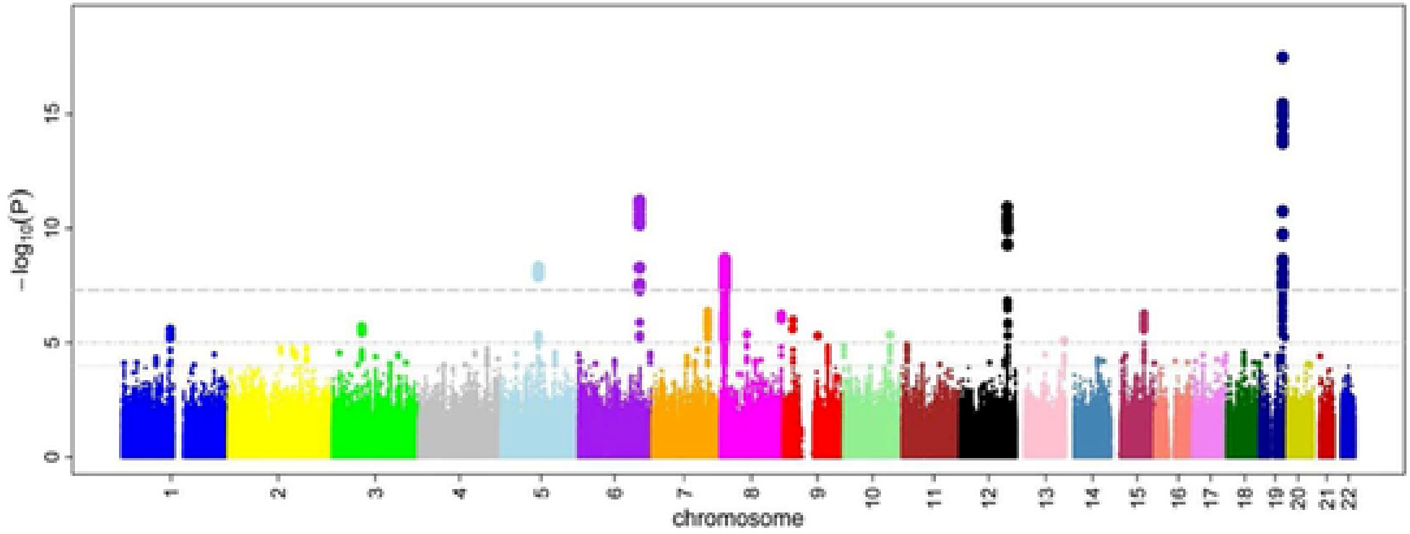
Polygenic Risk Scores for Predictive Medecine and Epidemiology

Florian Privé, Hugues Aschard and Michael Blum

IAB - June 15, 2018

Introduction

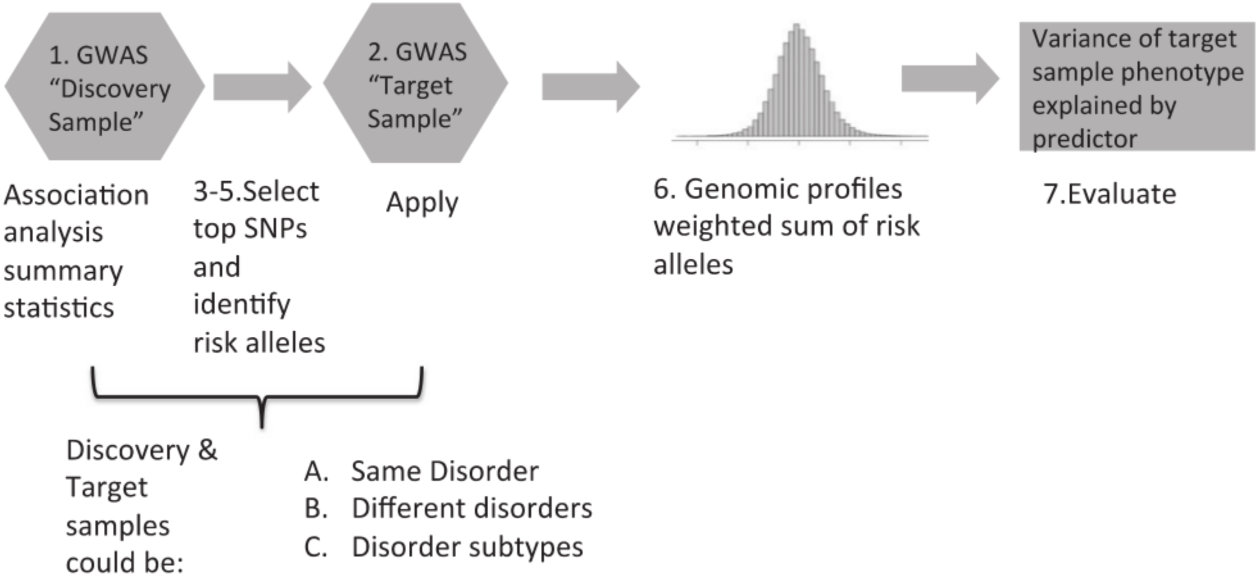
From genome-wide association studies (GWAS) to polygenic risk scores (PRS)



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

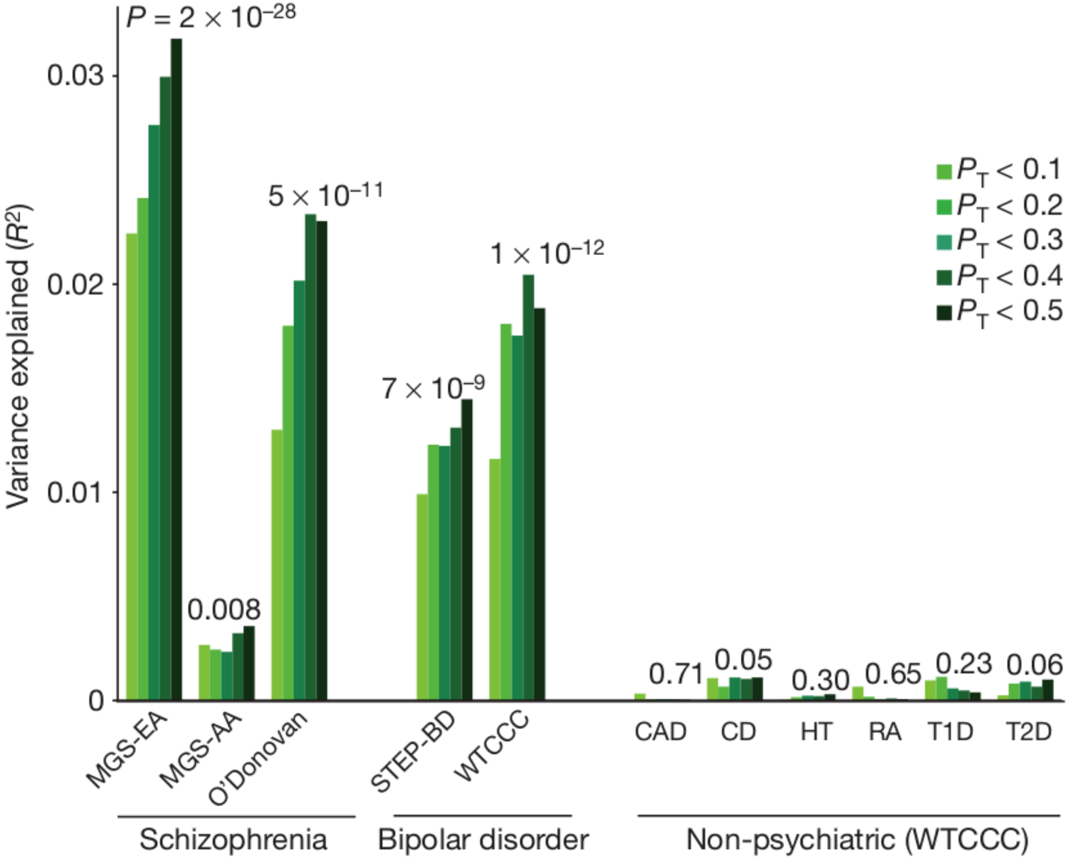
Polygenic Risk Scores (PRS) for epidemiology

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


1. GWAS “Discovery Sample”
2. GWAS “Target Sample”
Variance of target sample phenotype explained by predictor
Association analysis summary statistics
3-5.Select top SNPs and identify risk alleles
Apply
6. Genomic profiles weighted sum of risk alleles
7.Evaluate
Discovery & Target samples could be:
A. Same Disorder
B. Different disorders
C. Disorder subtypes

Source: 10.1111/jcpp.12295

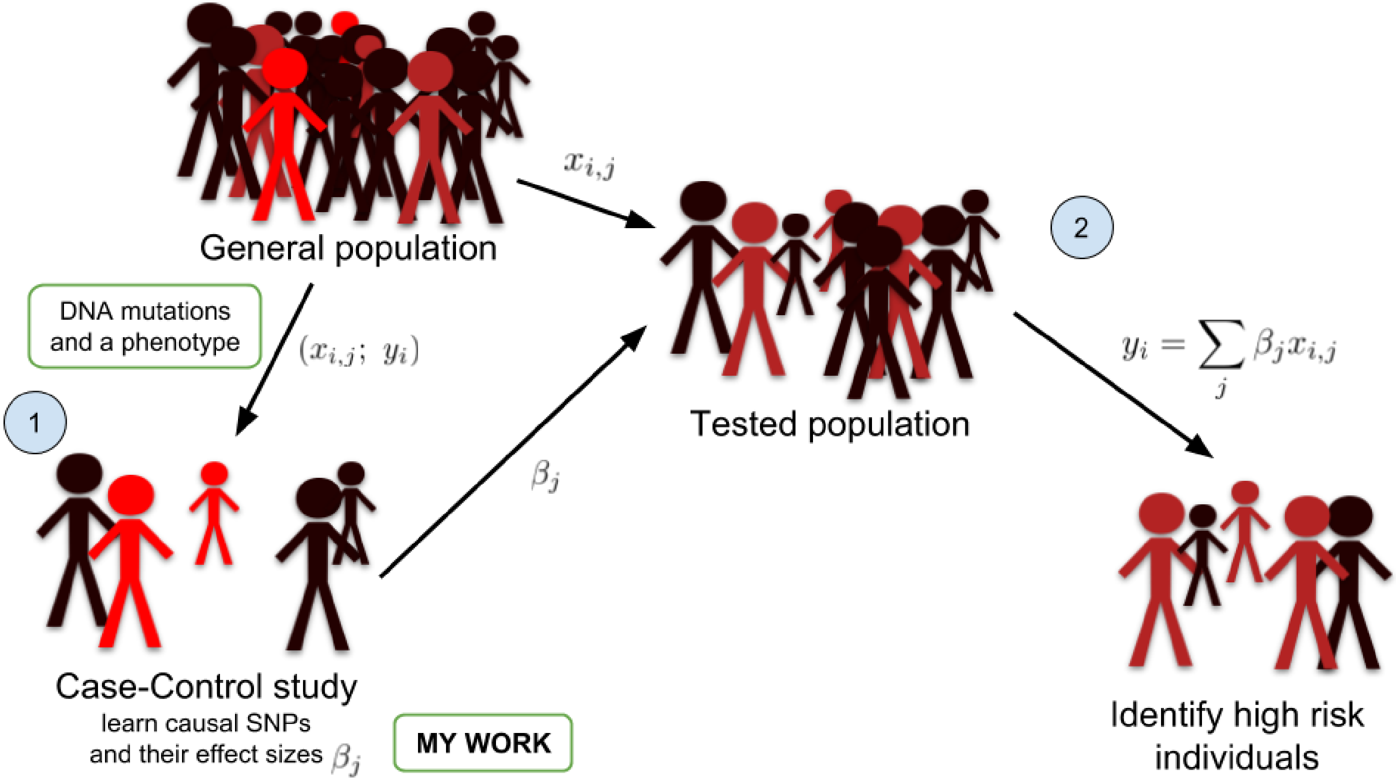
Polygenic Risk Scores (PRS) for epidemiology



Source: 10.1038/nature08185

Polygenic Risk Scores (PRS) for predictive medicine

Another application: to identify high risk individuals


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

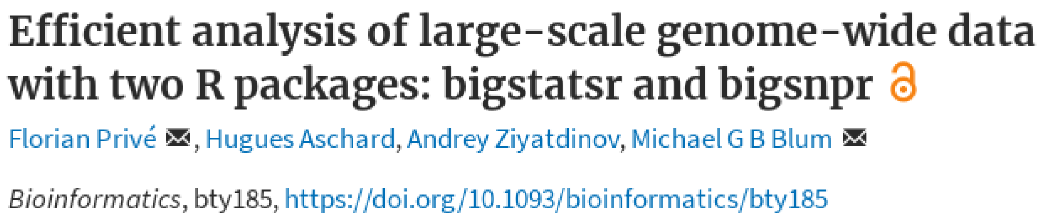
- previously: 15K x 280K, celiac disease (~30GB)
- currently: 500K x 500K, UK Biobank (~2TB)

But I still want to use R..

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

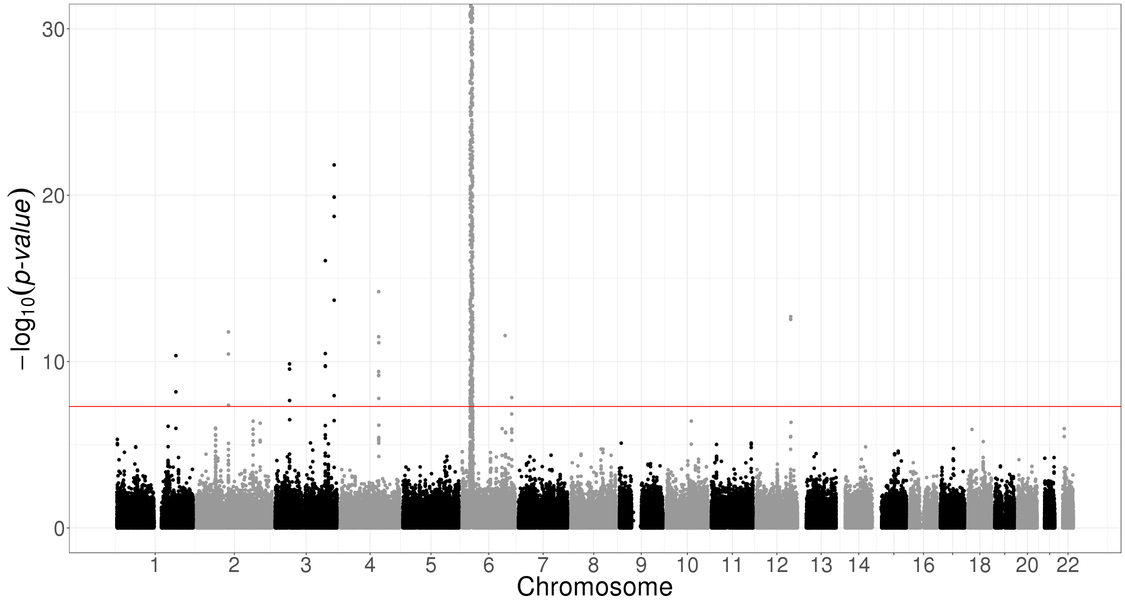
Package {bigstatsr} provides fast PCA, association and predictive models, etc.

How to predict disease status based on genotypes?

Standard PRS - part 1: estimating effects

Genome-wide association studies (GWAS)

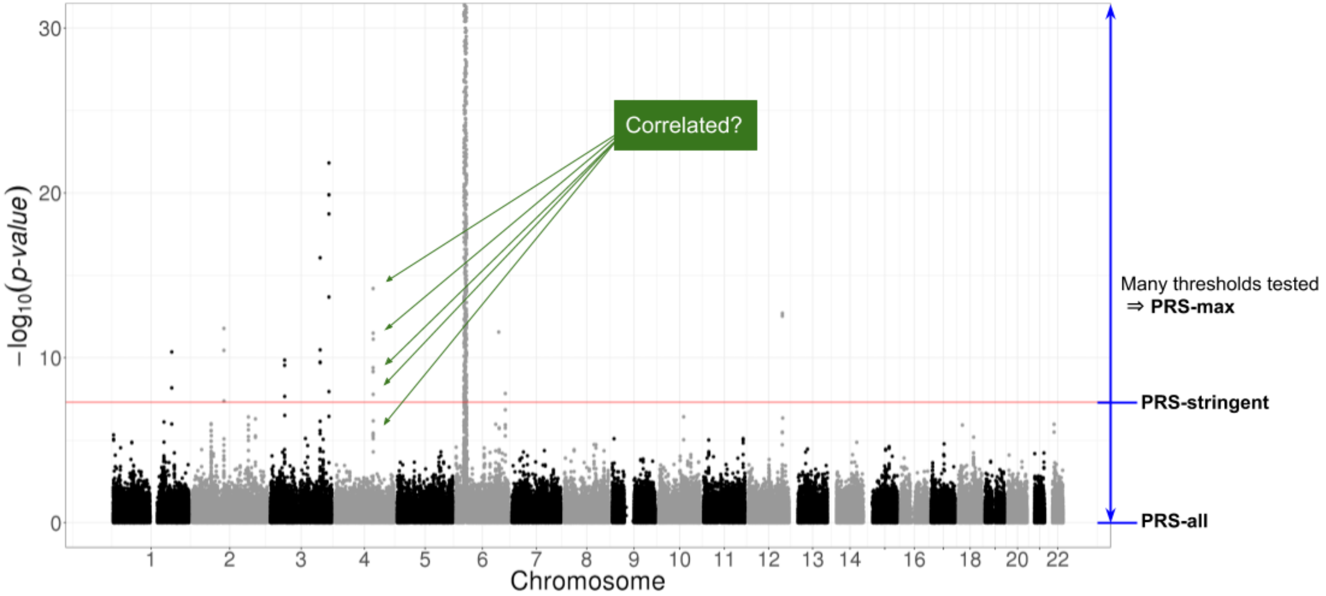
In a GWAS, each single-nucleotide polymorphism (SNP) is tested **independently**, resulting in one **effect size** $\hat{\beta}$ and one **p-value** p for each SNP.



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

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 litterature (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

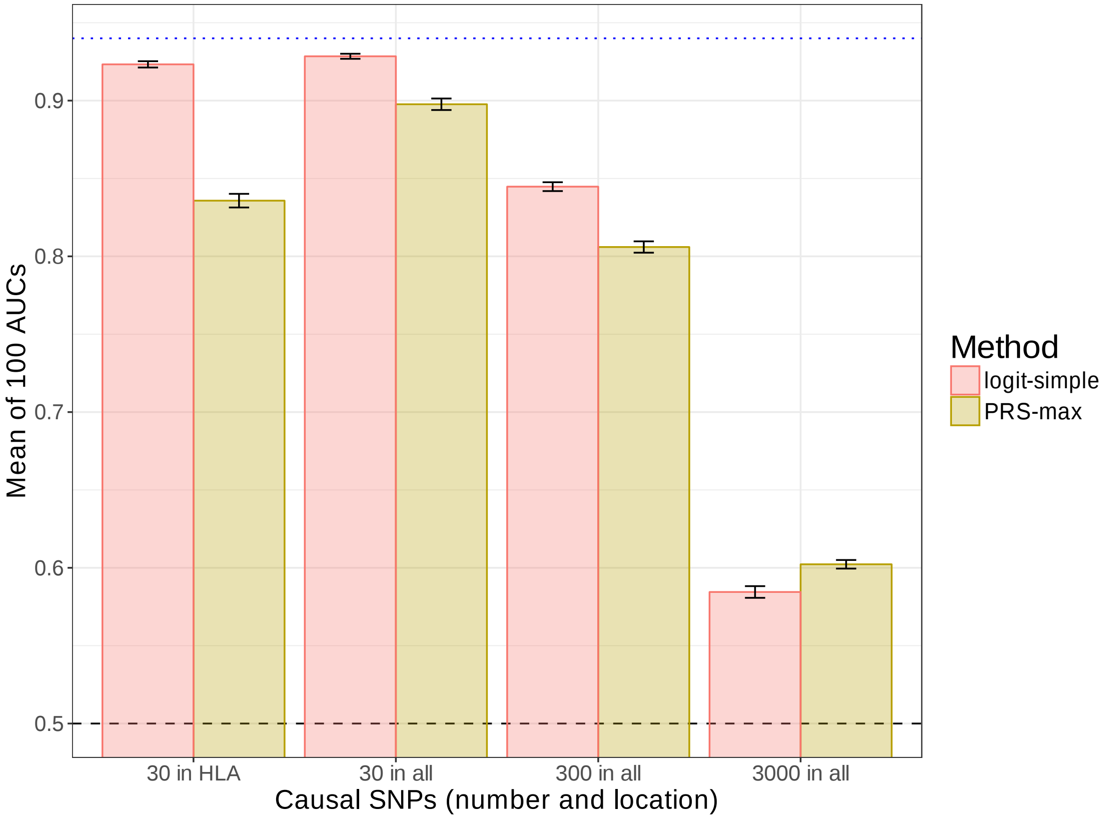
Comparison of methods for computing PRS

Penalized Logistic Regression

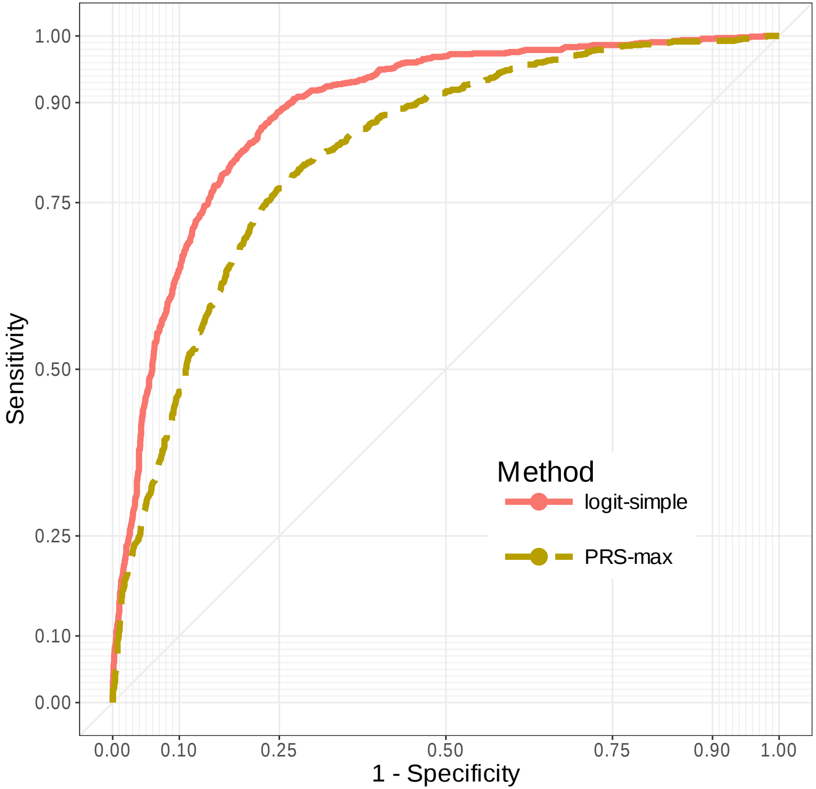
$$\underset{\beta_0,\beta}{\operatorname{argmin}}(\lambda,\alpha)\left\{\underbrace{\frac{1}{n}\sum_{i=1}^{n}\log\left(1+e^{-y_i(\beta_0+x_i^T\beta)}\right)}_{\text{Loss function}}+\underbrace{\lambda\left((1-\alpha)\frac{1}{2}\|\beta\|_2^2+\alpha\|\beta\|_1\right)}_{\text{Penalization}}\right\}$$

- 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 $0 \le \alpha \le 1$.

Higher predictive performance with penalized logistic regression

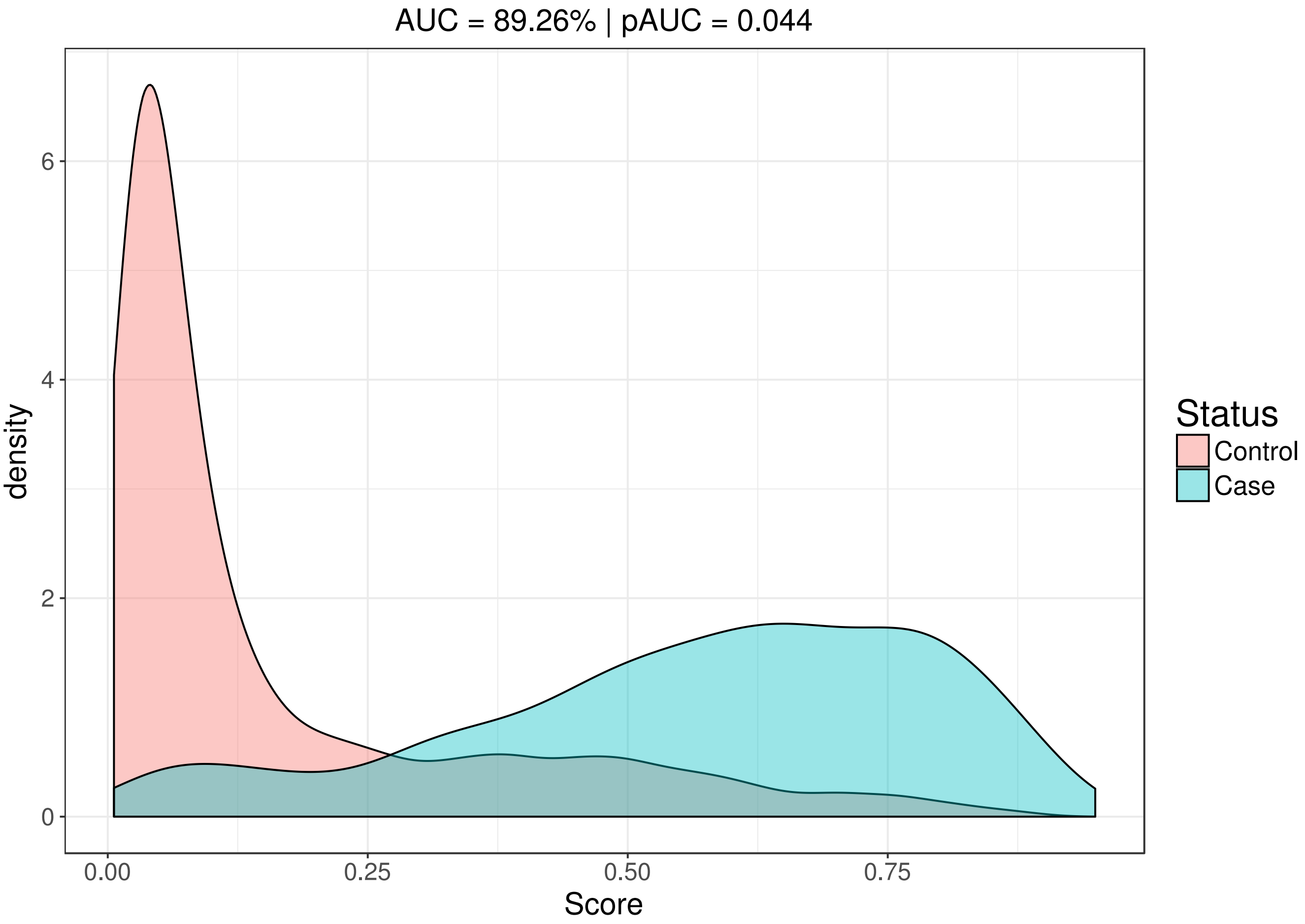


Results: real Celiac phenotypes



Method	AUC	pAUC	# predictors	Execution time (s)
PRS-max	0.824 (0.000704)	0.0286 (0.00016)	9850 (781)	148 (0.414)
logit-simple	0.888 (0.000468)	0.0414 (0.000164)	3220 (62)	83.8 (1.27)

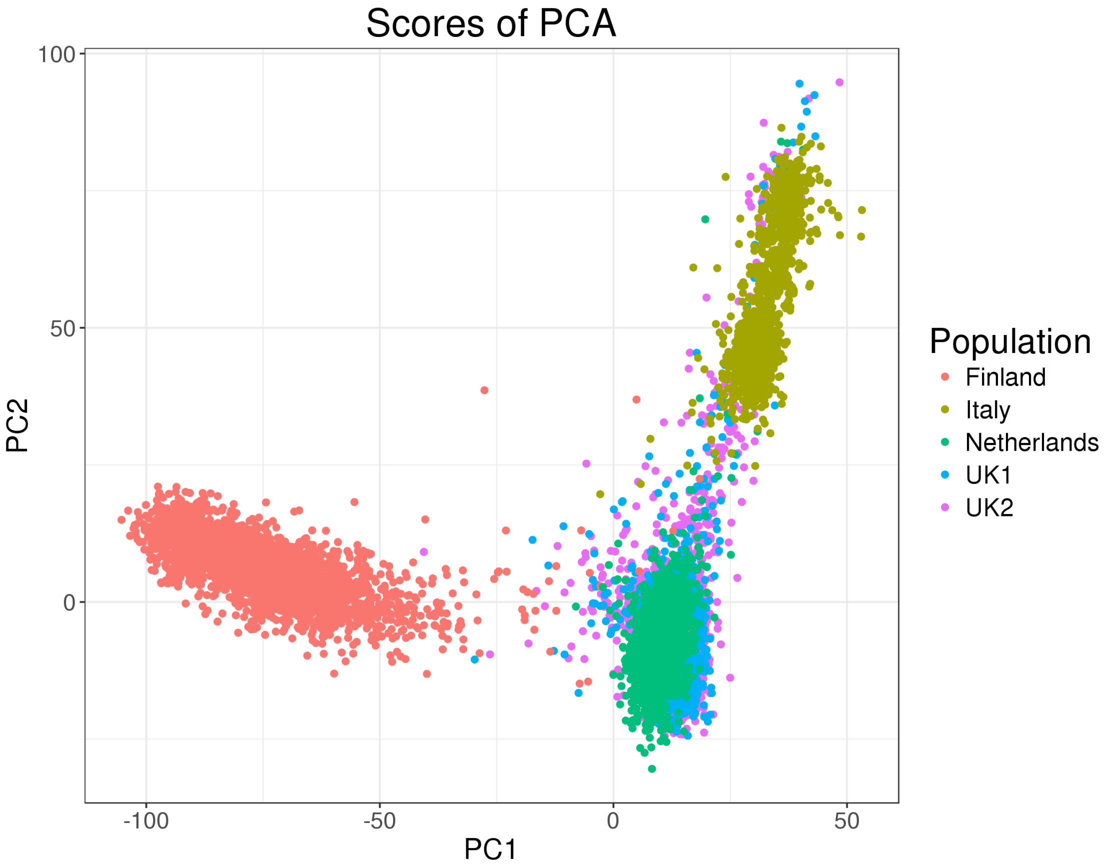
Results: real Celiac phenotypes



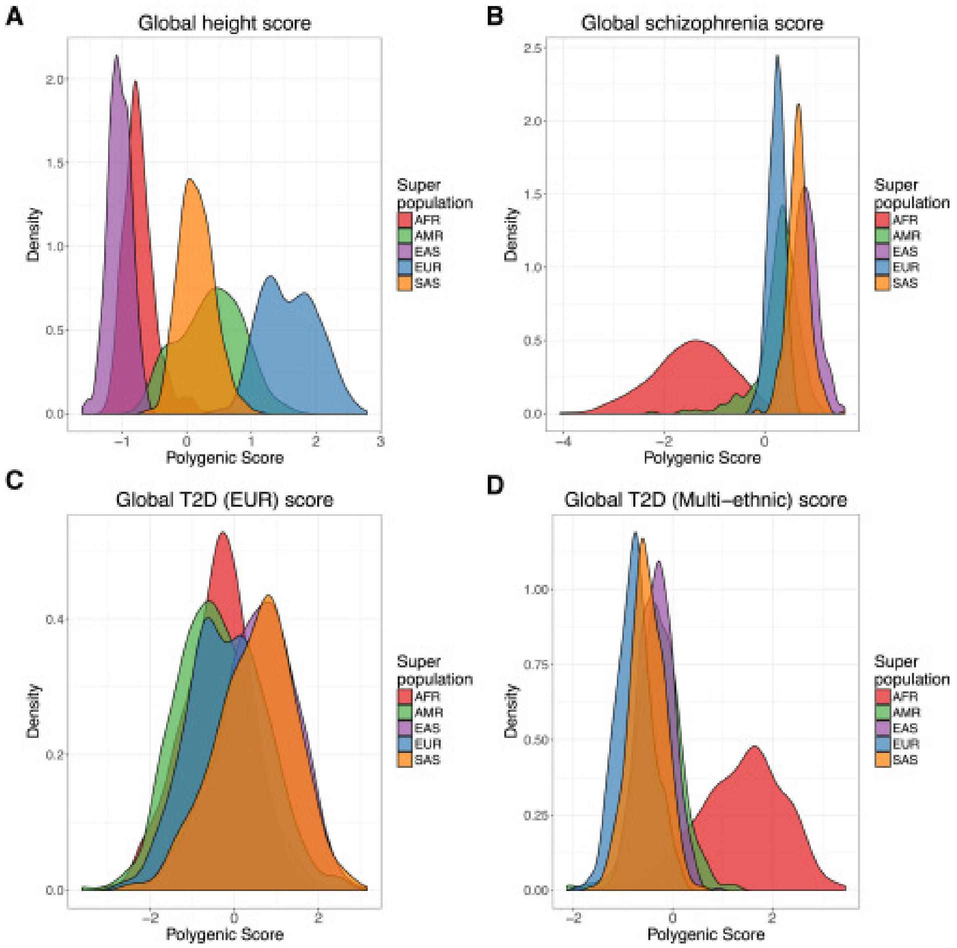
How to combine the information of multiple studies?

(possibly of different populations)

Genetics are different between populations

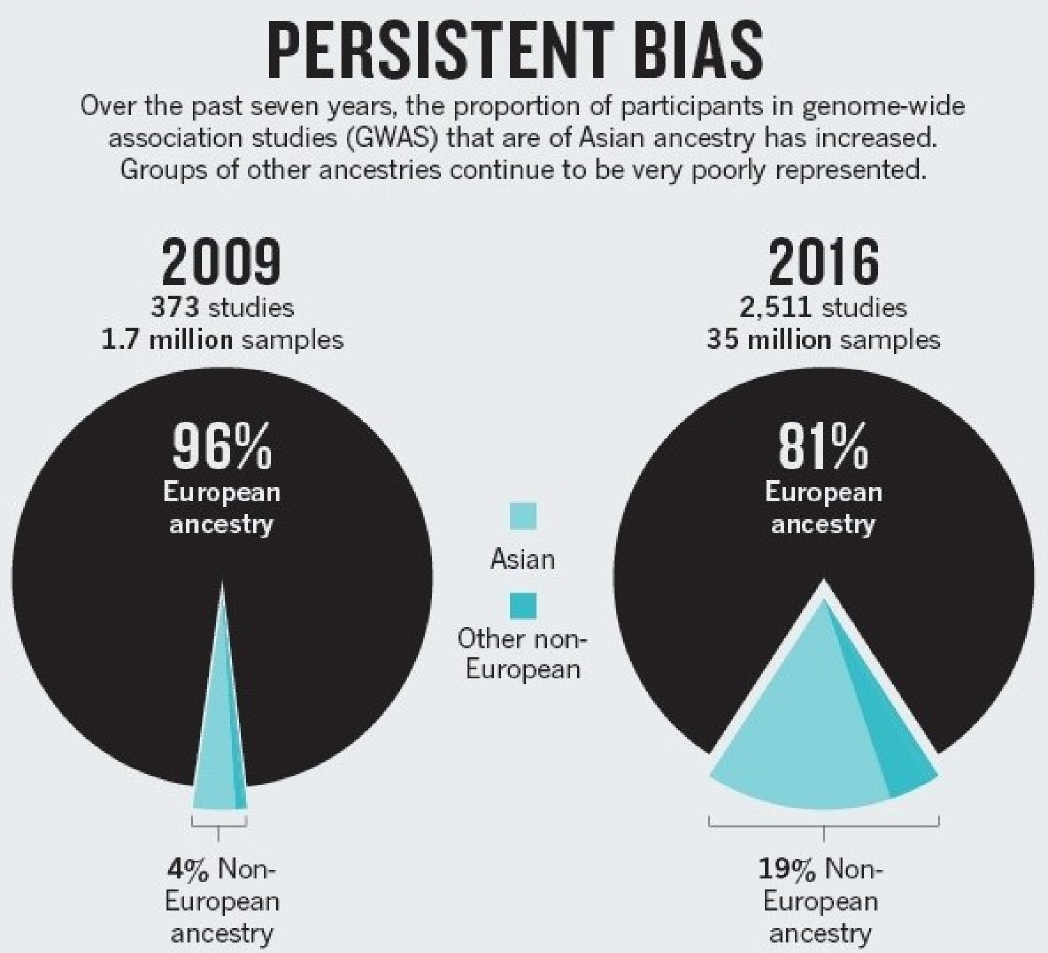


which makes predictions fail on external populations



Source: 10.1016/j.ajhg.2017.03.004

Genomics is failing on diversity



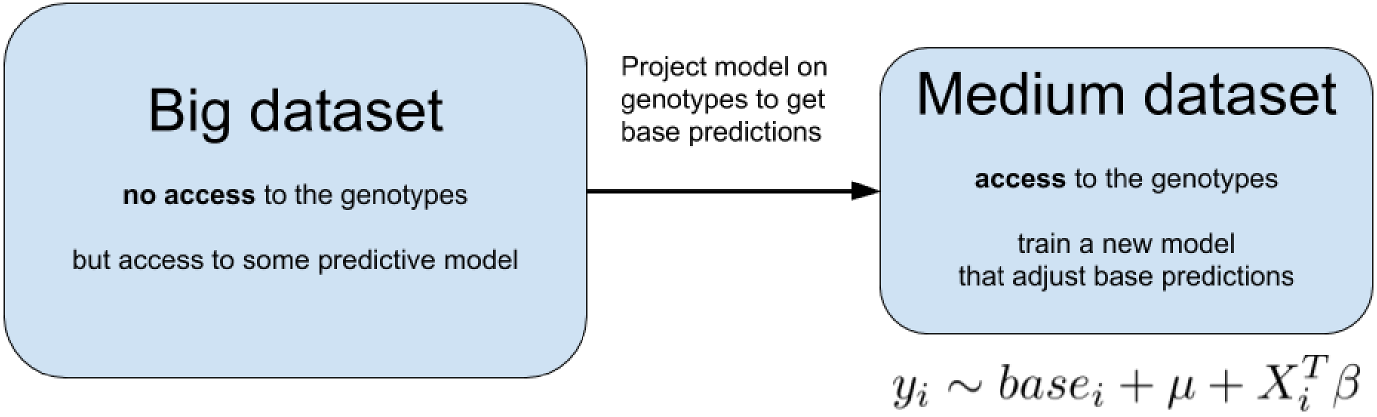
PERSISTENT BIAS

Over the past seven years, the proportion of participants in genome-wide association studies (GWAS) that are of Asian ancestry has increased. Groups of other ancestries continue to be very poorly represented.

Source: 10.1038/538161a

What can we do about it?

We can use information from other studies (possibly in other populations)



Big dataset

no access to the genotypes

but access to some predictive model

Project model on genotypes to get base predictions

Medium dataset

access to the genotypes

train a new model that adjust base predictions

$$y_i \sim base_i + \mu + X_i^T \beta$$

Will this improve prediction?

Can we learn more than just prediction?

1. Imagine you learn a model on a large european population
2. You project this predictive model on an african population in order to get a base predictor
3. You learn another model on this african population to adjust from this base predictor

$$y_i \sim base_i + \mu + X_i^T \beta$$

What can we tell about the SNPs that are used in the new model?

Thanks!

Presentation available at

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

privefl privefl F. Privé

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