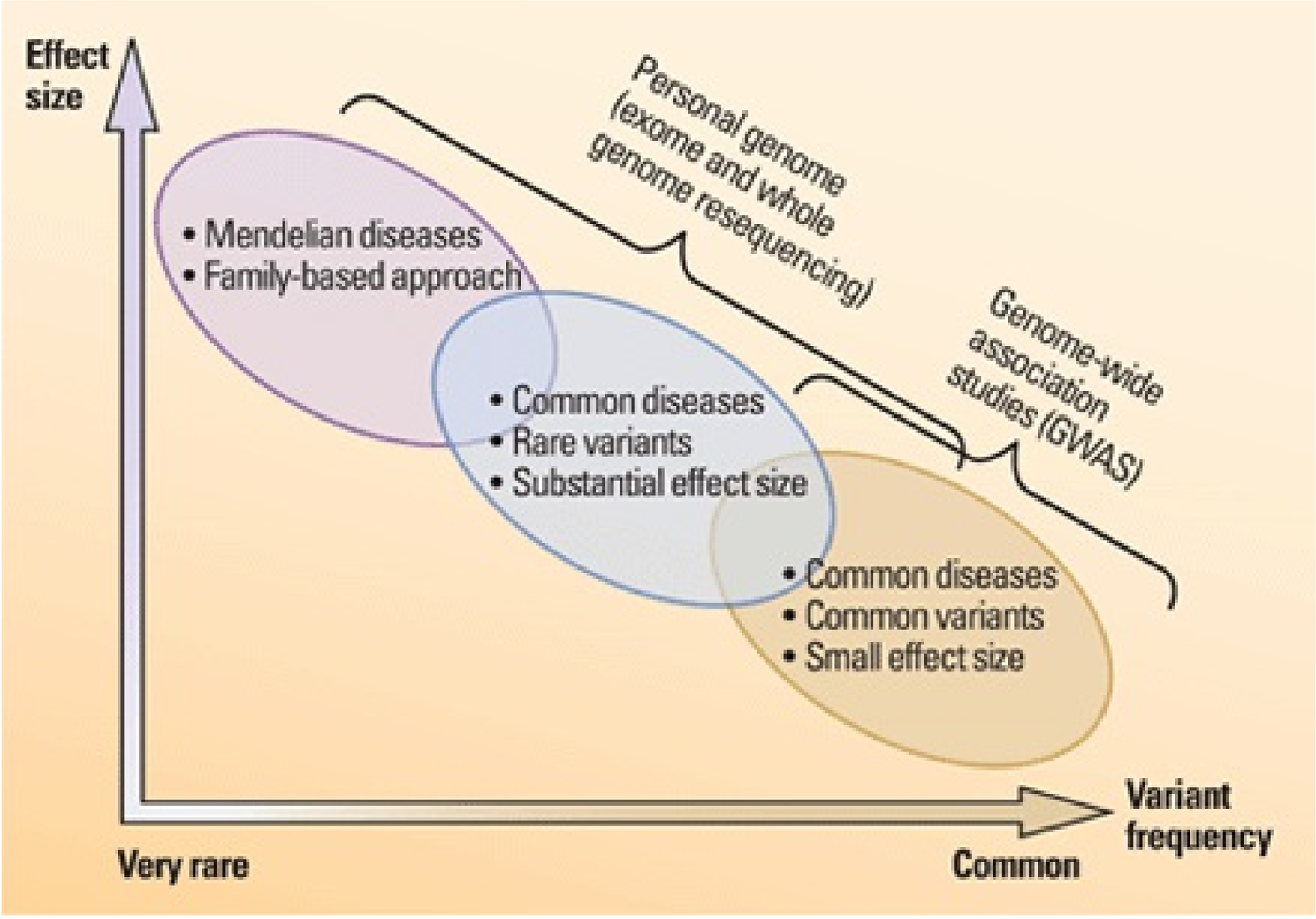
Predicting complex diseases: performance and robustness

Florian Privé

March 23, 2018

Introduction

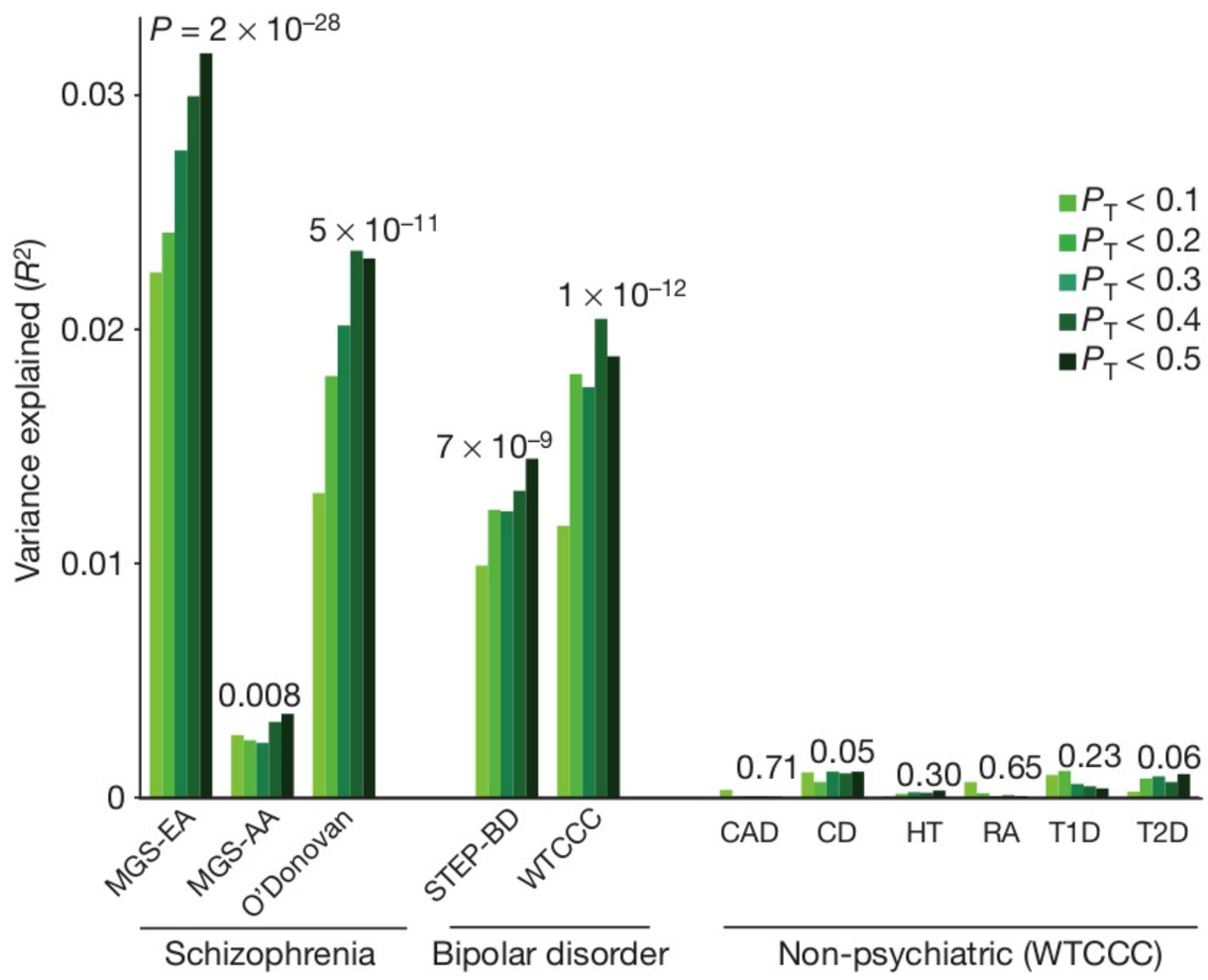
Disease architectures



Source: 10.1126/science.338.6110.1016

Polygenic Risk Scores (PRS)

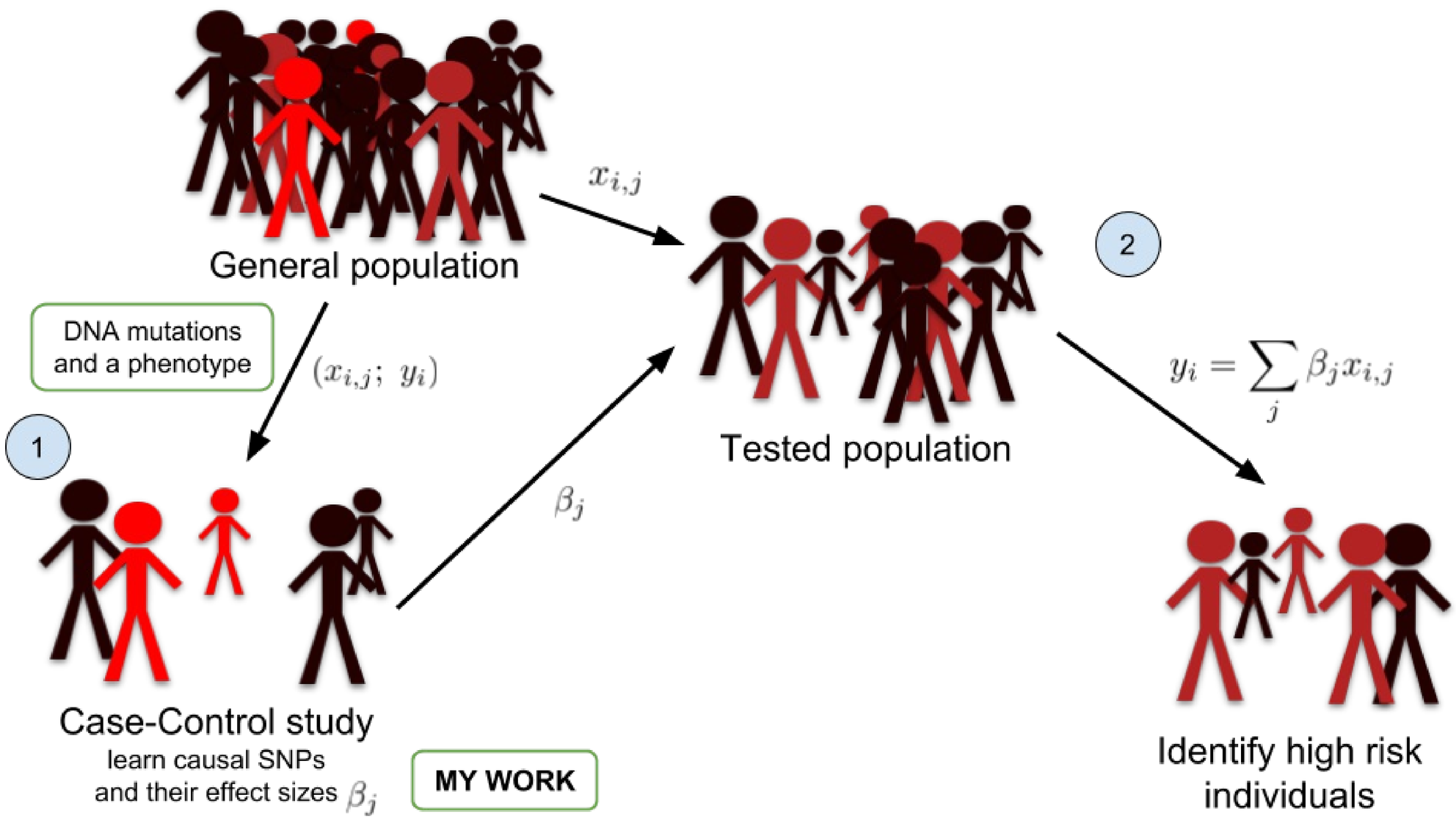
One application: to provide genetic evidence



Source: 10.1038/nature08185

Polygenic Risk Scores (PRS)

Another application: to identify high risk individuals



Predictive methods

Methods already developed by other people

- **GWAS** + **Clumping** + **Thresholding** (C+T)
- Linear Mixed Models
- Statistical Learning such as
  - **Logistic Regression**
  - Support Vector Machine
  - Decision tree methods such as Random Forests

Our 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

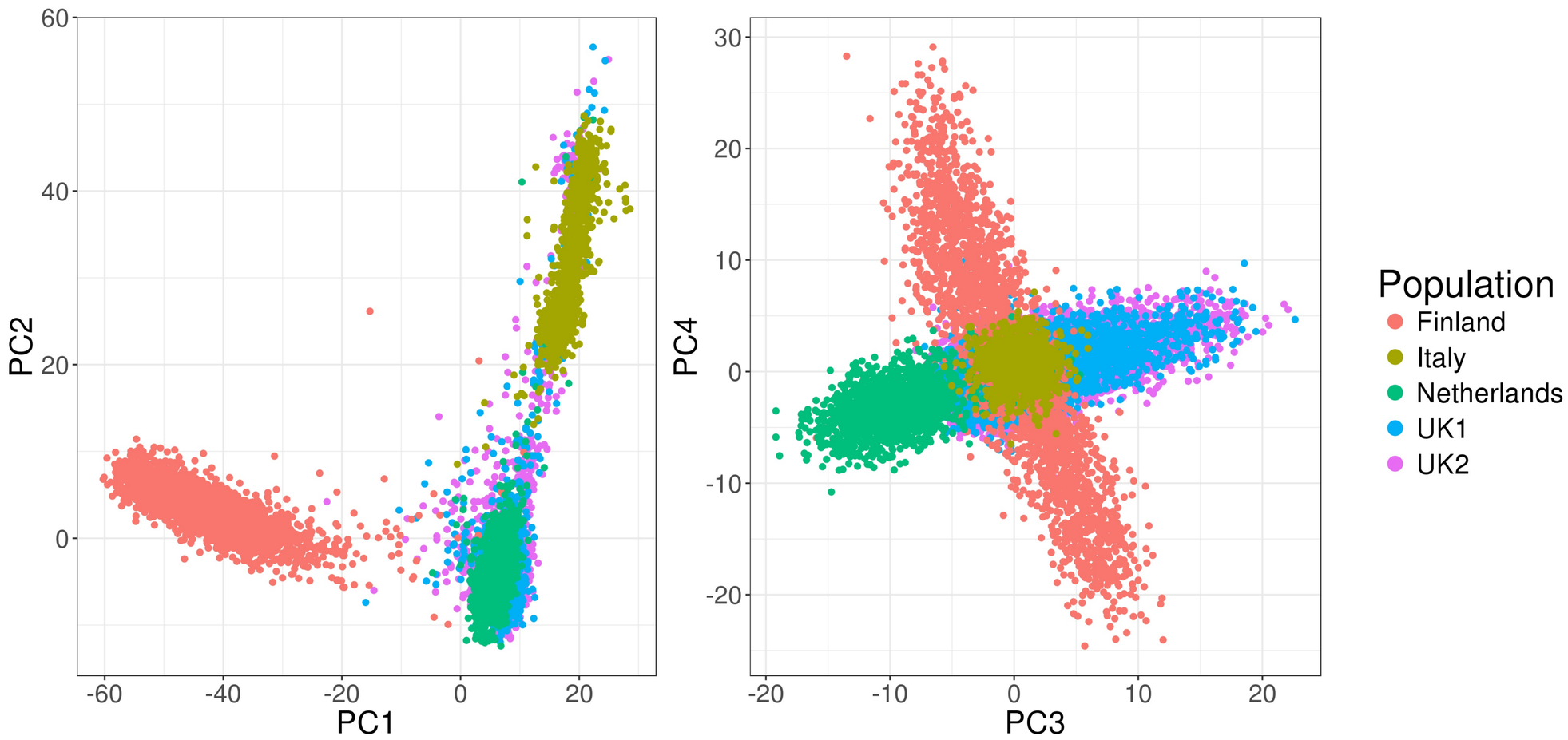
Package {bigstatsr} provides a **fast penalized logistic regression**.

ACTION	STATUS	ID	TITLE	SUBMITTED	DECISIONED
	• Accept after Review (22-Mar-2018) • With production editor	BIOINF-2017-1798.R1	Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr View Submission	02-Feb-2018	22-Mar-2018

Methods

Real genotype data

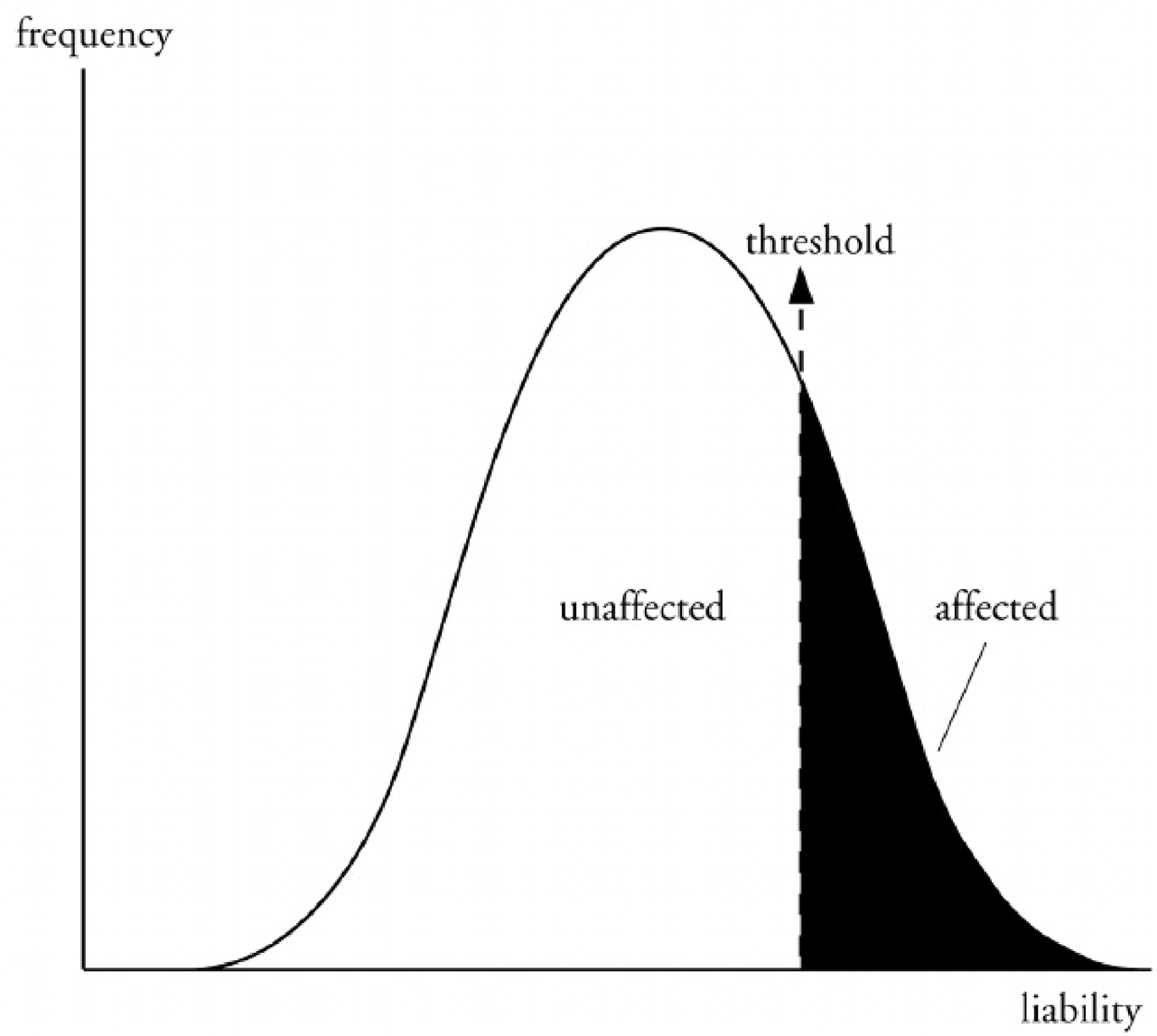
Use real data from a case-control study for the Celiac disease.



Keep only **controls** from the **UK** and **not deviating from the robust Malahanobis distance**.

Simulate new phenotypes

The liability-threshold model



Two models of liability

A "simple" model

$$y_i = \underbrace{\sum_{j \in S_{\text{causal}}} w_j \cdot \widetilde{G_{i,j}}}_{\text{genetic effect}} + \underbrace{\epsilon_i}_{\text{environmental effect}}$$

A "fancy" model

$$y_i = \underbrace{\sum_{j \in S^{(1)}_{\text{causal}}} w_j \cdot \widetilde{G_{i,j}}}_{\text{linear}} + \underbrace{\sum_{j \in S^{(2)}_{\text{causal}}} w_j \cdot \widetilde{D_{i,j}}}_{\text{dominant}} + \underbrace{\sum_{\substack{k=1 \\ j_1 = e_k^{(3.1)} \\ j_2 = e_k^{(3.2)}}}^{k = \left|S^{(3.1)}_{\text{causal}}\right|} w_{j_1} \cdot \widetilde{G_{i,j_1} G_{i,j_2}}}_{\text{interaction}} + \epsilon_i$$

- w_j are **weights** (generated with a Gaussian or a Laplace distribution)
- $G_{i,j}$ is the **allele count** of individual i for SNP j
- $D_{i,j} = 1\left\{G_{i,j} \neq 0\right\}$

Comprehensive simulations

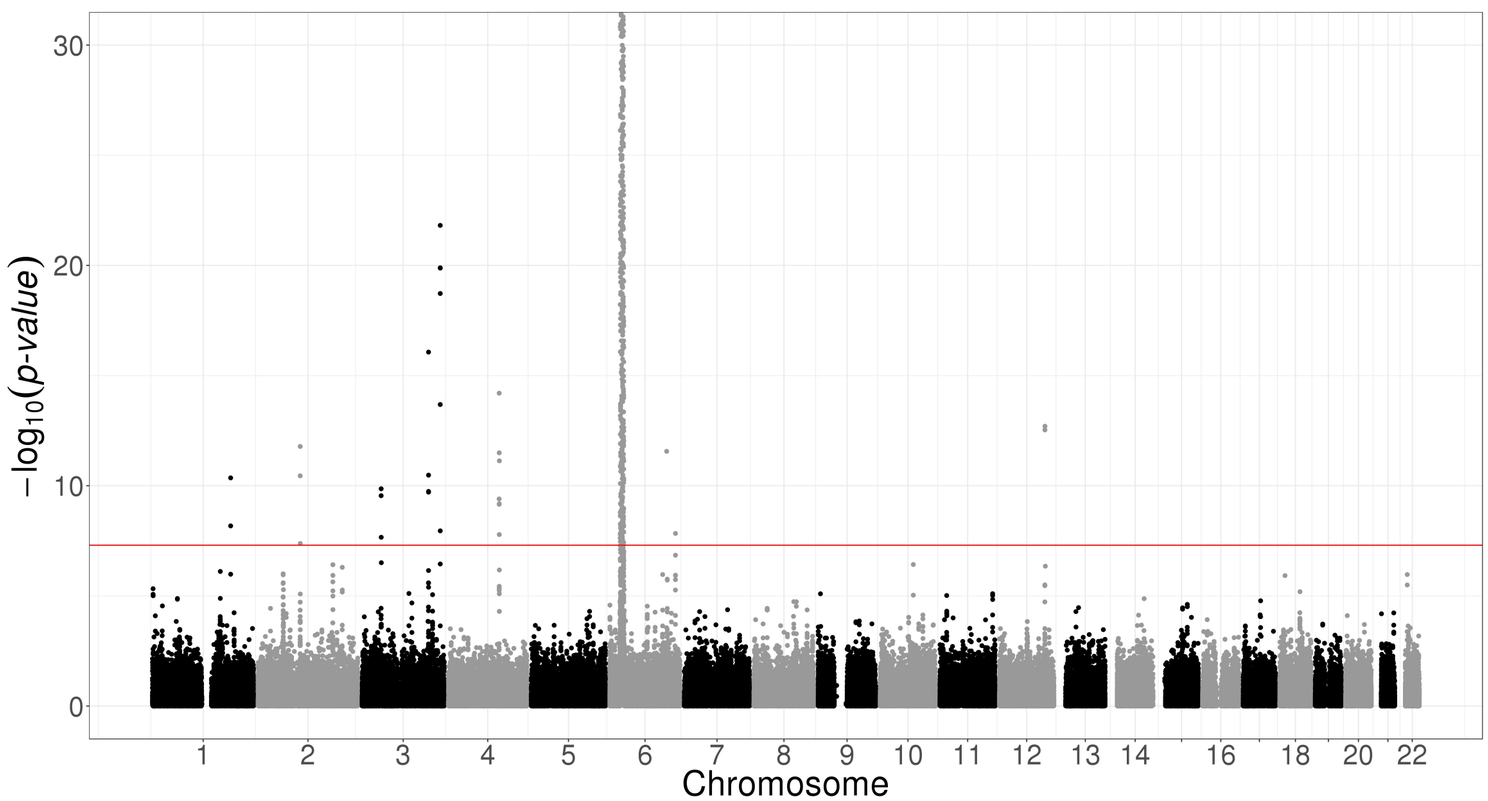
Varying many parameters

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	simple fancy	PRS logit-simple logit-triple (T-Trees)
2	Chromosome 6 only	-	-	-	-	simple	PRS logit-simple
3	All 22 chromosomes	1000 2000 3000 4000 5000	300 in all	-	-	-	-

Methods compared

The C+T method, from GWAS results

$$PRS_i = \sum_{j \in S_{\text{clumping}}} 1\{p_j < p_T\} \cdot \beta_j \cdot G_{i,j}$$



Pitfalls: weights learned independently and heuristics for correlation and regularization.

Methods compared

T-Trees (Trees inside Trees)

- an algorithm derived from random forests
- takes into account the correlation structure among the genetic markers implied by linkage disequilibrium in GWAS data

	qc		*wtccc*	
	rf	**tt**	**rf**	**tt**
BD	0.743	**0.813**	0.918	**0.959**
CAD	0.756	**0.814**	0.998	**0.999**
HT	0.807	**0.866**	0.938	**0.969**
RA	0.806	**0.830**	0.993	**0.996**
T1D	0.860	**0.870**	0.900	**0.940**
T2D	0.758	**0.834**	0.959	**0.979**

Predictive power of RF and TT on two variants of the 6 other wtccc datasets. The $_{qc}$ columns corresponds to the "qc"-like filtered variant and the $_{wtccc}$ to the weakly filtered variant. (Parameters settings: RF: $T=1000$, $K=10000$, $N_{min}=250$; TT: $T=1000$, $K=1000$, $IC=5$, $N_{min}=2000$).
doi:10.1371/journal.pone.0093379.t004

Methods compared

Penalized Logistic Regression

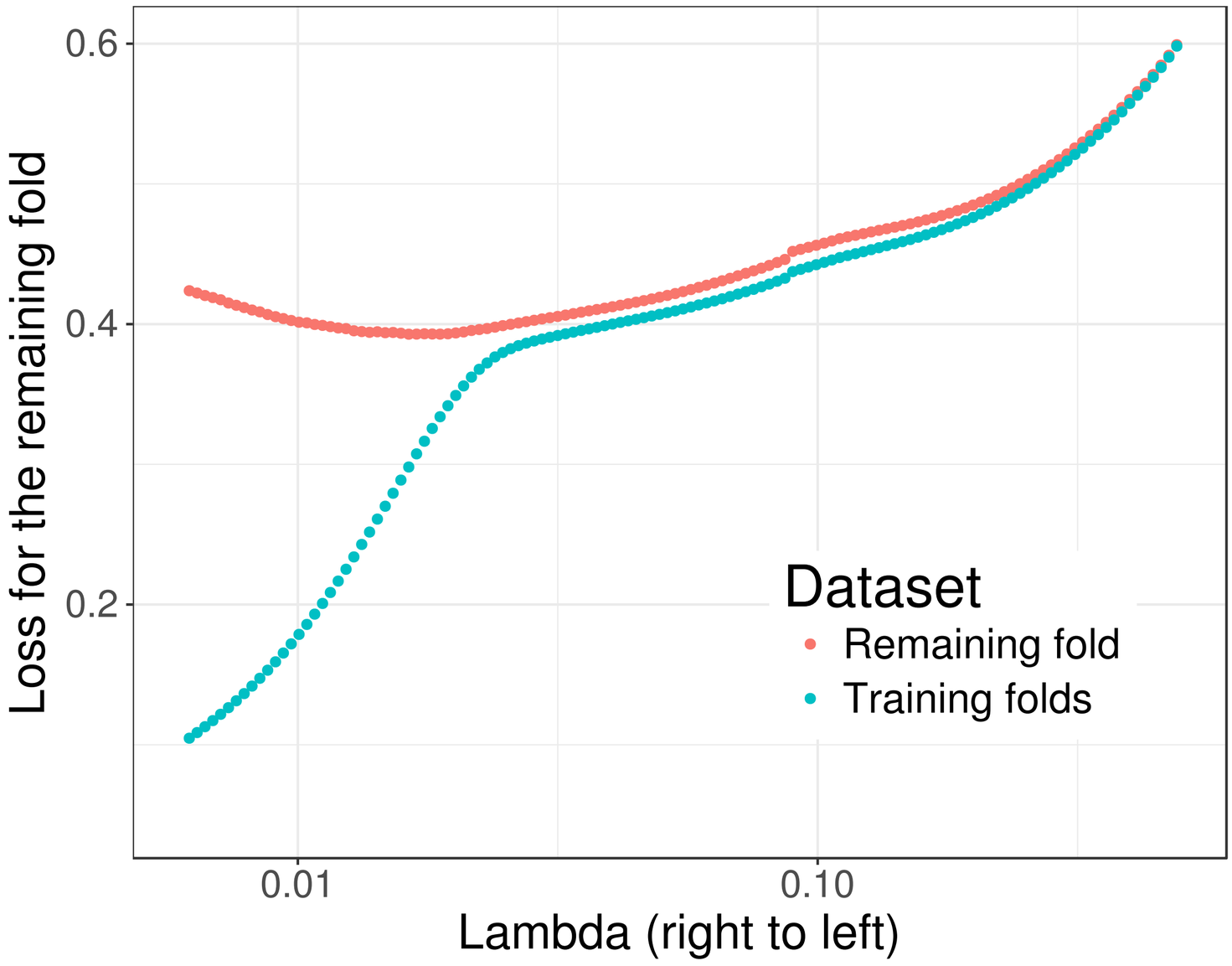
$$\underset{\beta_0,\beta}{\operatorname{argmin}}(\lambda,\alpha)\left\{\underbrace{\frac{1}{n}\sum_{i=1}^{n}\log\left(1+e^{-y_i(\beta_0+x_i^T\beta)}\right)}_{\text{Loss function}}+\underbrace{\lambda\left((1-\alpha)\frac{1}{2}\|\beta\|_2^2+\alpha\|\beta\|_1\right)}_{\text{Penalization}}\right\}$$

- 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 $0 \leq \alpha \leq 1$.

Efficient algorithm

- Strong rules for discarding predictors in lasso-type problems (Tibshirani et al., 2012)
- implemented in R package {biglasso} (Zeng et al., 2017)
- reimplemented in R package {bigstatsr} (Privé et al., 2017) with *Cross-Model Selection and Averaging (CMSA)*:

CMSA: maximization of one model



Extension via feature engineering

We construct a separate dataset with, for each SNP variable, two more variables coding for recessive and dominant effects.

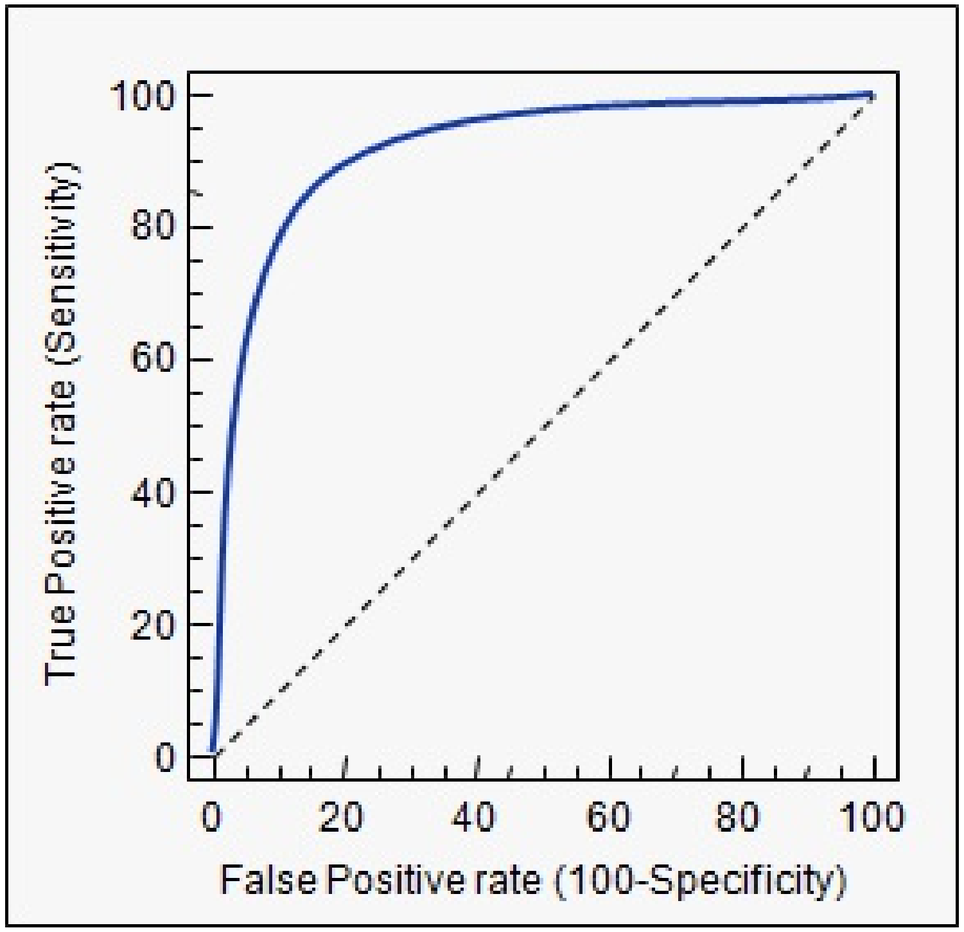
	SNP1	SNP2
[1,]	0	2
[2,]	0	1
[3,]	1	1
[4,]	0	2
[5,]	0	0
[6,]	1	0
[7,]	1	1
[8,]	0	1
[9,]	0	1
[10,]	0	2

	SNP1.1	SNP1.2	SNP1.3	SNP2.1	SNP2.2	SNP2.3
[1,]	0	0	0	2	1	1
[2,]	0	0	0	1	1	0
[3,]	1	1	0	1	1	0
[4,]	0	0	0	2	1	1
[5,]	0	0	0	0	0	0
[6,]	1	1	0	0	0	0
[7,]	1	1	0	1	1	0
[8,]	0	0	0	1	1	0
[9,]	0	0	0	1	1	0
[10,]	0	0	0	2	1	1

We call these two methods "logit-simple" and "logit-triple".

Predictive performance measures

AUC (Area Under the ROC Curve) is used.

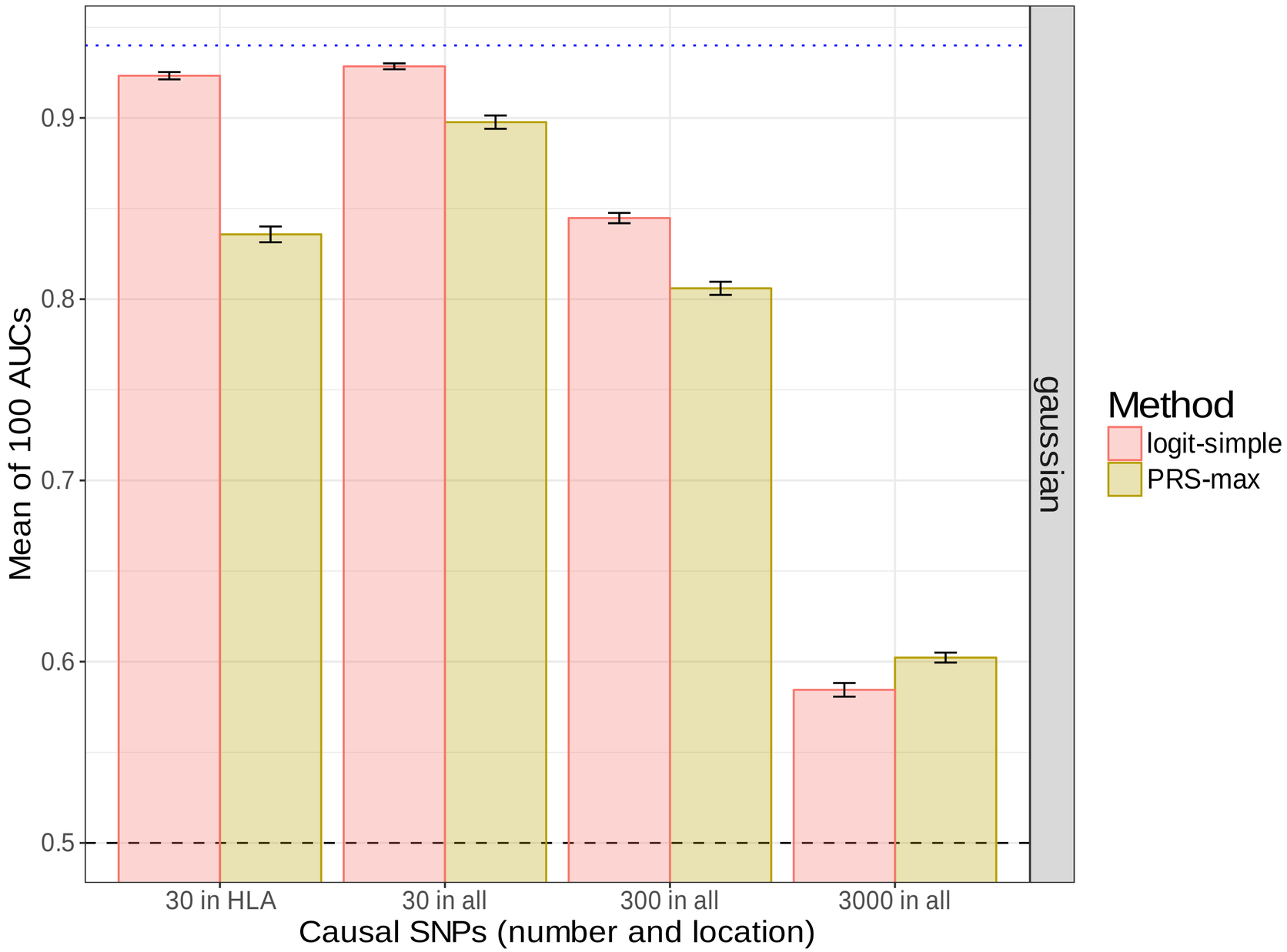


$$\text{AUC} = P(S_{\text{case}} > S_{\text{control}})$$

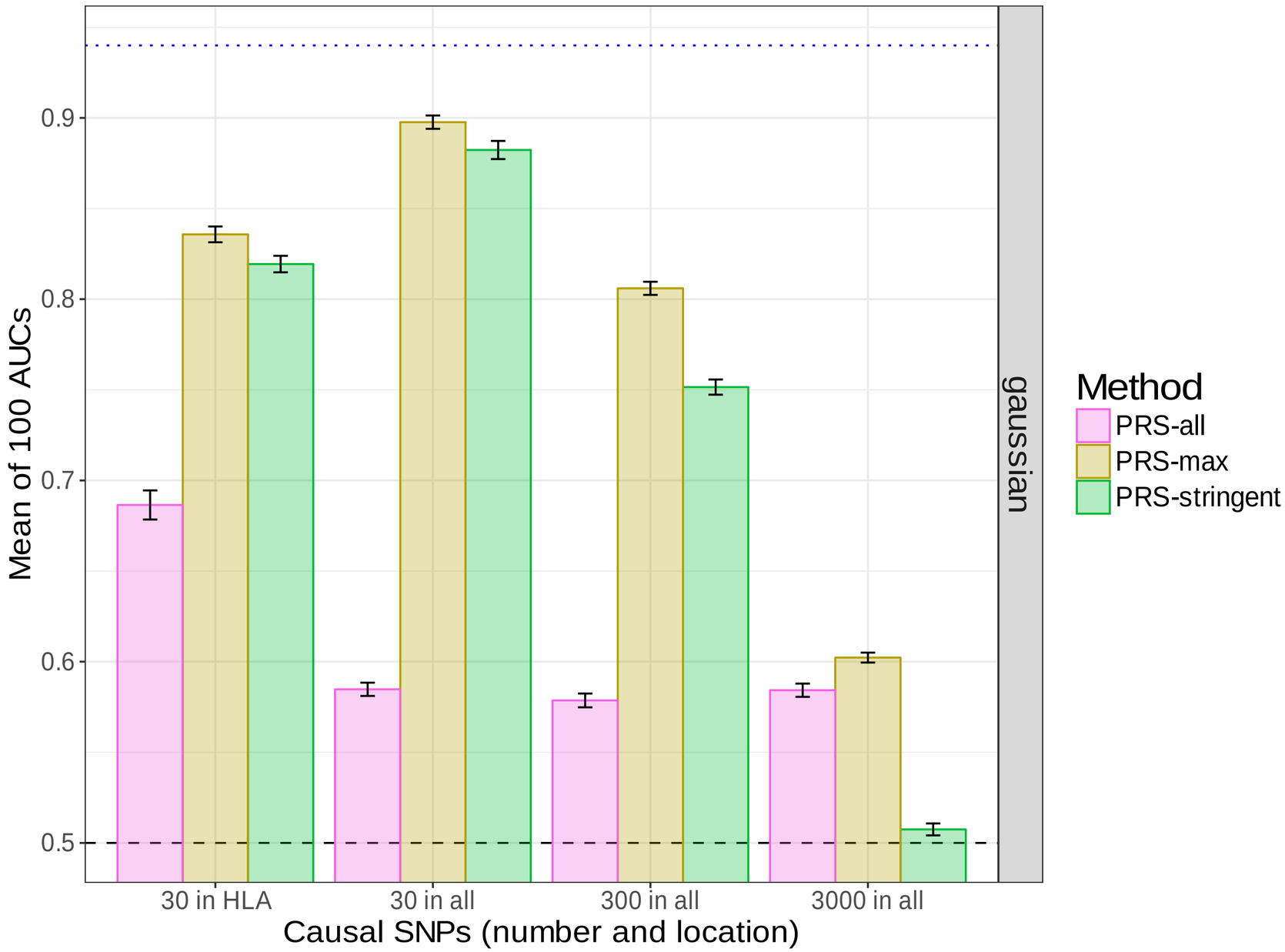
As a second measure, the partial AUC for specificities between 90% and 100% is also reported.

Results

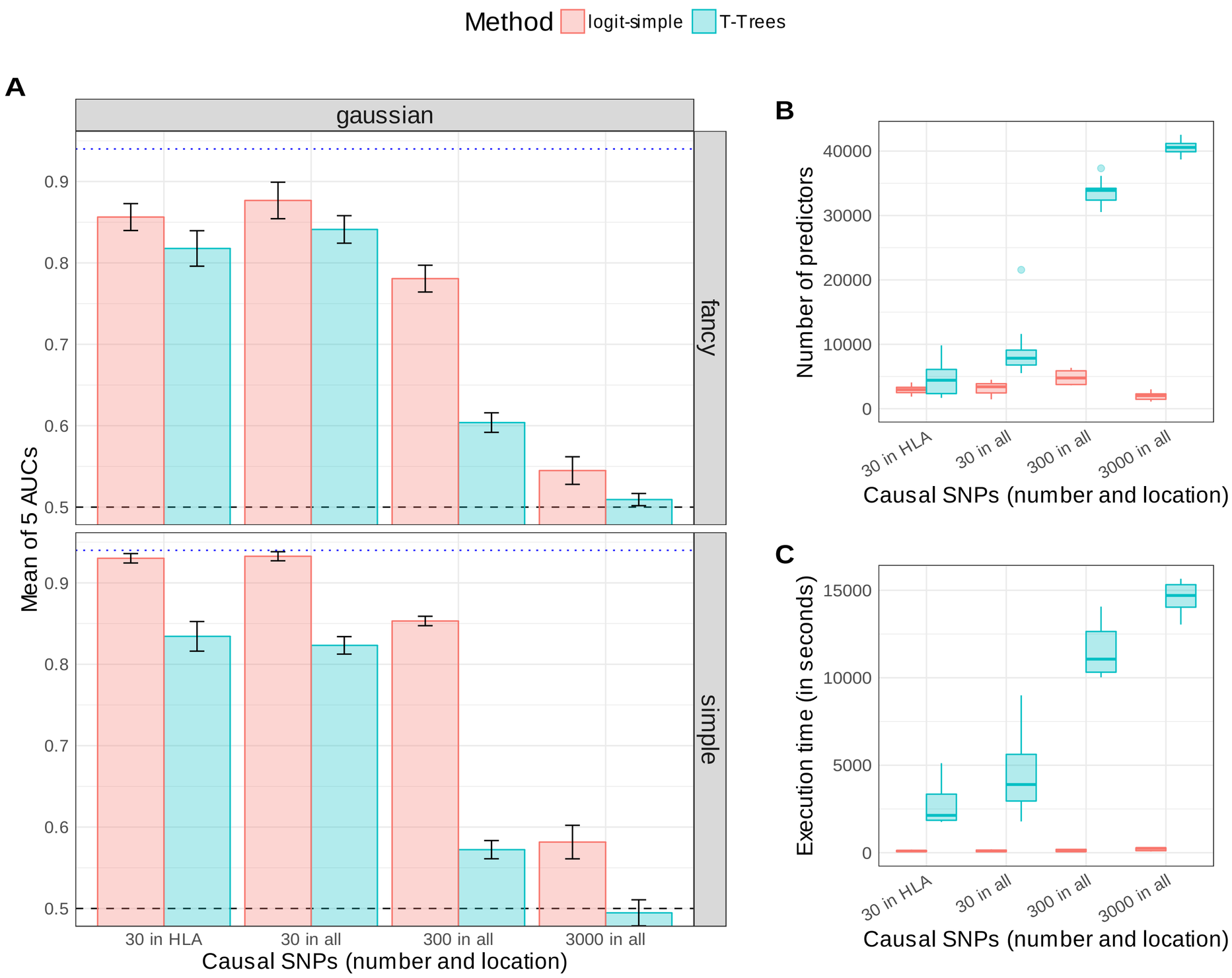
Higher predictive performance with logit-simple



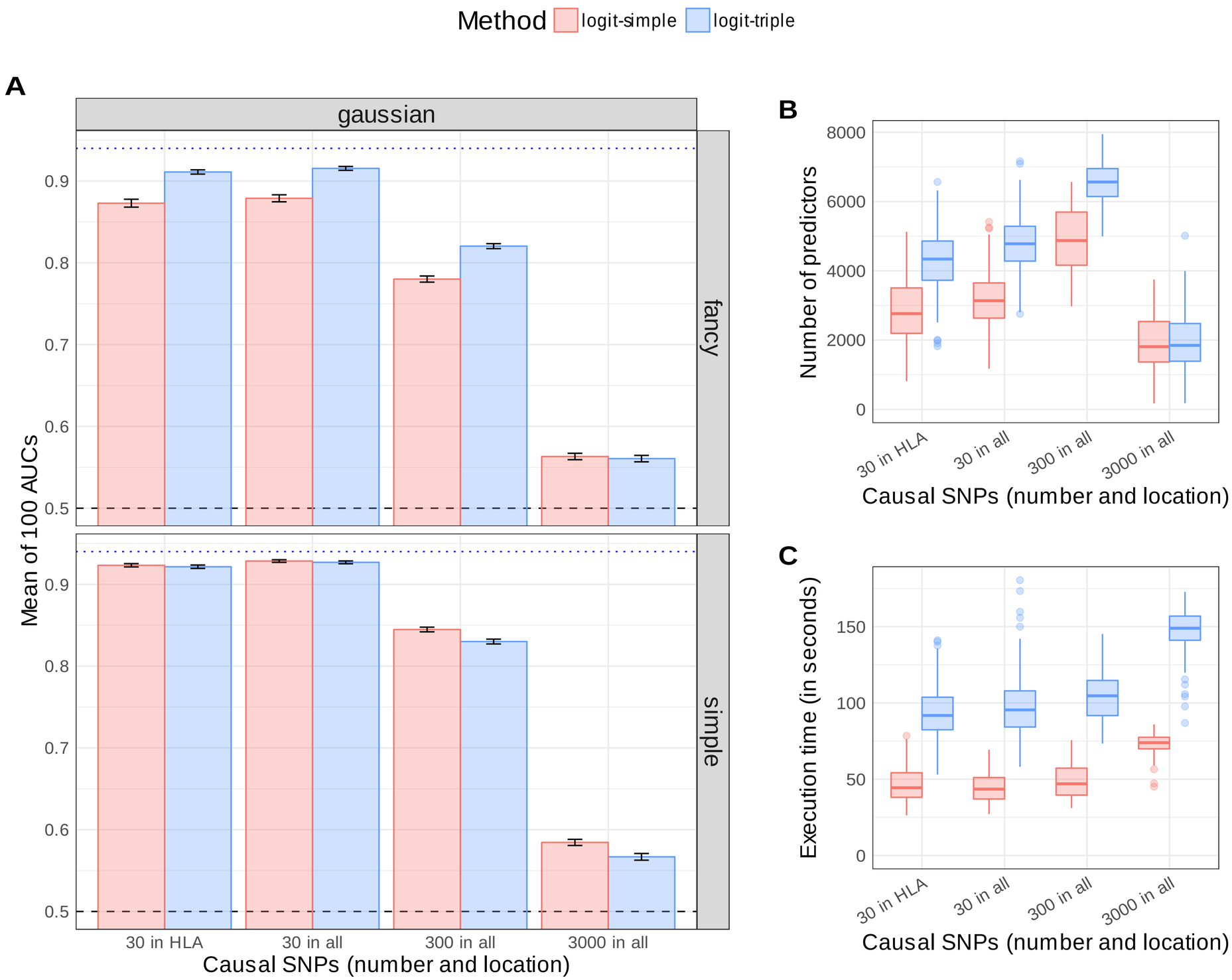
Predictive performance of C+T method varies with threshold



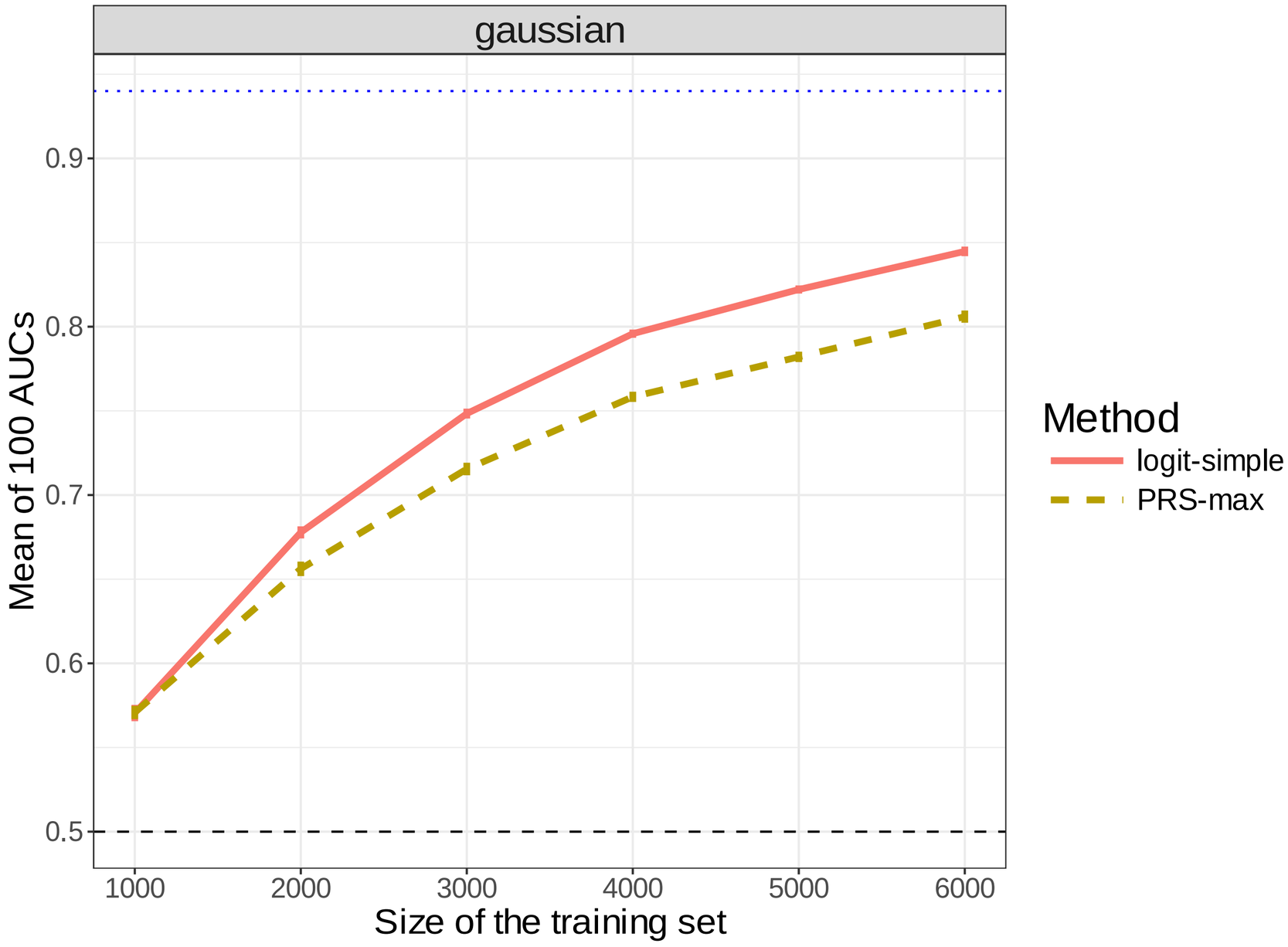
T-Trees, not performant enough



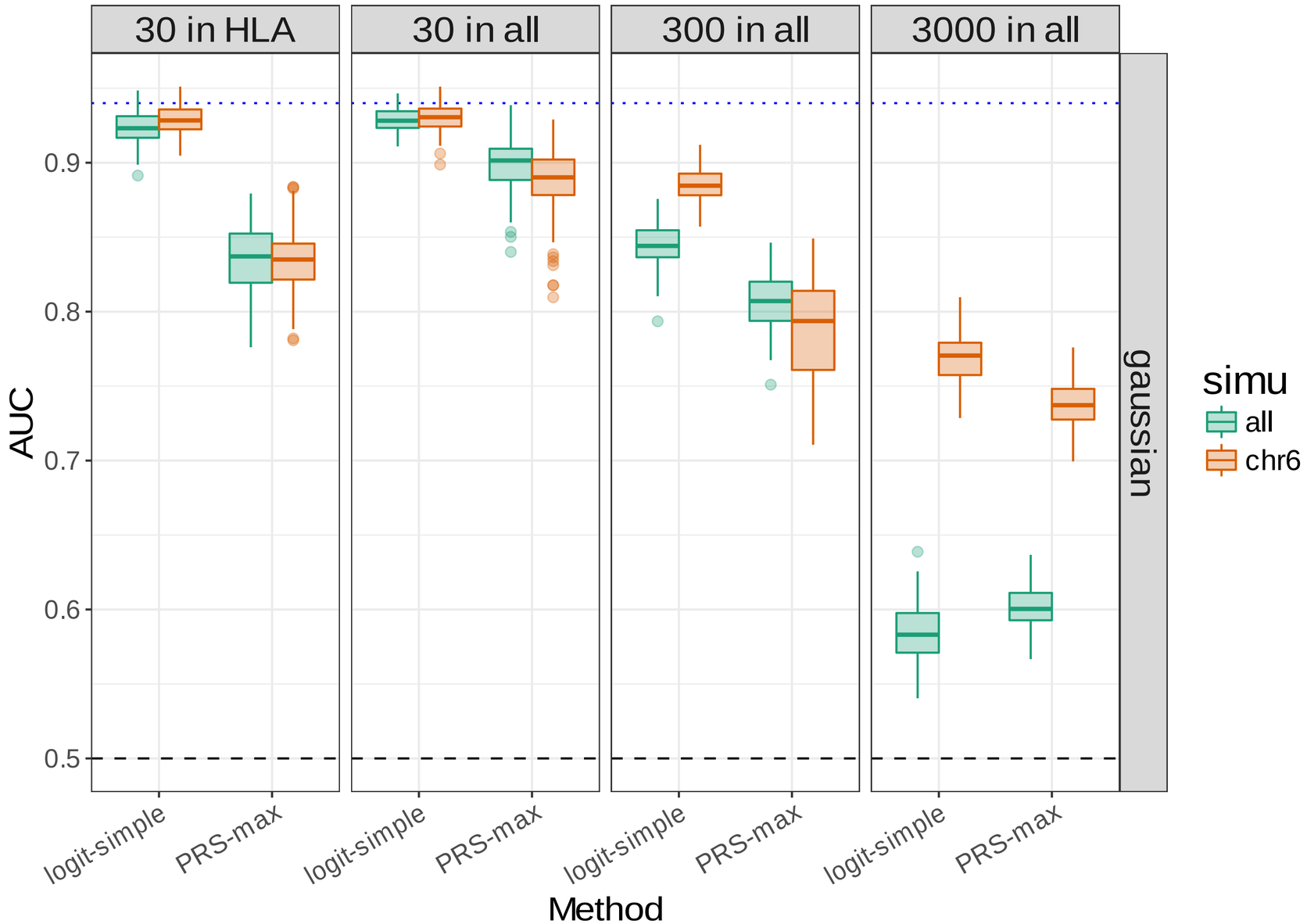
Feature engineering improves prediction



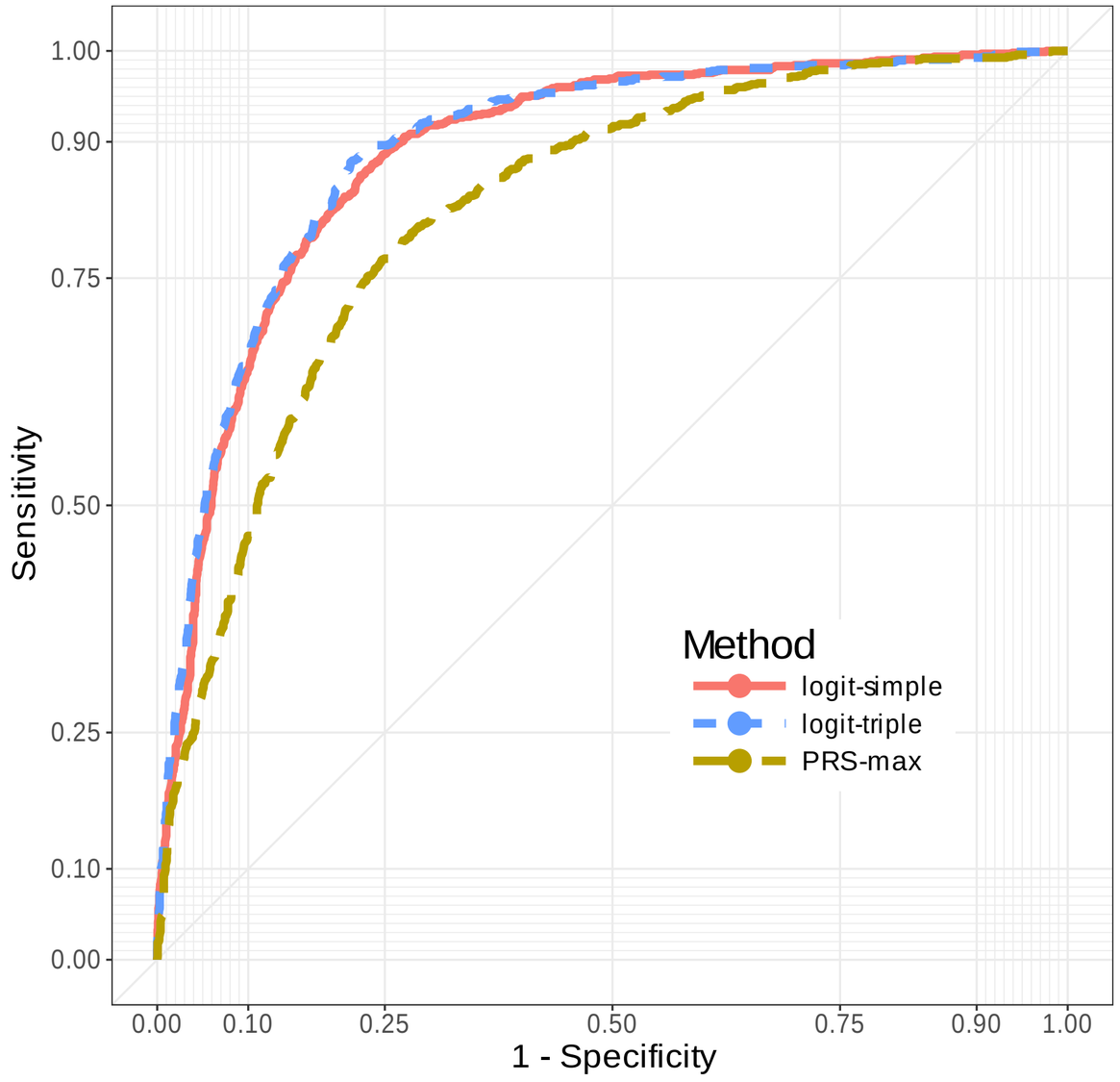
Prediction with logit-simple is improving faster (1/2)



Prediction with logit-simple is improving faster (2/2)



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)
logit-triple	0.892 (0.000488)	0.0429 (0.000174)	4470 (80.6)	141 (1.85)

Discussion

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

- A more optimal approach for predicting complex diseases
- linear solution and really sparse
- even faster
- no need to choose the regularization parameter
- can be extended to capture also recessive and dominant effects

Prospects: future work

- use of summary statistics
- generalization on external population
- integration of clinical and environmental data

Future work: UK Biobank

UK Biobank 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/paper2.html

privefl privefl F. Privé

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