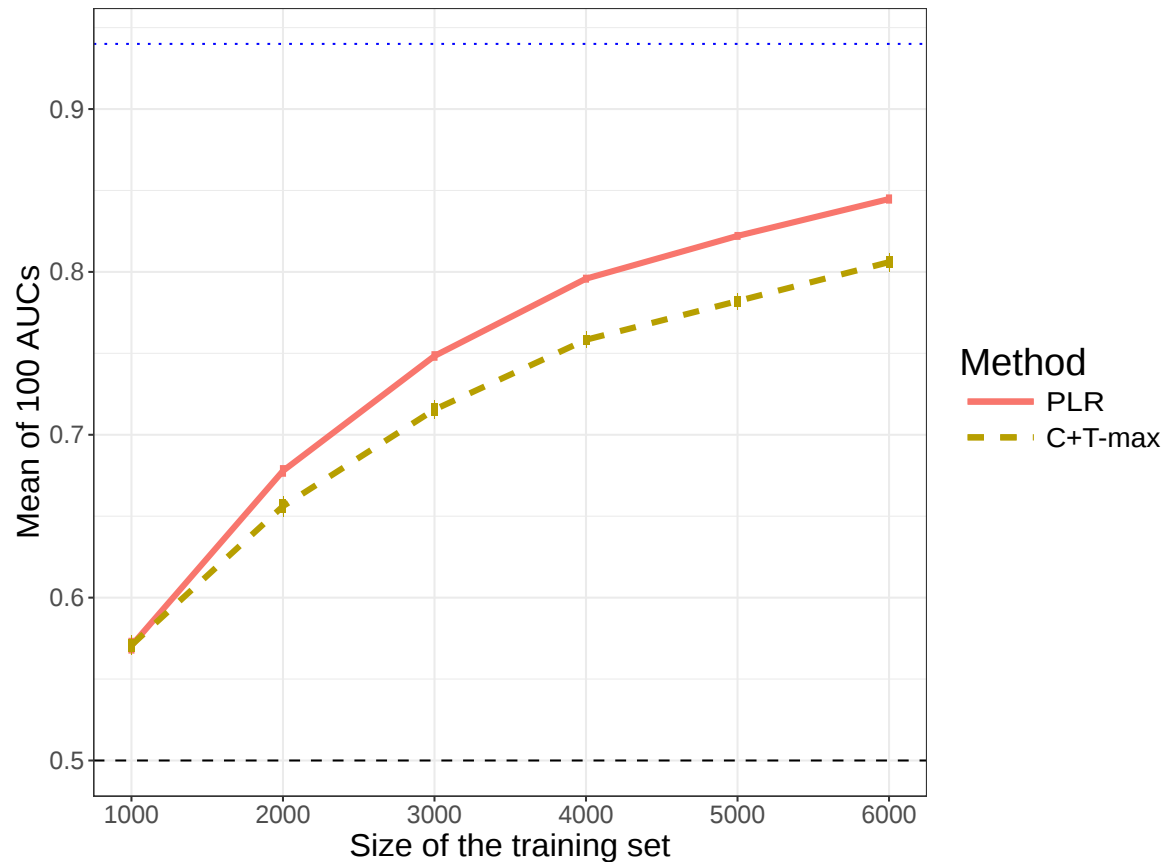
- a memory- and computation-efficient implementation to be used for biobank-scale data
- an automatic choice of the regularization hyper-parameter
- a comprehensive comparison for different disease architectures

Penalized Logistic Regression (PLR)

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

-
- $p_i = 1 / (1 + \exp(-(\beta_0 + x_i^T \beta)))$
 - 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 $0 \leq \alpha \leq 1$.

Prediction with PLR is improving faster



"Efficient implementation of penalized regression for genetic risk prediction" -
- under revision

Real data

Celiac disease

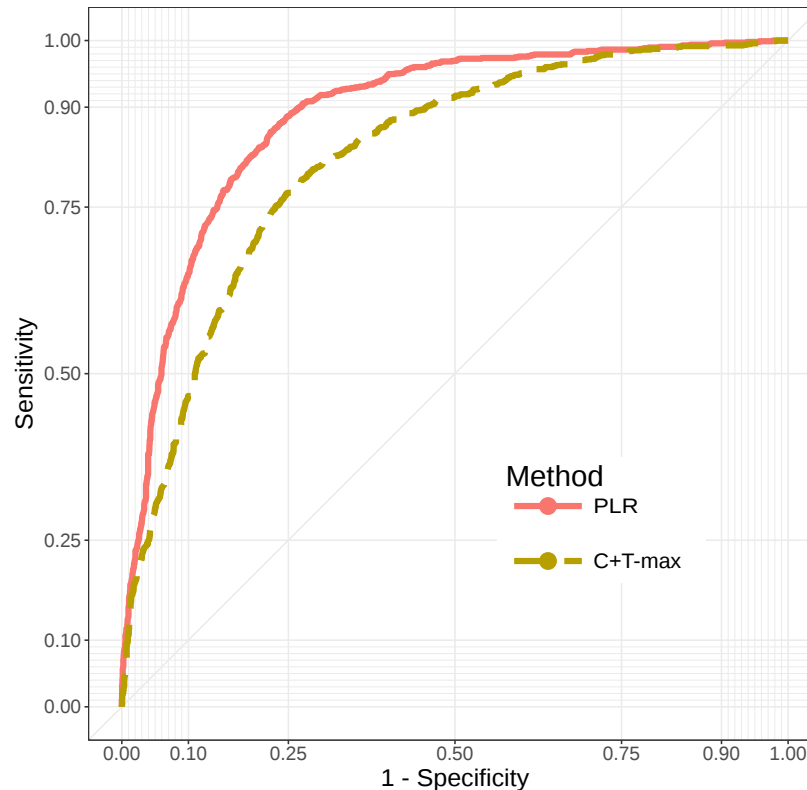
- intolerance to gluten
- only treatment: gluten-free diet
- heritability: 57-87% (Nisticò et al. 2006)
- prevalence: 1-6%

Case-control study for the celiac disease (WTCCC, Dubois et al. 2010)

- ~15,000 individuals
- ~280,000 SNPs
- ~30% cases

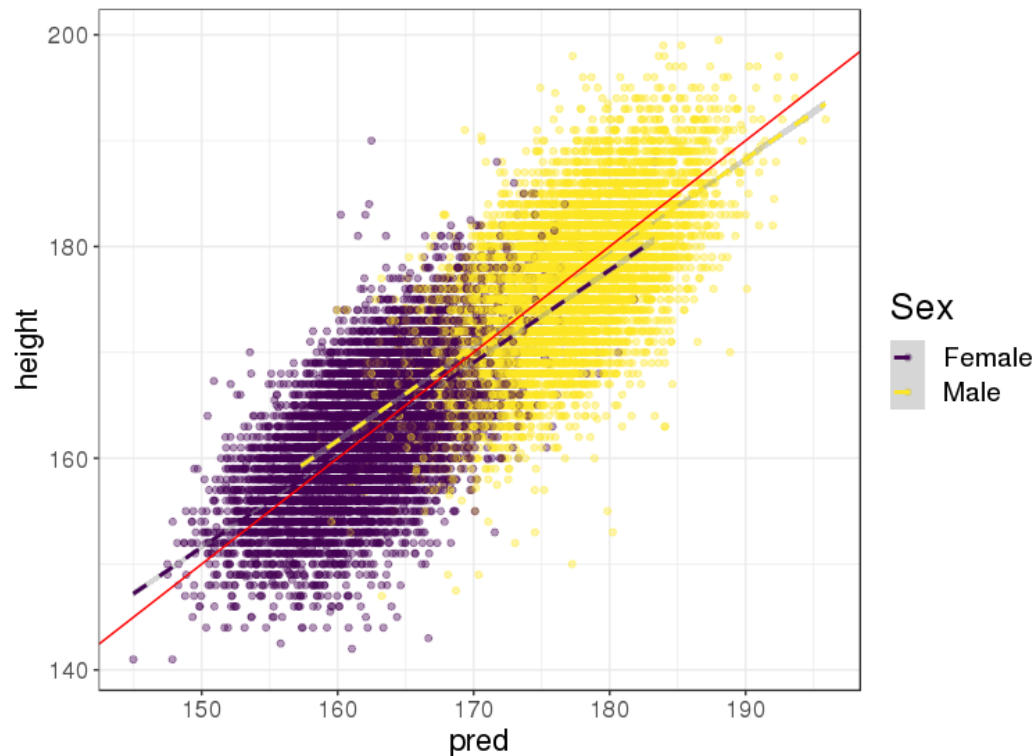
Results: real Celiac phenotypes

Method	AUC	pAUC	# predictors	Execution time (s)
C+T-max	0.825 (0.000664)	0.0289 (0.000187)	8360 (744)	130 (0.143)
PLR	0.887 (0.00061)	0.0411 (0.000224)	1570 (46.4)	190 (1.21)
PLR3	0.891 (0.000628)	0.0426 (0.000219)	2260 (56.1)	296 (2.03)



LASSO for predicting height

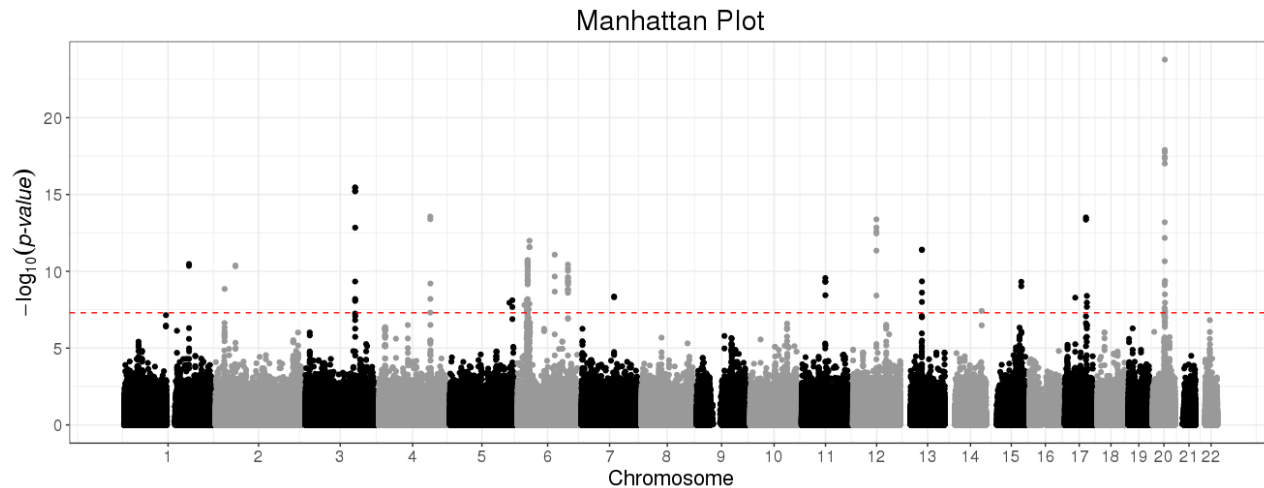
- 350K individuals x 656K SNPs in less than one day
- Within each sex category, 65.5% of correlation between predicted and true height (56% with C+T-max)



Using summary statistics

Standard PRS: C+T

In a GWAS, each SNP is tested independently, resulting in one **effect size** $\hat{\beta}$ and one **p-value** p for each SNP (**summary statistics**).



$$PRS_i = \sum_{\substack{j \in S_{\text{clumping}} \\ p_j < p_T}} \hat{\beta}_j \cdot G_{i,j}$$

Predictive methods based on summary statistics

When you have only summary statistics (and a small reference panel), you can use:

- C+T
- LDpred (*Vilhjálmsen, Bjarni J., et al. "Modeling linkage disequilibrium increases accuracy of polygenic risk scores." The American Journal of Human Genetics 97.4 (2015): 576-592.*)
- lassosum (*Mak, Timothy Shin Heng, et al. "Polygenic scores via penalized regression on summary statistics." Genetic epidemiology 41.6 (2017): 469-480.*)
- NPS (*Chun, Sung, et al. "Non-parametric polygenic risk prediction using partitioned GWAS summary statistics." bioRxiv (2018): 370064.*)

The idea of LDpred, lassosum and NPS is to use a reference panel to **account for correlation** between SNPs, instead of pruning (removing) SNPs. Lassosum also adds some sparsity.

Could those models be improved?

- take into account quality of imputation?

1	?	?	?	1	?	1	?	0	2	2	?	?	2	?	0
0	?	?	?	2	?	2	?	0	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	0
1	?	?	?	2	?	1	?	1	2	2	?	?	2	?	0
2	?	?	?	2	?	2	?	1	2	1	?	?	2	?	0
1	?	?	?	1	?	1	?	1	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	1
2	?	?	?	1	?	1	?	1	2	1	?	?	2	?	1
1	?	?	?	0	?	0	?	2	2	2	?	?	2	?	0

- support PLINK bed files only; what about BGEN files such as for the UK Biobank?
- scalable to which extent?
- combine with individual-level datasets? (of possibly different populations)

One idea: stacking

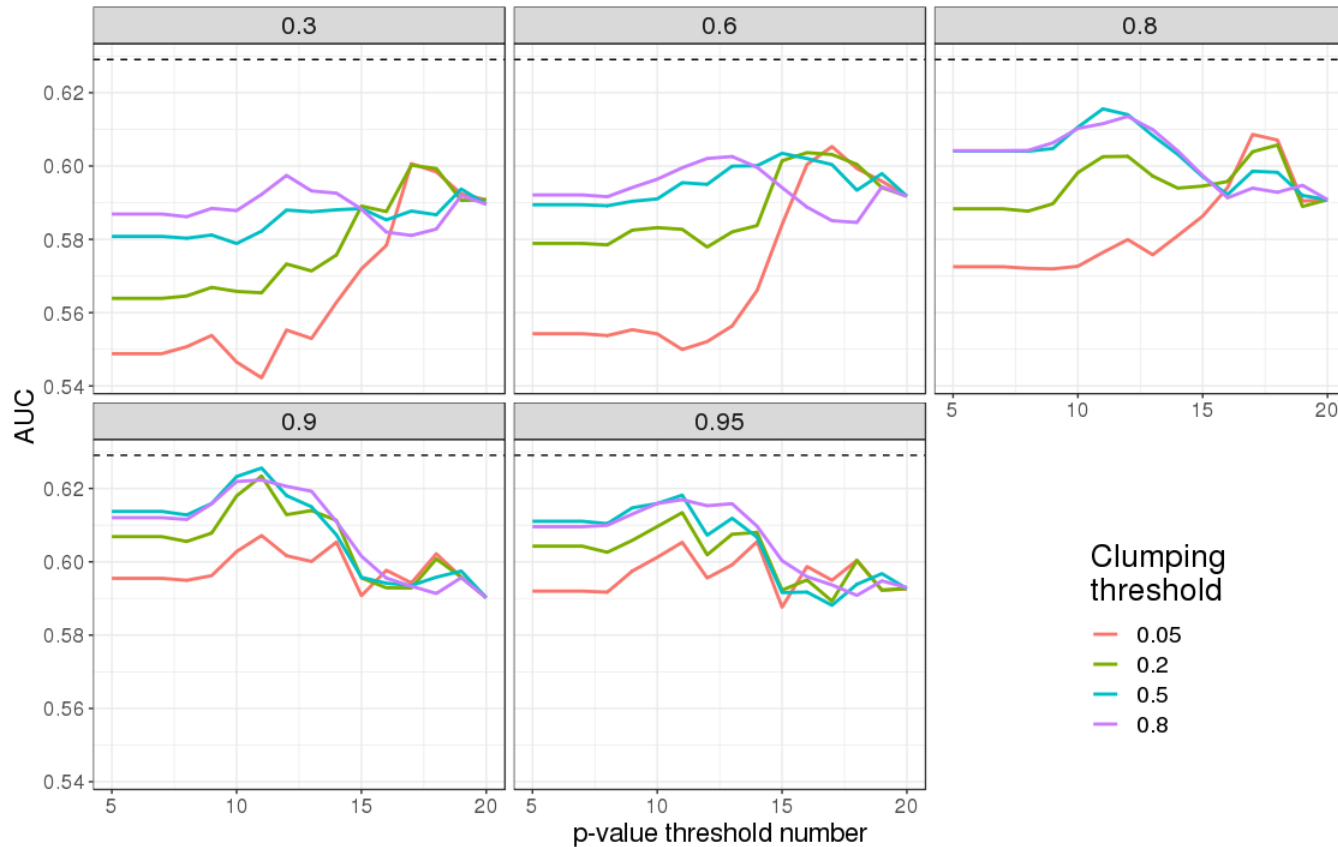
For a grid of combinations of 3 parameters:

- threshold of clumping ($r_{clumping}^2$)
- p-value threshold (p_T)
- threshold of imputation quality score ($r_{imputation}^2$)

Get a C+T prediction for each combination of parameters.

Then use those predictions as variables in a (penalized) regression (stacking of models).


Stacking: preliminary results



Prediction for breast cancer.
Facets are representing different imputation quality thresholds.

Conclusion

My thesis work

1. Developing two  packages for the analysis of large-scale genomic data.

(<https://doi.org/10.1093/bioinformatics/bty185>)

Package bigstatsr can be used for any data encoded as matrices.

2. Including an implementation of penalized regression for very large individual-level datasets + assess the potential gain in prediction over the simple standard model (C+T).

(<https://doi.org/10.1101/403337>)

3. Including summary statistics from large GWAS to improve prediction.

(TODO)

Make sure to grab an hex sticker



Thanks!

Presentation available at

<https://privefl.github.io/thesis-docs/aarhus.html>



privefl



privefl



F. Privé

Slides created via R package **xaringan**.