Polygenic Risk Scores based on Statistical Learning

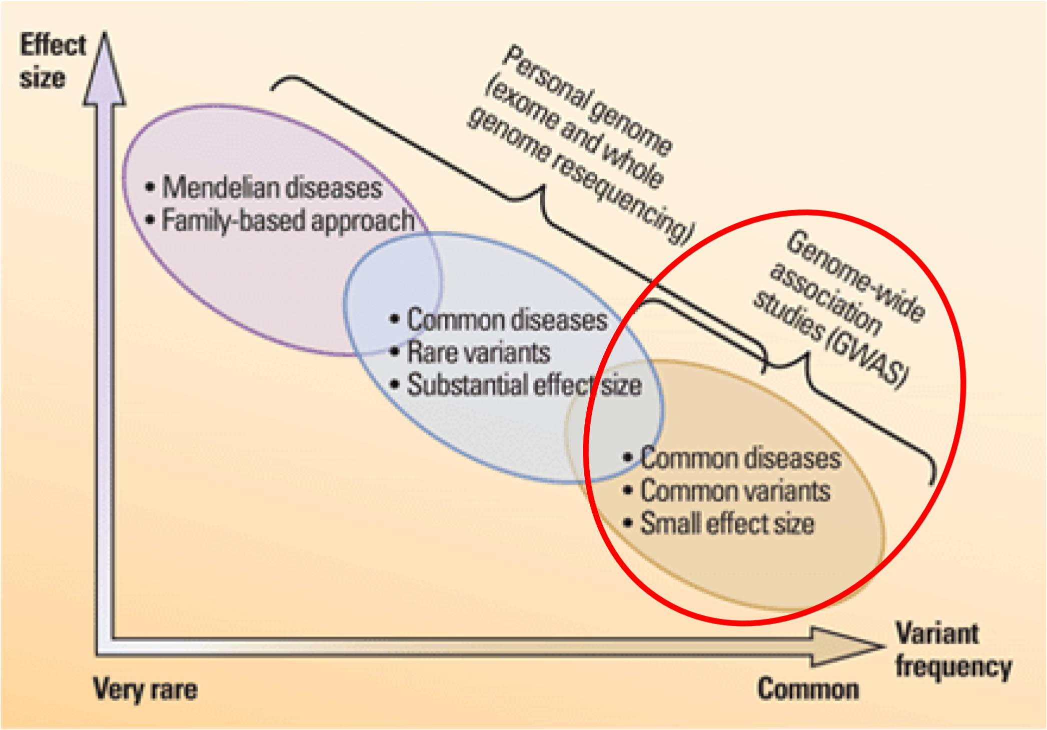
Florian PRIVÉ

thesis supervised by Michael BLUM (Univ. Grenoble Alpes)
and co-supervised by Hugues ASCHARD (Institut Pasteur)

Introduction & Motivation

Data, application and research interest

Disease architecture

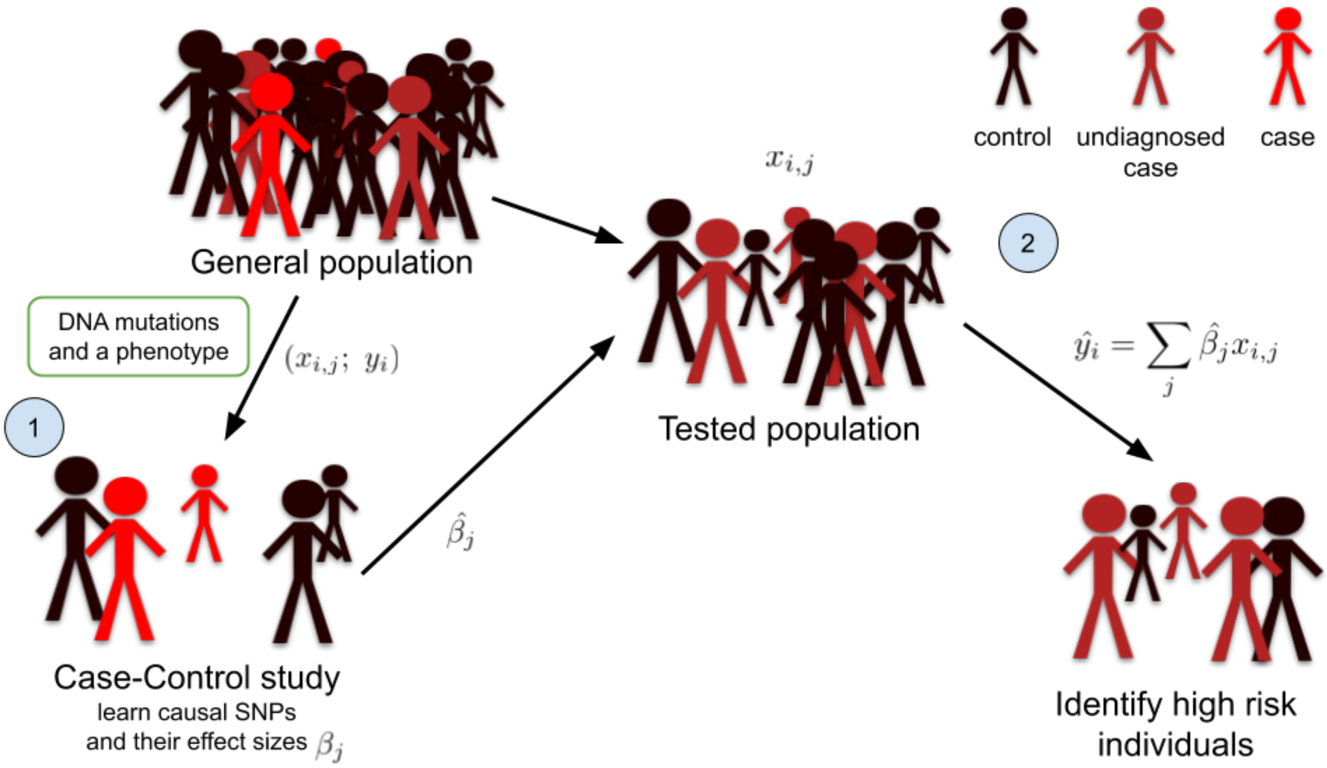


Source: 10.1126/science.338.6110.1016

Polygenic Risk Scores (PRS)

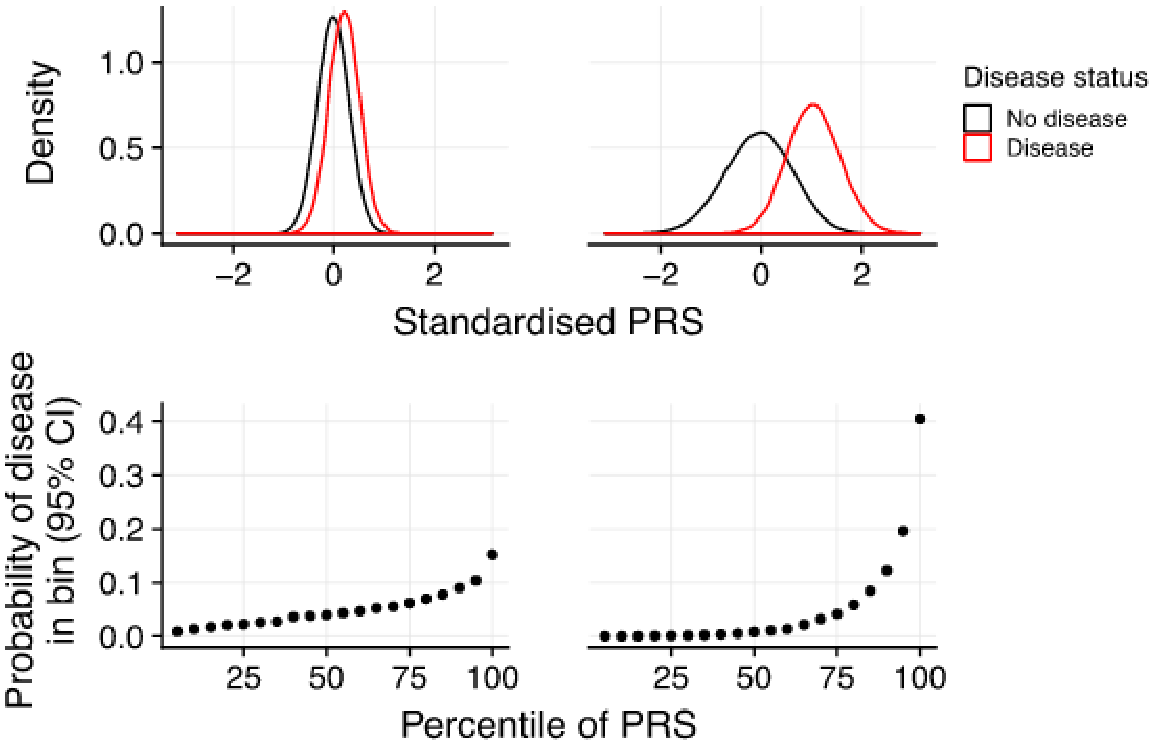
A simple model: $y_i = \sum_j \beta_j x_{i,j} + \epsilon$

y_i: phenotypes, $x_{i,j}$: genotypes, β_j: effect sizes, ϵ: environmental effect.



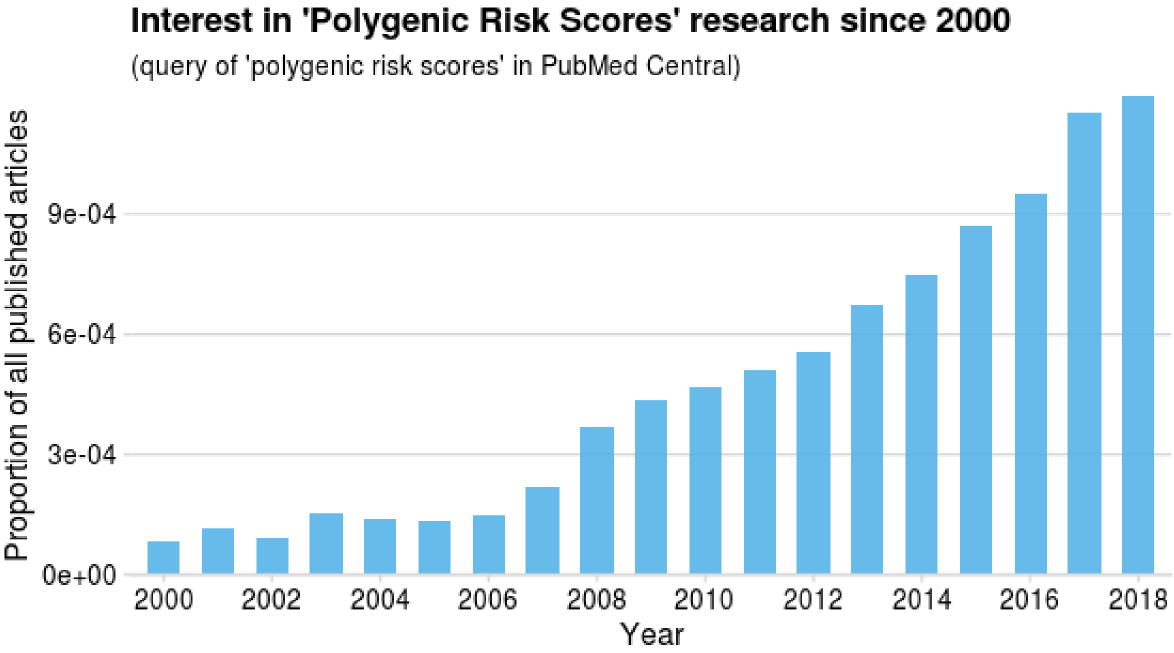
control
undiagnosed case
case
General population
DNA mutations and a phenotype
$(x_{i,j};\ y_i)$
1
Case-Control study
learn causal SNPs and their effect sizes β_j
$\hat{\beta}_j$
$x_{i,j}$
Tested population
2
$\hat{y}_i = \sum_j \hat{\beta}_j x_{i,j}$
Identify high risk individuals

Identify high-risk individuals



Source: 10.1093/hmg/ddz187

Interest in Polygenic Risk Scores (PRS)



However, current predictions fall short from clinical utility.

We need larger sample sizes and more optimal predictions.

Data: very large genotype matrices

Matrices of genetic variants (DNA mutations)

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

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

Data I analyzed:

- celiac disease: 15K x 280K (~30GB)
- UK Biobank: 500K x 800K (~3TB)

But I still want to use R..

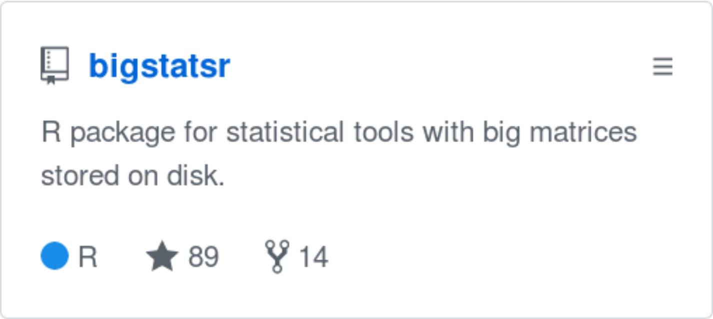
How to analyze large genomic data?

Privé, F., Aschard, H., Ziyatdinov, A., & Blum, M. G.B. (2018). *Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr.* Bioinformatics, 34(16), 2781-2787.

Our two R packages: bigstatsr and bigsnpr

Smooth and fast data analysis with big matrices stored on disk

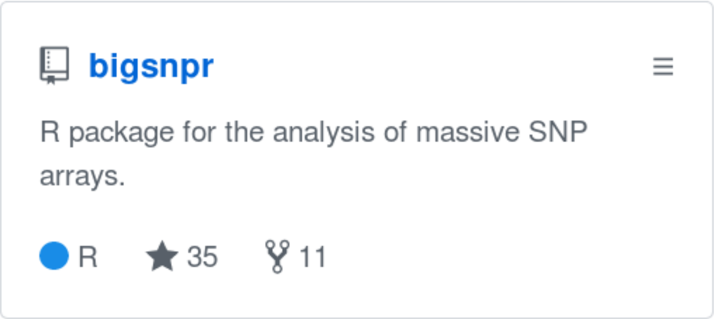
- {bigstatsr} for many types of matrix, to be used by any field of research
- {bigsnpr} for functions that are specific to the analysis of genetic data



bigstatsr

R package for statistical tools with big matrices stored on disk.

● R ★ 89 ⑂ 14



bigsnpr

R package for the analysis of massive SNP arrays.

● R ★ 35 ⑂ 11

How to predict disease status based on genotypes?

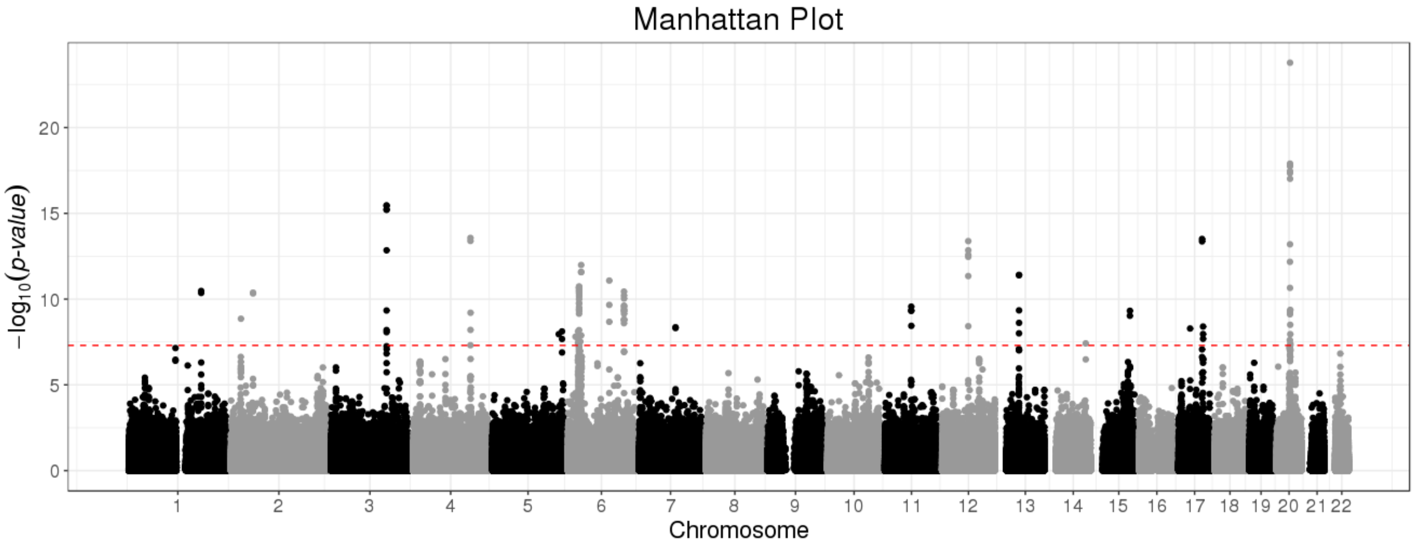
Prediction using individual-level data

Privé, F., Aschard, H., & Blum, M. G.B. (2019).
Efficient implementation of penalized regression for genetic risk prediction. Genetics, 212(1), 65-74.

Standard PRS - part 1: estimating effects

Genome-wide association studies (GWAS)

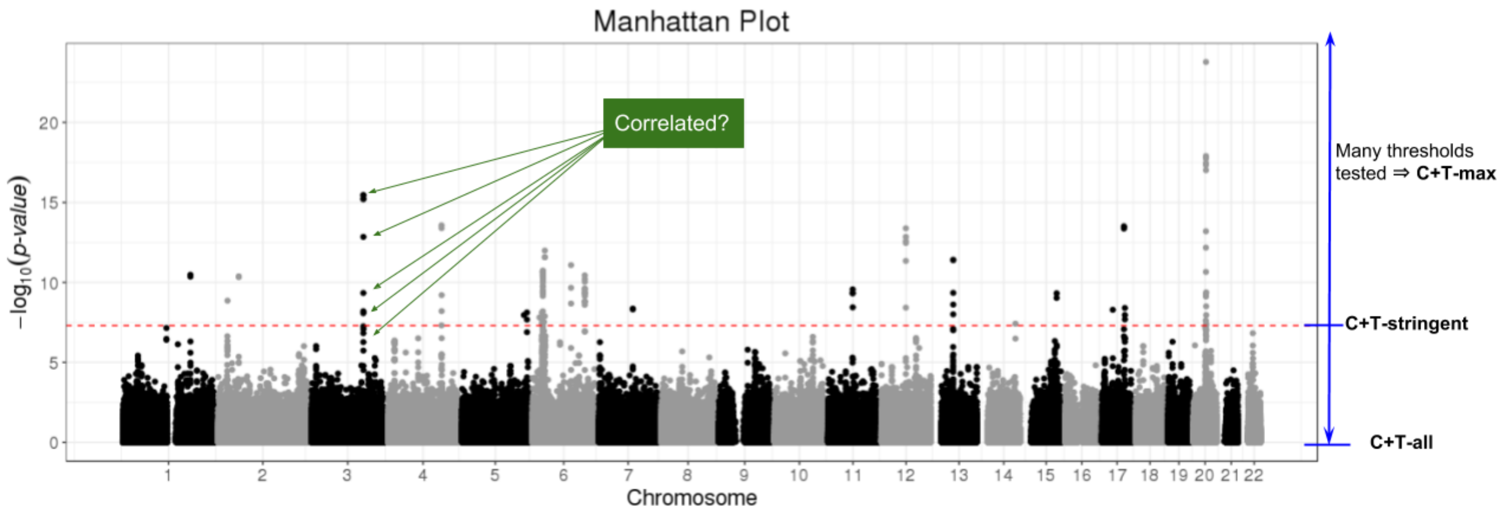
In a GWAS, each genetic variant is tested **independently**, resulting in one **effect size** $\hat{\beta}$ and one **p-value** p for each variant.



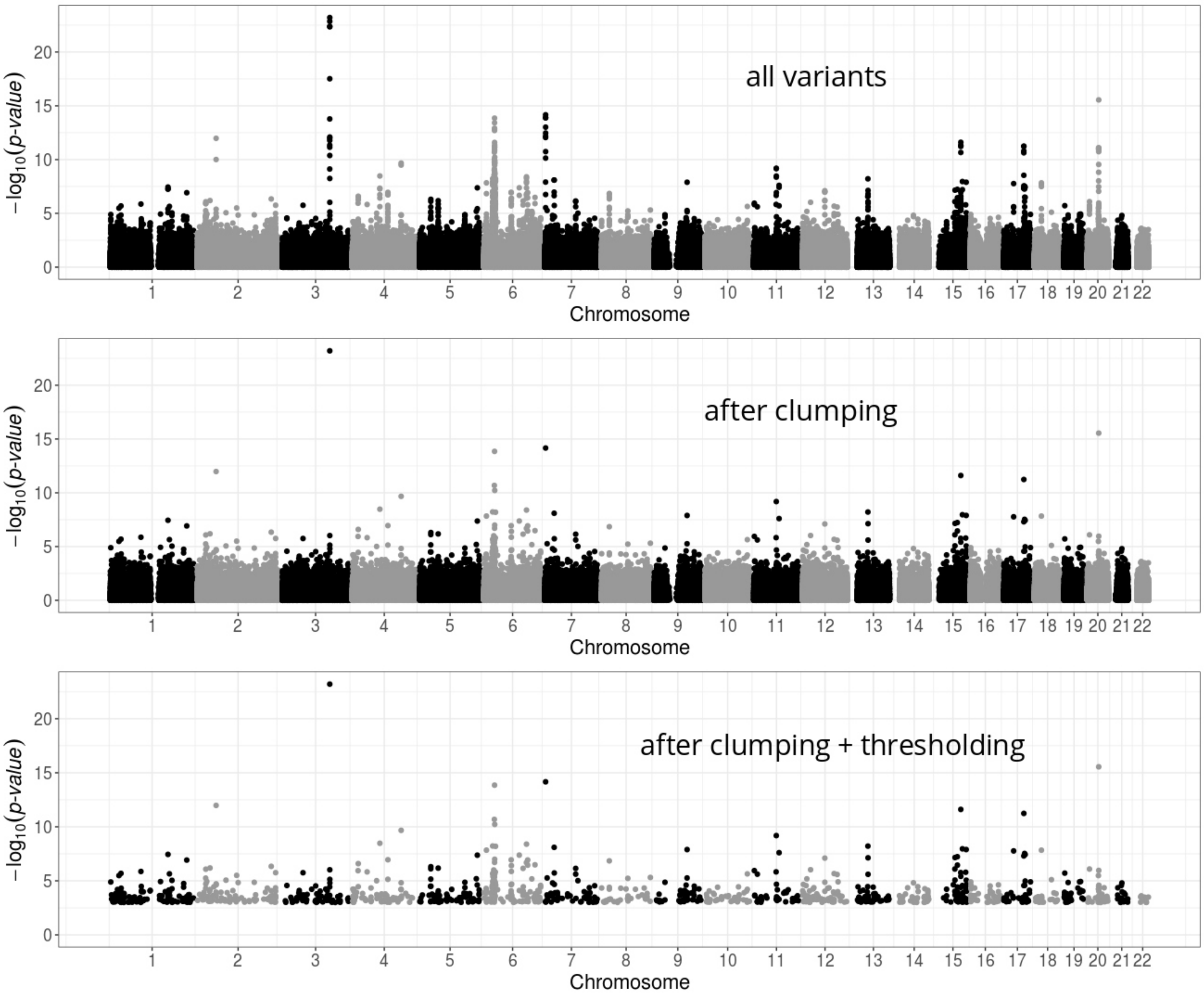
Easy combining: $PRS_i = \sum_j \hat{\beta}_j \cdot G_{i,j}$

Standard PRS - part 2: restricting predictors

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



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


all variants
after clumping
after clumping + thresholding
−log10(p-value)
0
5
10
15
20
Chromosome
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

A more optimal approach to computing PRS?

In C+T, weights are learned independently and we use heuristics for correlation and regularization.

Statistical learning

- joint models of all variants 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

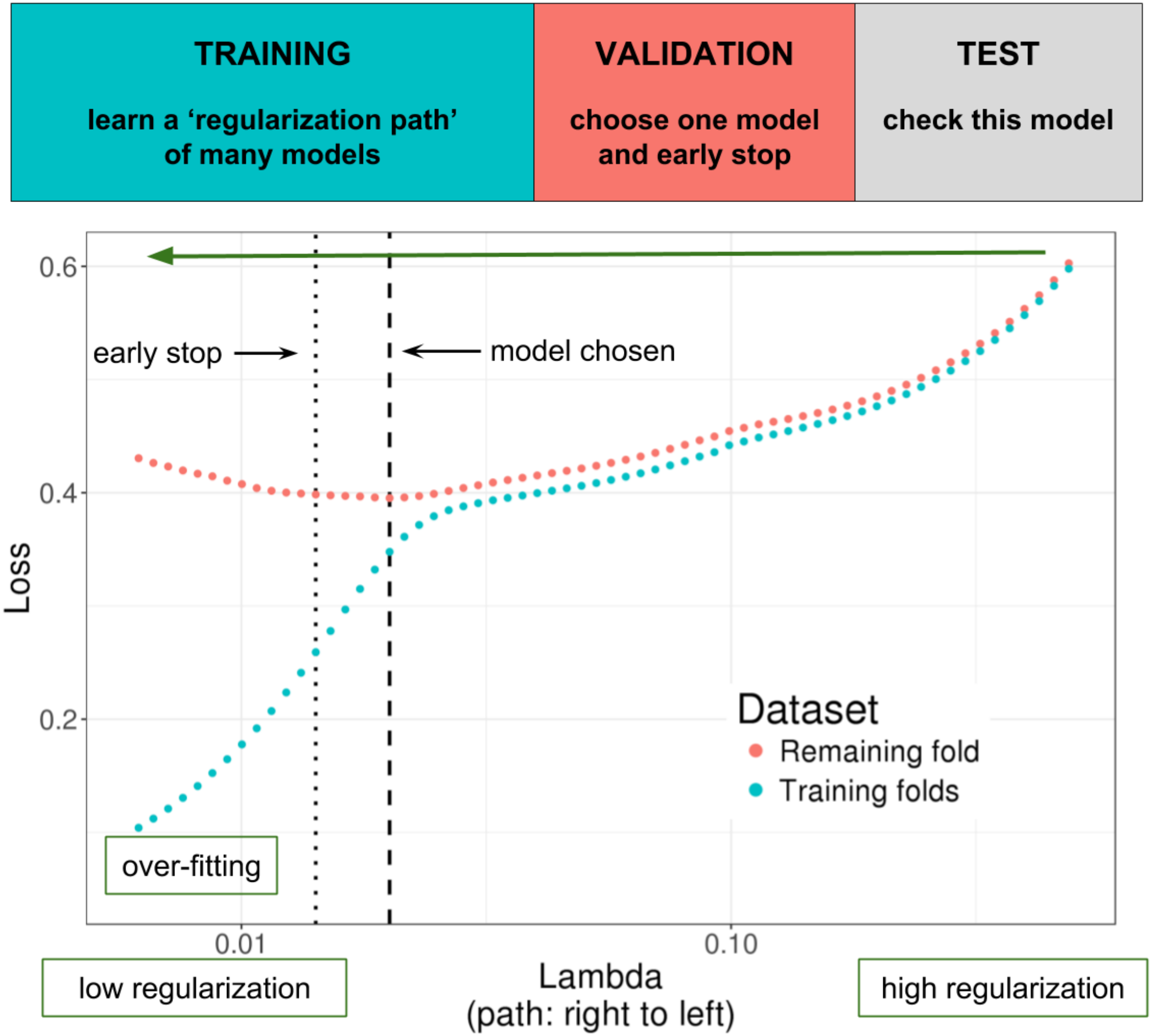
- a memory- and computation-efficient implementation of penalized regressions to be used for biobank-scale data
- an automatic choice of the regularization hyper-parameter
- a comprehensive comparison for different disease architectures

Penalized Logistic Regression (PLR)

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

- $p_i = 1/\left(1 + \exp\left(-(\beta_0 + x_i^T \beta)\right)\right)$
- x is denoting the **genotypes** and covariates (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 $0 \leq \alpha \leq 1$.

Choice of the hyper-parameter λ



TRAINING

learn a 'regularization path' of many models

VALIDATION

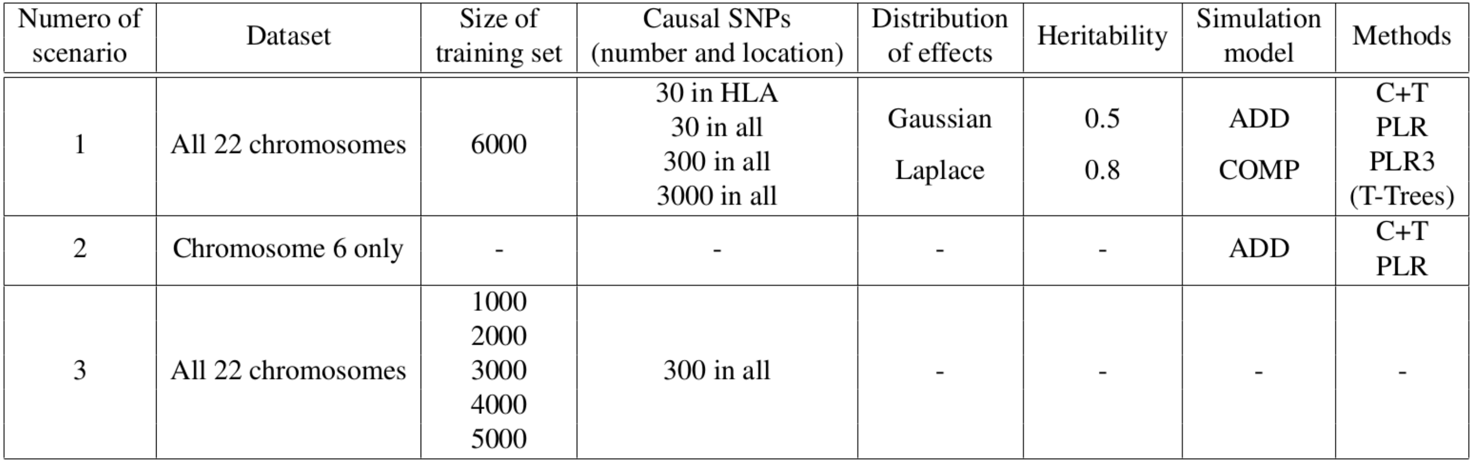
choose one model and early stop

TEST

check this model

Comprehensive simulations: varying many parameters

Simulation models (real genotypes & simulated phenotypes)



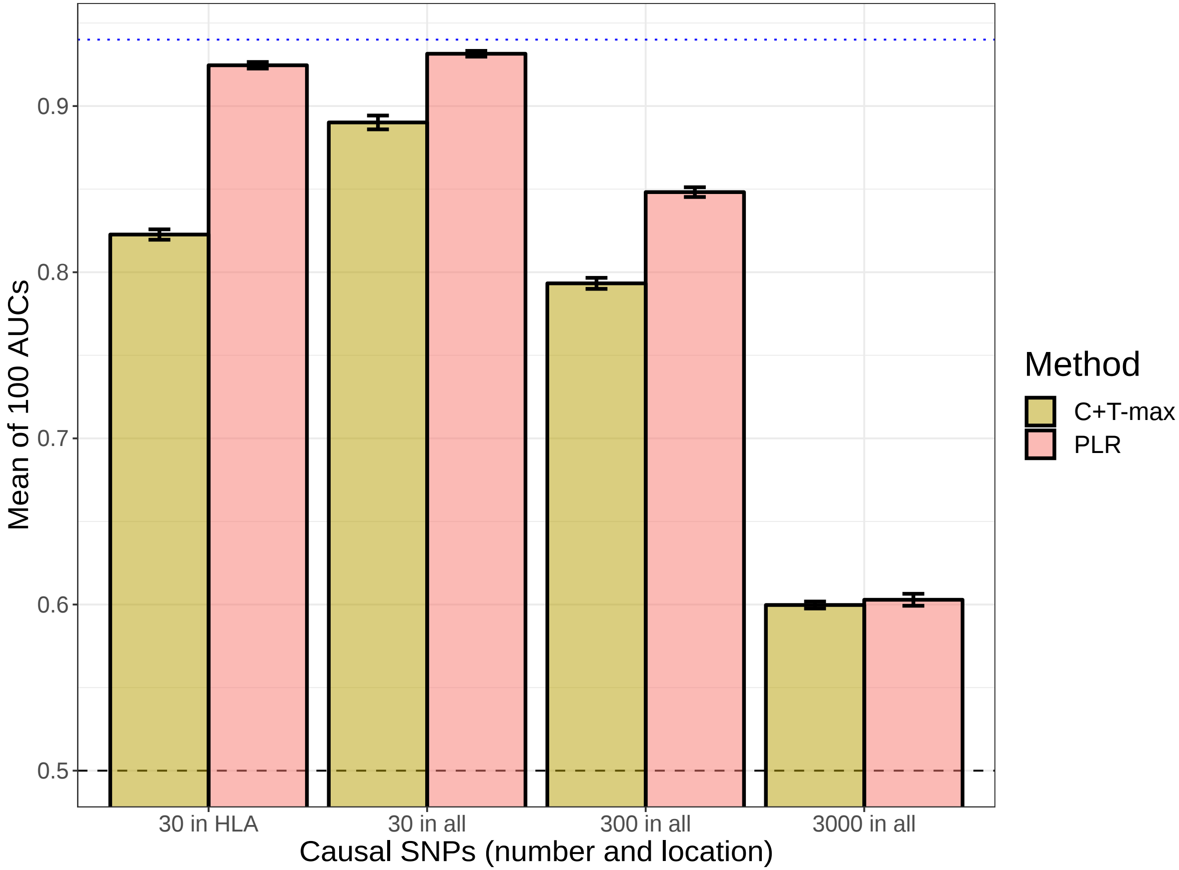
Numero of scenario	Dataset	Size of training set	Causal SNPs (number and location)	Distribution of effects	Heritability	Simulation model	Methods
1	All 22 chromosomes	6000	30 in HLA 30 in all 300 in all 3000 in all	Gaussian Laplace	0.5 0.8	ADD COMP	C+T PLR PLR3 (T-Trees)
2	Chromosome 6 only	-	-	-	-	ADD	C+T PLR
3	All 22 chromosomes	1000 2000 3000 4000 5000	300 in all	-	-	-	-

Methods compared

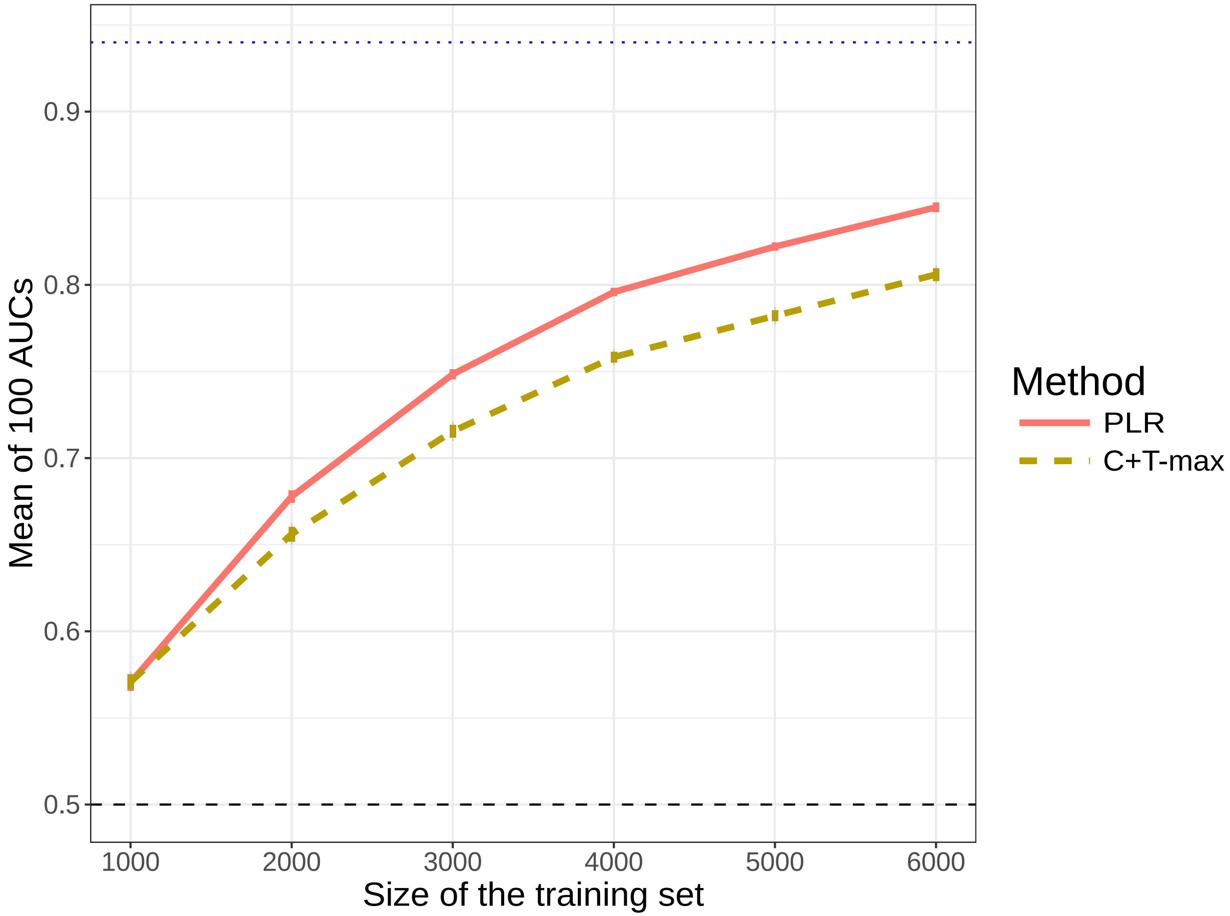
- C+T-max: best prediction for all thresholds, considered as an upper-bound
- PLR: penalized logistic regression with automatic selection of hyper-parameters
- (T-trees and PLR3)

Prediction in different simulation scenarios

$$\text{AUC (Area Under the ROC Curve)} = Prob(PRS_{\text{case}} > PRS_{\text{control}})$$



Prediction with PLR is improving faster



Real data

Celiac disease

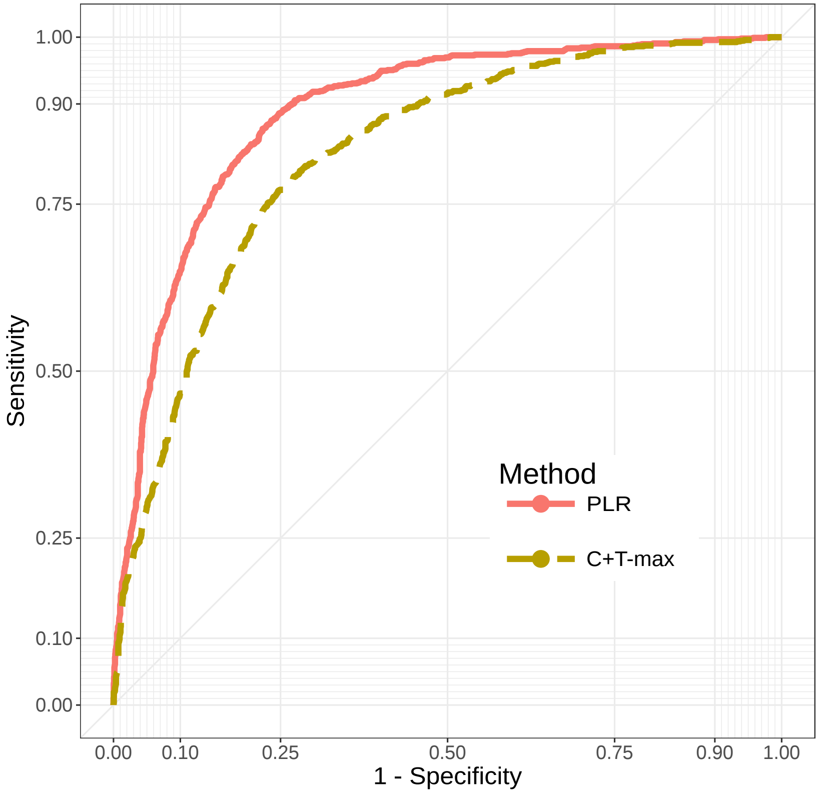
- intolerance to gluten
- only treatment: gluten-free diet
- heritability: 57-87% (Nisticò et al. 2006)
- prevalence: 1-6%

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

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

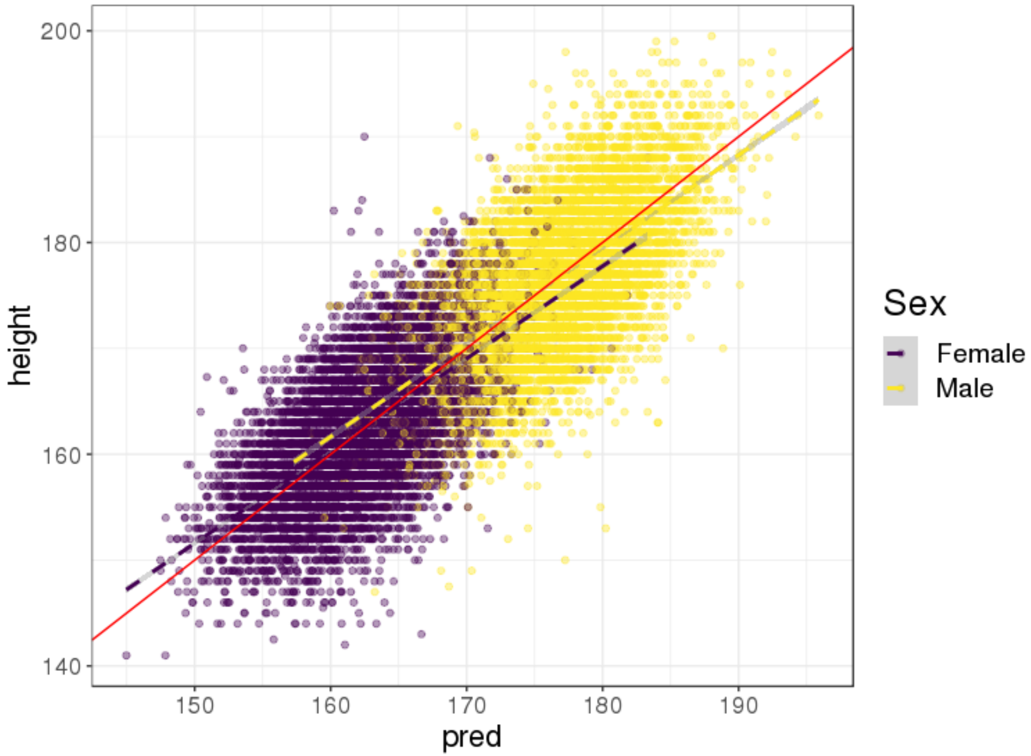
Results: real Celiac phenotypes

Method	AUC	pAUC	Execution time (s)
C+T-max	0.825 (0.0007)	0.029 (0.0002)	130 (0.14)
PLR	0.887 (0.0006)	0.041 (0.0002)	190 (1.2)



PLR for predicting height

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



Summary of our penalized regression as compared to the C+T method

- A more **optimal** approach for predicting complex diseases, providing more predictive models as long as one of
    - there are moderate effects,
    - there is some correlation between causal variants
    - sample size if large enough
- models are **linear** and **sparse**
- very **fast** and scalable to very large datasets such as the UK Biobank
- **automatic choice** for the two hyper-parameters of PLR
- can be extended to capture also recessive and dominant effects
- can be extended to integrate external summary statistics information

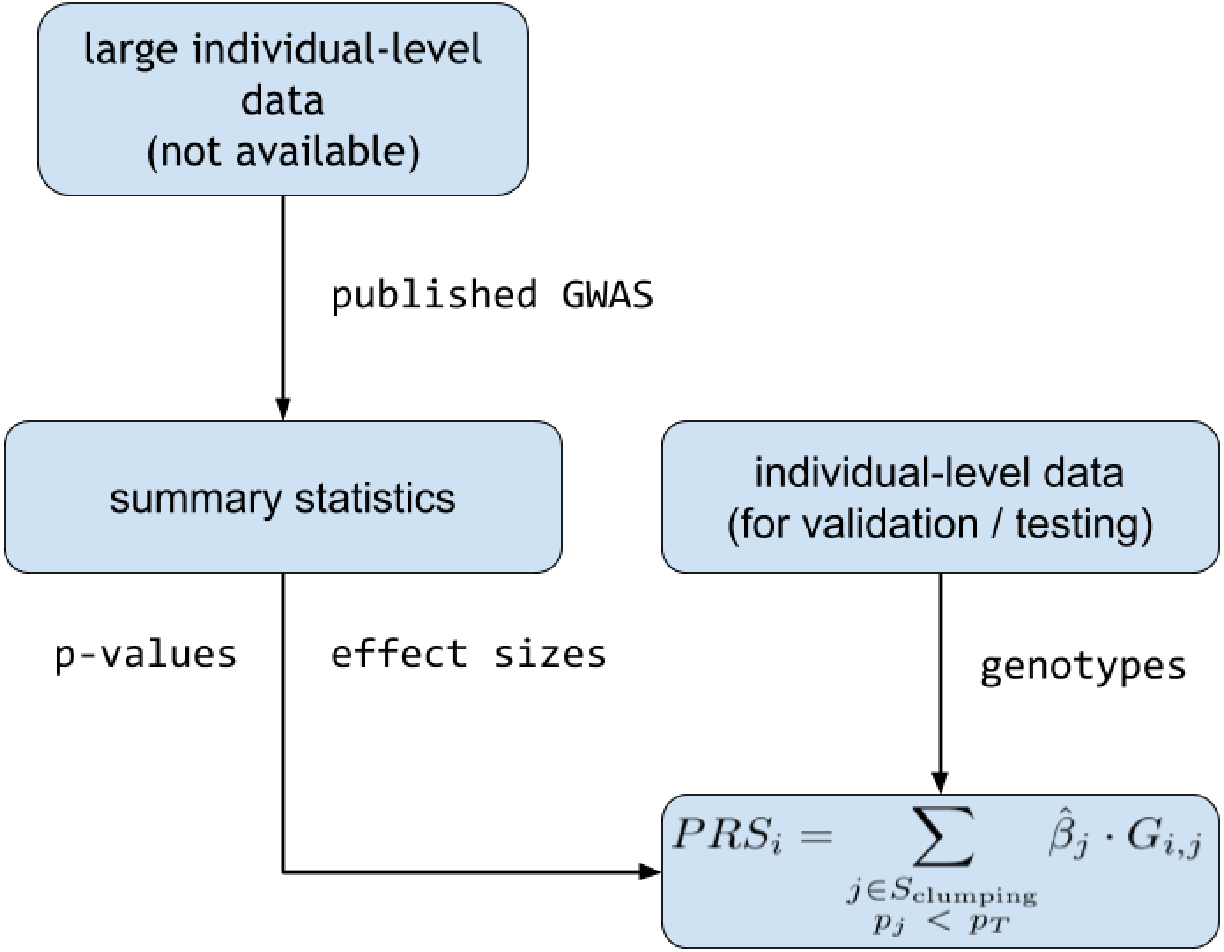
However, need to have access to large individual-level data.

Prediction using summary statistics

Privé, F., Vilhjálmsson, B. J., Aschard, H., & Blum, M. G. (2019).
Making the most of Clumping and Thresholding for polygenic scores.
bioRxiv, 653204.

[in revision in the American Journal of Human Genetics]

Using summary statistics from large GWAS


large individual-level data (not available)
published GWAS
summary statistics
individual-level data (for validation / testing)
p-values
effect sizes
genotypes
$PRS_i = \sum_{\substack{j \in S_{\text{clumping}} \\ p_j \; < \; p_T}} \hat{\beta}_j \cdot G_{i,j}$

Predictive methods based on summary statistics

When you have only summary statistics (and a small individual-level dataset), you can use:

- C+T

- LDpred (*Vilhjálmsson, 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.*)

- Other methods in development, such as NPS, PRS-CS and SBayesR.

The idea of LDpred, lassosum and the other methods is to use a reference panel to **account for correlation** between variants, instead of clumping (removing) variants.

Making the most of C+T

Hyper-parameters in C+T

- threshold of imputation quality score ($INFO_T \sim 0.3$)
- threshold on squared correlation of clumping ($r_c^2 \sim 0.2$) and window size for LD computation ($w_c \sim 500kb$)
- p-value threshold (p_T between 1 and 10^{-8} and choose the best one)

$\Longrightarrow$ *stdCT* (standard C+T)

Our contribution

- an efficient implementation to compute many C+T scores for different hyper-parameters (**5600 sets of hyper-parameters** × 22 chromosomes)
  $\Longrightarrow$ *maxCT* (maximized C+T)
- going further by **stacking** (*Breiman, Leo. "Stacked regressions." Machine learning 24.1 (1996): 49-64.*) with a linear combination of all C+T models (instead of just choosing the best model)
  $\Longrightarrow$ *SCT* (Stacked C+T)

Grid of hyper-parameters and Stacking

We compute C+T scores *for each chromosome separately* and for several parameters:

- **Threshold on imputation** INFO score INFO_T within **{0.3, 0.6, 0.9, 0.95}**.
- Squared correlation **threshold of clumping** r_c^2 within **{0.01, 0.05, 0.1, 0.2, 0.5, 0.8, 0.95}**.
- Base **size of clumping window** within {50, 100, 200, 500}. The window size w_c is then computed as the base size divided by r_c^2. For example, for $r_c^2 = 0.2$, we test values of w_c within {250, 500, 1000, 2500} (in kb).
- A sequence of **50 thresholds on p-values** between the least and the most significant p-values, equally spaced on a log-log scale.

Then, we **stack these 123,200 C+T scores** by using them as variables in the efficient penalized regressions we implemented previously.

Data (simulations)

Real genotypes

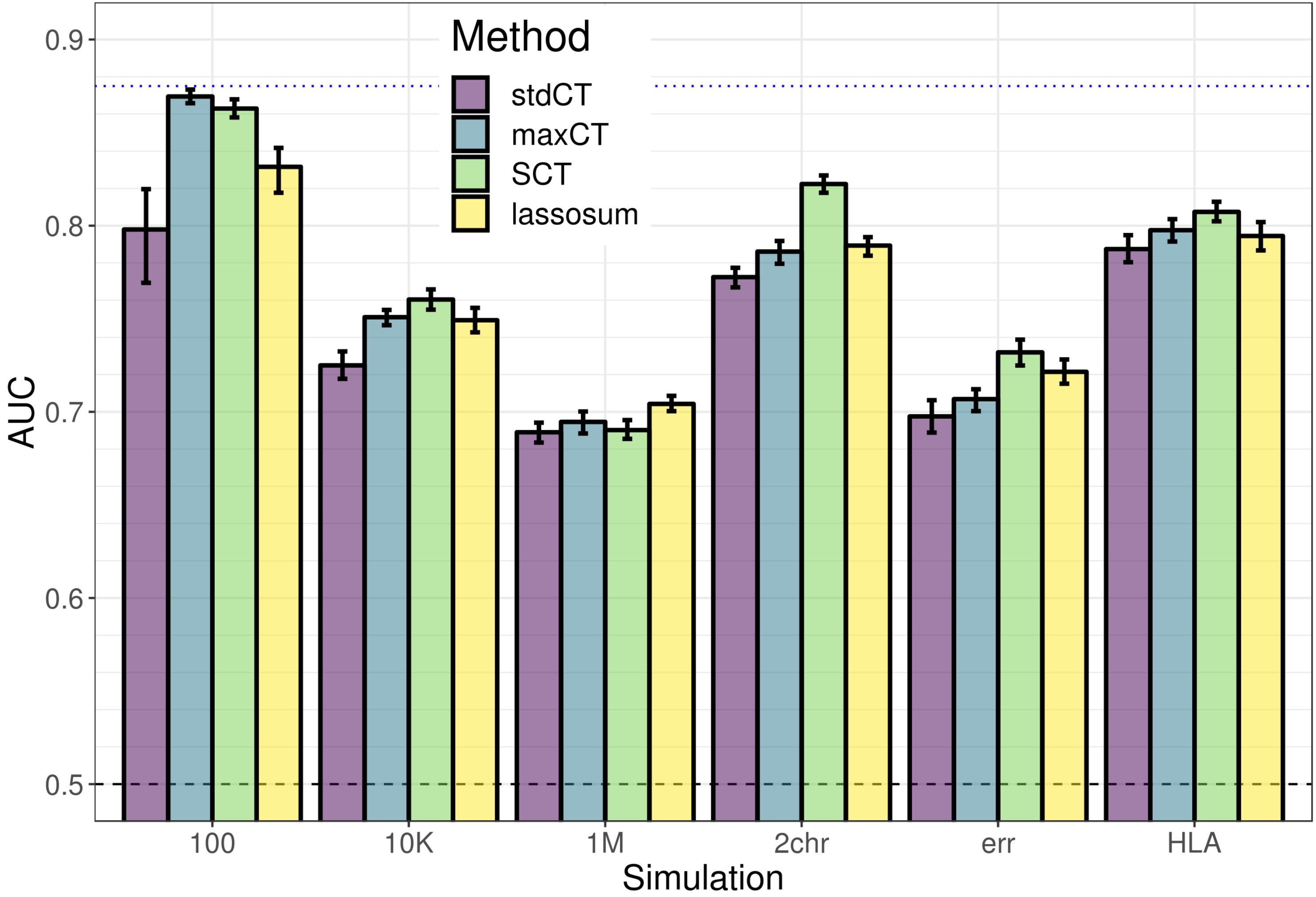
UK Biobank data for 1M variants and:

- 315,609 individuals for computing summary statistics (GWAS),
- a set of 10,000 individuals for training hyper-parameters and lastly
- a test set of 10,000 individuals for evaluating models.

Simulate new phenotypes

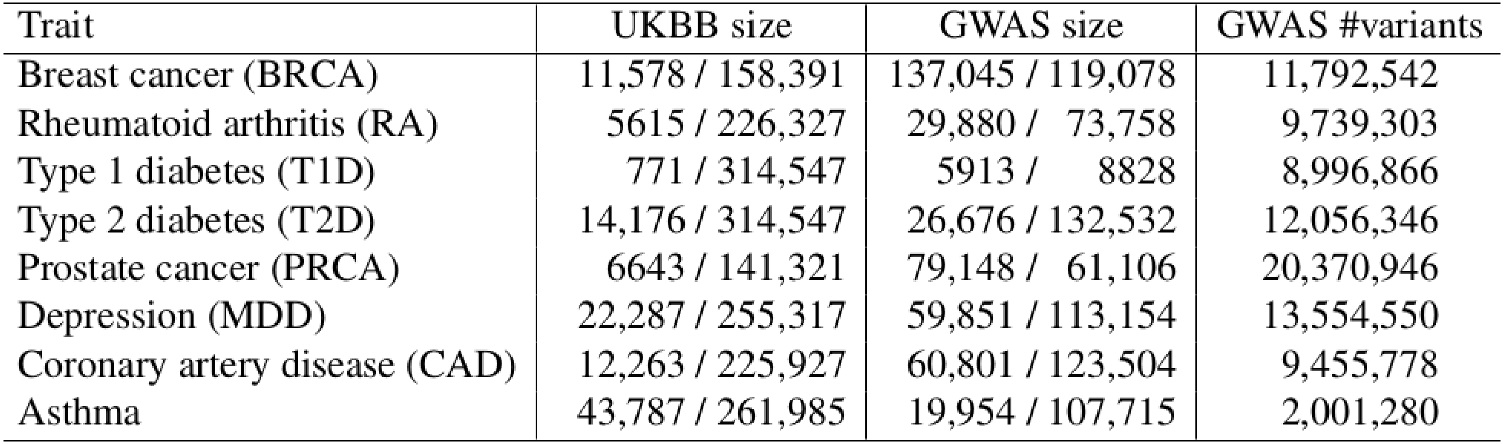
- 100, 10K, or 1M random causal variants with Gaussian effects
- Three additional scenarios with more complex architectures:
  - "2chr": 100 variants of chromosome 1 and all variants of chromosome 2 are causal
  - "err": (not presented)
  - "HLA": 7105 causal variants are chosen in one long-range LD region

Results (simulations)



Data (real phenotypes)

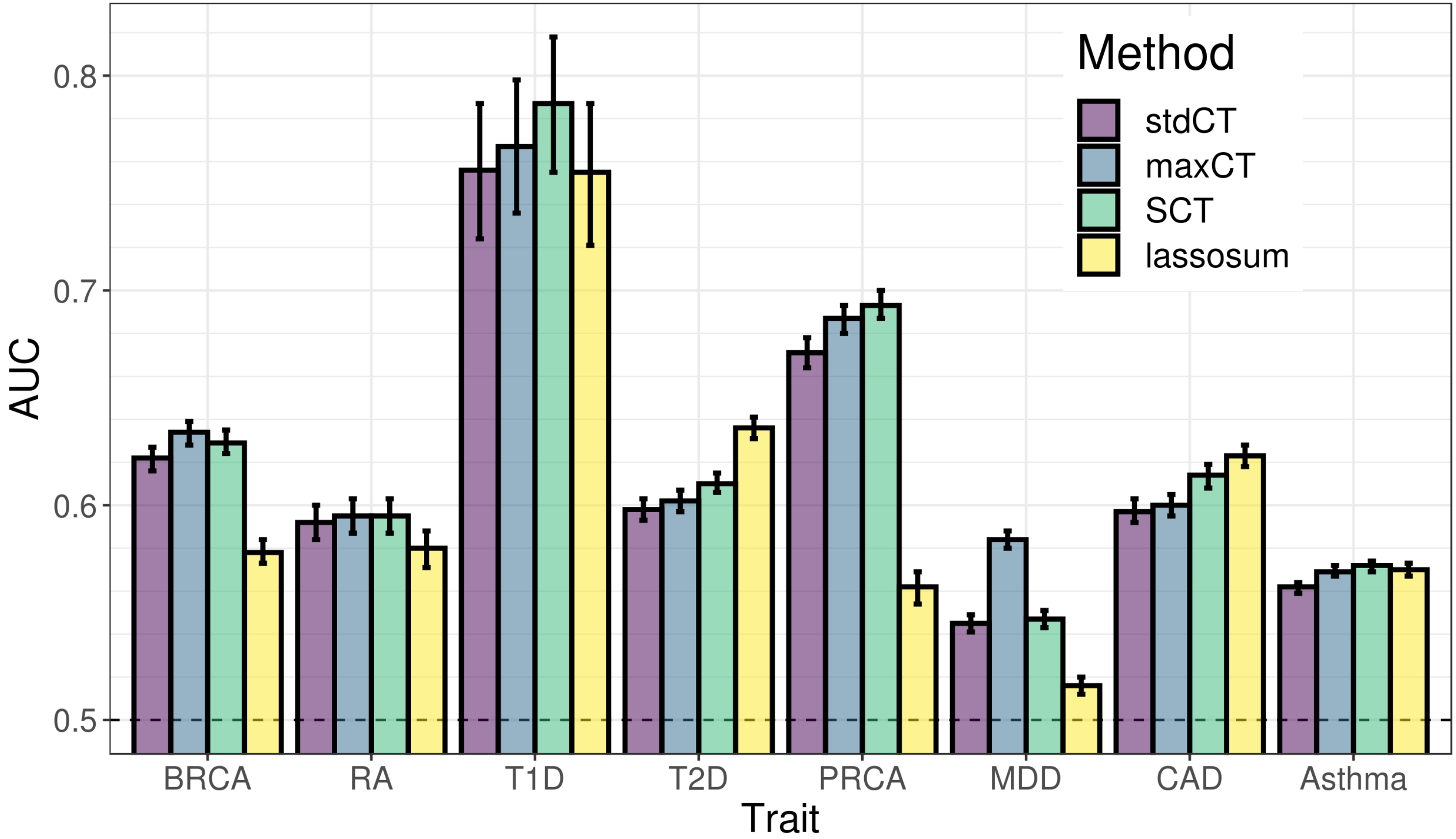
- Include 8 common disorders
- Real genotypes + phenotypes (UK Biobank) for training/validation/test
- External published summary statistics (that did not use UK Biobank)



Trait	UKBB size	GWAS size	GWAS #variants
Breast cancer (BRCA)	11,578 / 158,391	137,045 / 119,078	11,792,542
Rheumatoid arthritis (RA)	5615 / 226,327	29,880 / 73,758	9,739,303
Type 1 diabetes (T1D)	771 / 314,547	5913 / 8828	8,996,866
Type 2 diabetes (T2D)	14,176 / 314,547	26,676 / 132,532	12,056,346
Prostate cancer (PRCA)	6643 / 141,321	79,148 / 61,106	20,370,946
Depression (MDD)	22,287 / 255,317	59,851 / 113,154	13,554,550
Coronary artery disease (CAD)	12,263 / 225,927	60,801 / 123,504	9,455,778
Asthma	43,787 / 261,985	19,954 / 107,715	2,001,280

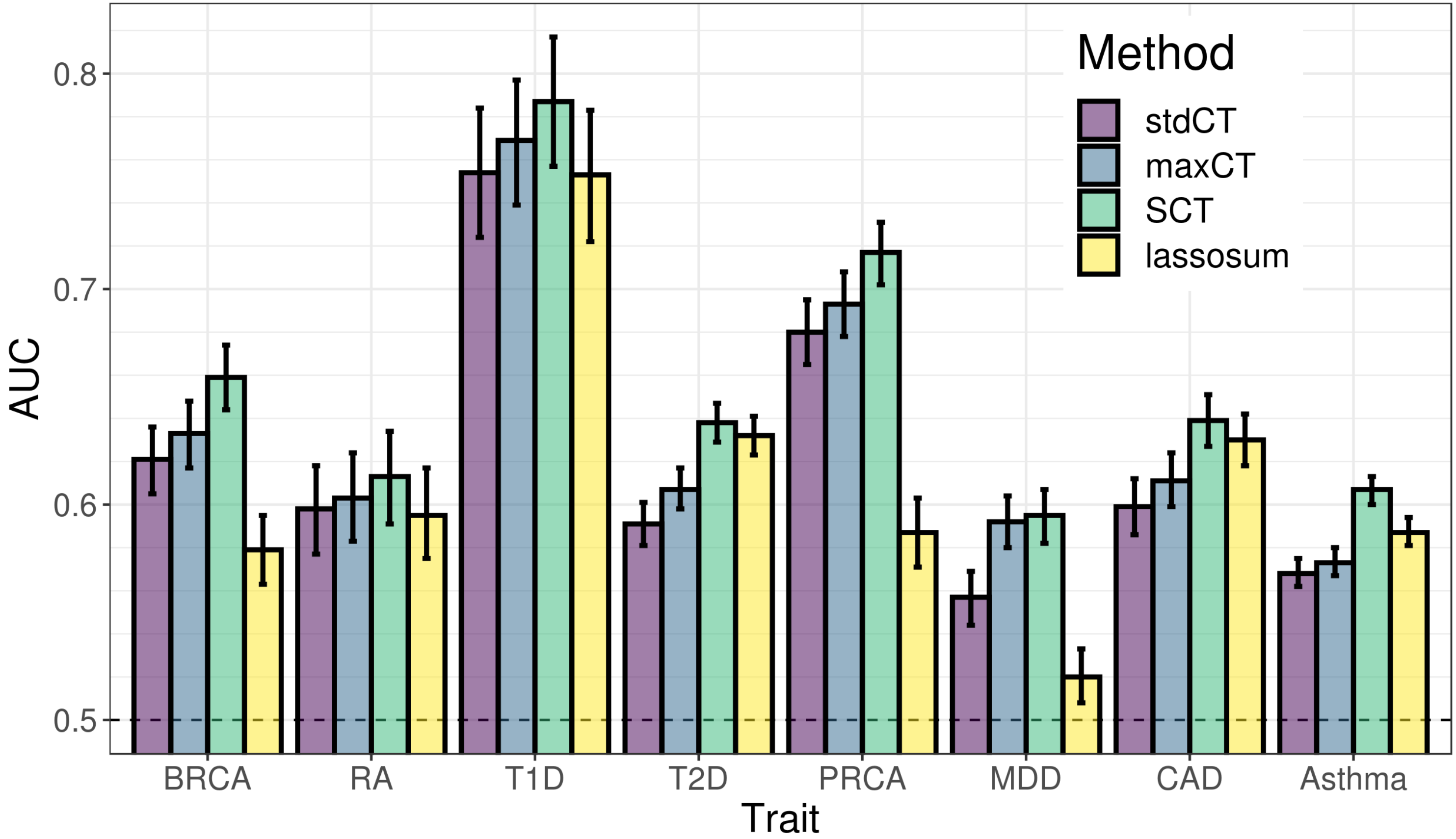
Results (small training set)

500 cases and 2000 controls in training



Results (large training set)

Between 120K and 350K individuals in training



Summary

- We improved C+T by tuning more hyper-parameters
- maxCT is on par with lassosum, while being more robust (no model)
- stacking makes C+T more flexible and potentially much more predictive
- predictive power of SCT is increasing with sample size
- can extend SCT to account for other parameters (e.g. MAF)
- can extend SCT to use multiple summary statistics

Conclusion

My thesis work

1. Developping two R 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 (in bigstatsr) 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/10.1534/genetics.119.302019)

3. Extending the set of parameters tested in C+T (implemented in bigsnpr) to achieve higher predictive performance with C+T. Extension via stacking. Comparison with standard C+T, lassosum (and LDpred).

   (https://doi.org/10.1101/653204)

Directions of future work

- Revisions for C+T/SCT paper
    - add LDpred to the comparisons
    - investigate MAF parameter
- Coding in bigsnpr
    - clumping and PCA directly on PLINK files with missing values
    - improving autoSVD algorithm, including automatic detection of outlier samples on top of long-range LD regions
- multi-phenotype prediction with SCT (e.g. for schizophrenia, bipolar disorder and depression)
- testing of different scaling functions in penalized regressions
- inclusion of summary statistics information in penalized regressions
- coding of penalized Cox regression
- comparison of PRS methods (via data challenge?)

I thank you for your attention

Presentation available at

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

privefl privefl F. Privé

Slides created via R package **xaringan**.