Improve Health Equality for Polygenic Risk Score (PRS) by Joint Penalized Regression of GWAS Summary Statistics from Two Ancestries

Peng Liu

March 29, 2022

Genetic Disorders

- Mendelian disorders are caused by specific mutations in single gene. e.g. cystic fibrosis
- Complex genetic disorders result from combined effect of tens to hundreds of gene mutations each with a small effect.

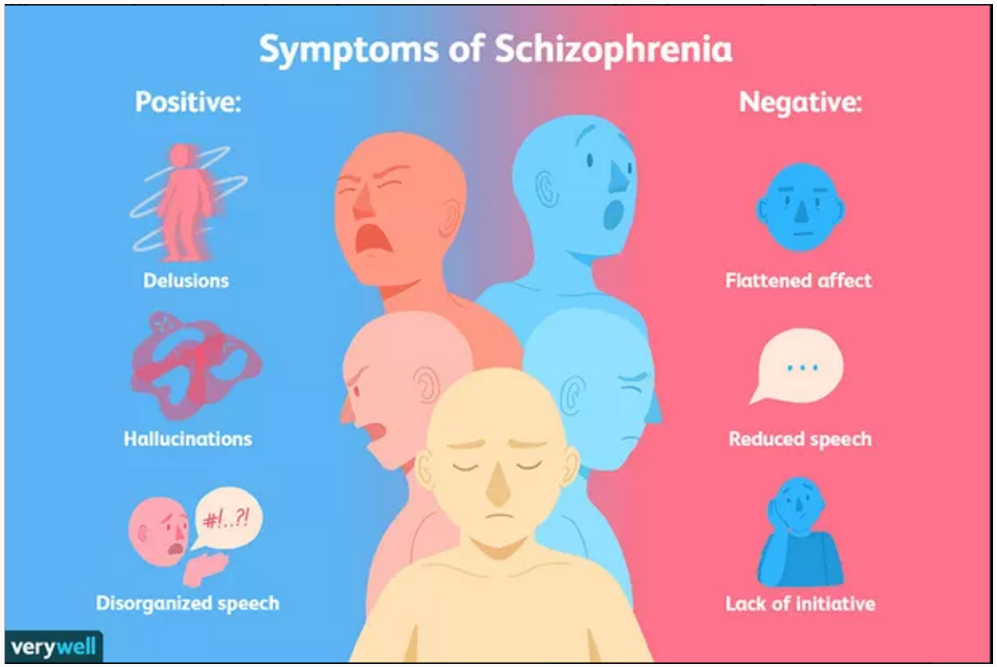
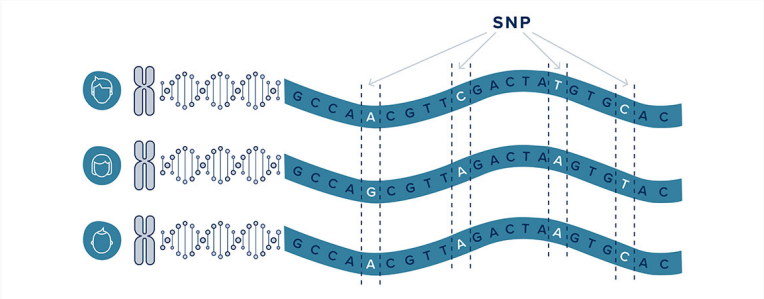
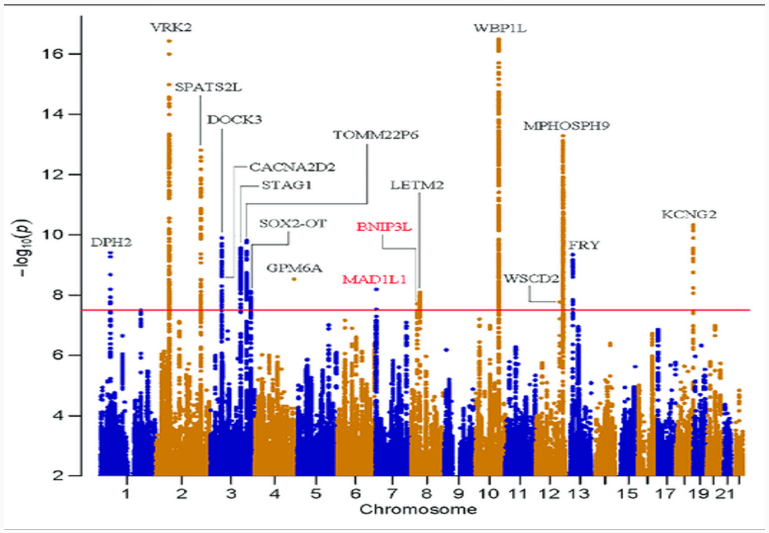


Figure 1: Schizophrenia

Genome-Wide Association Studies (GWAS)



(a) Single Nucleotide Polymorphism (SNP)



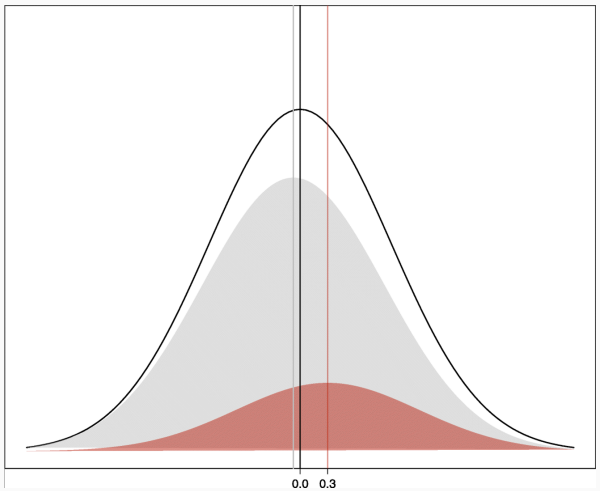
(b) Manhattan plot for genome-wide association studies (GWAS)

- The genetic risk is most often assessed through the polygenic risk score (PRS), a weighted sum of a number of risk alleles.

Weighting up the Risks

A polygenic score (PGS) or polygenic risk score (PRS) is a weighted sum of the risk alleles

$$PRS_i = \sum_j \beta_j x_{ij} \tag{1}$$



(a) Distribution of PRS

SNP (j)

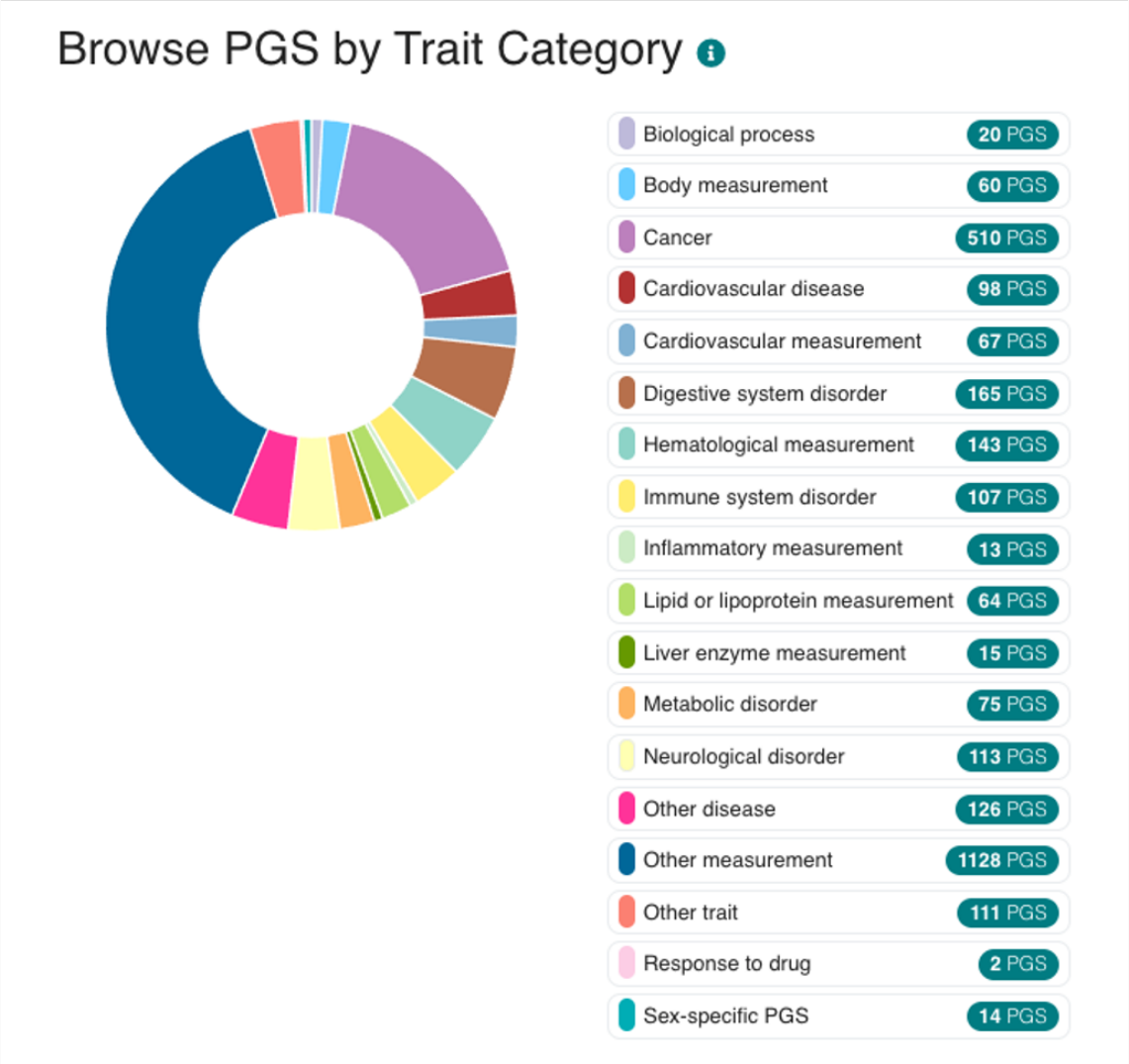
Subject (i)	Genotype	1	2	3	4
	1	0	1	0	1
	2	0	0	2	1
	3	1	2	0	0

(b) Genotype matrix

Lewis & Vassos (2020) Genome Med.

Polygenic Risk Score (PRS)

In the current PGS Catalog, 2163 polygenic scores over 527 traits have been developed. (This is only part of the whole picture!)



Polygenic Risk Score (PRS)

The scores were shown to have potential for broad-scale clinical use

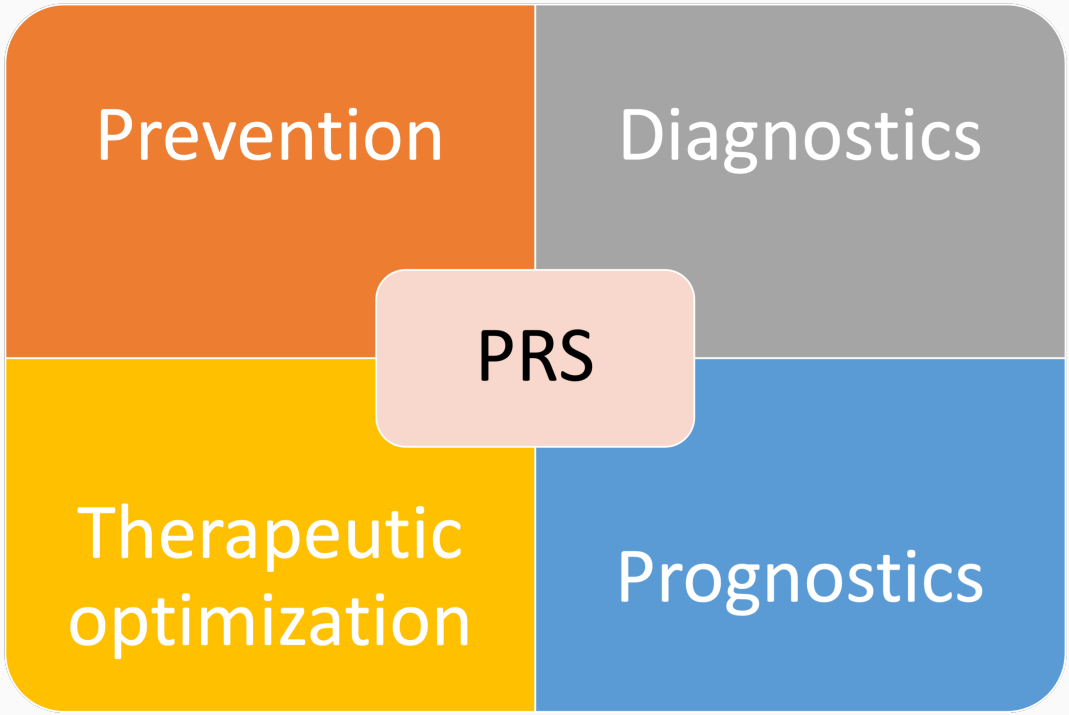
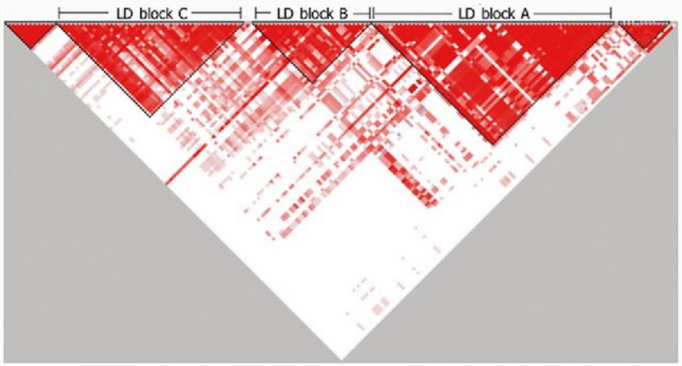


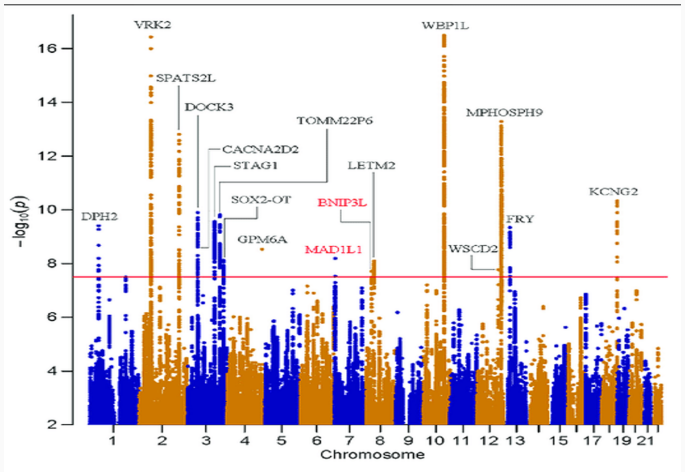
Figure 5: Clinical translation

Key Issues for PRS Calculation

- Selection of SNPs to be included in PRS derivation.
- Estimation of βs
- Incorporating the correlation of SNPs called Linkage disequilibrium (LD).



(a) LD blocks of SNPs



(b) GWAS results

Statistical Methods for PRS Derivation

- Pruning and thresholding (P+T)
- Frequentest approaches: lassosum
- Bayesian approaches: LDpred2, SbayesR

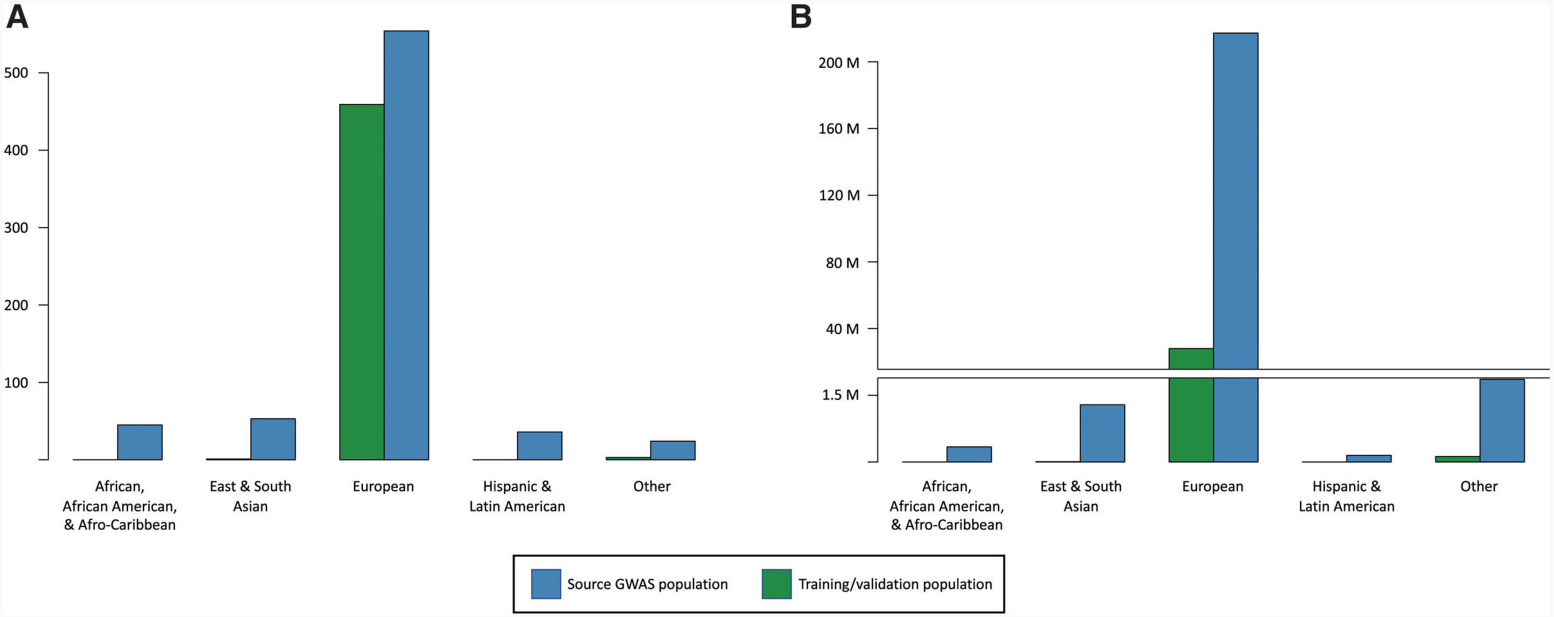
Motivation

Heath disparities of PRS

- The scores are largely been calculated from European population
- Clinical uses of PRS today would systematically afford greater improvement for European-descent populations.

Health Disparities of PRS

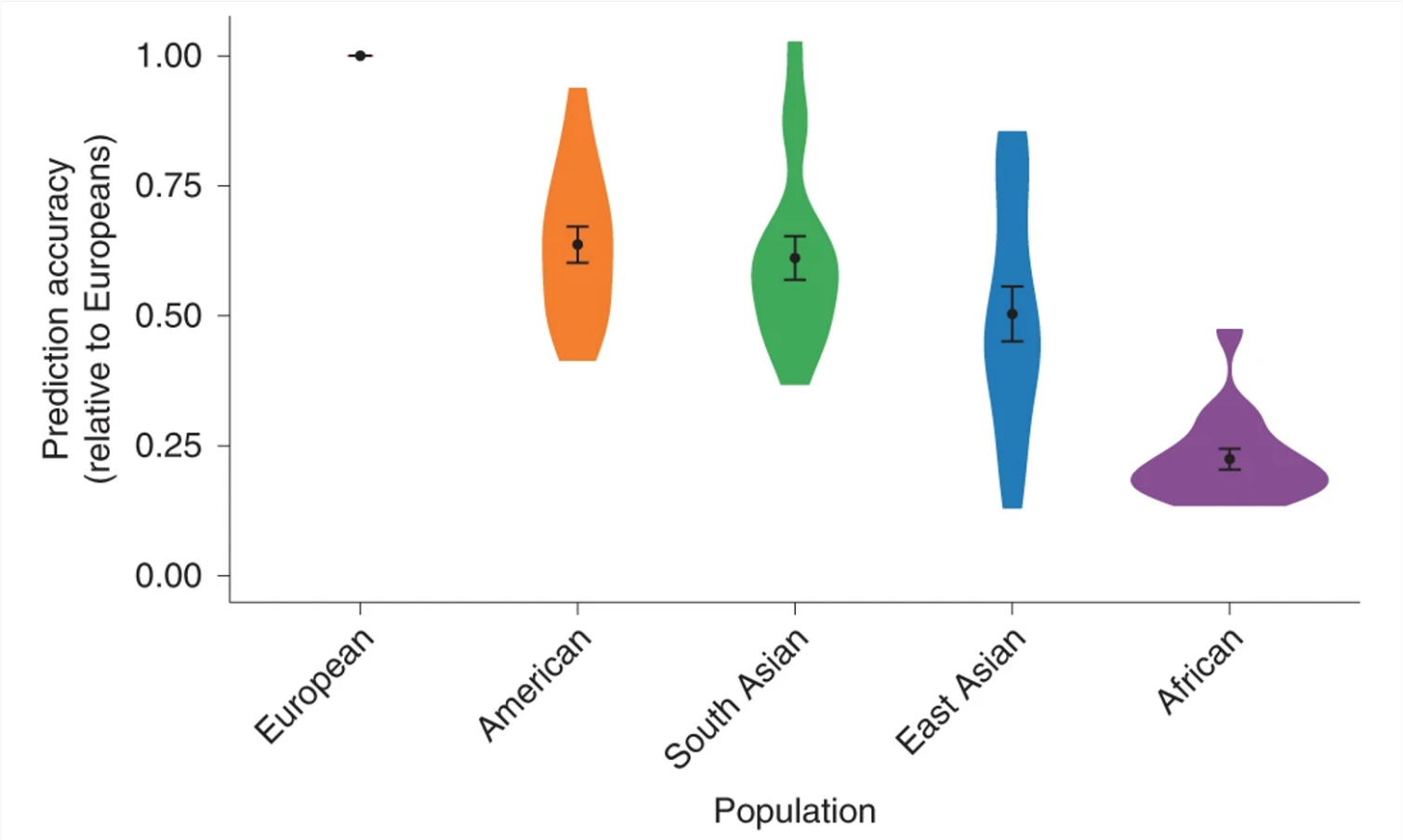
The scores are largely been calculated from European population.



Clarke et al. (2021) CIRC-GENOM

Challenges in PRS Portability

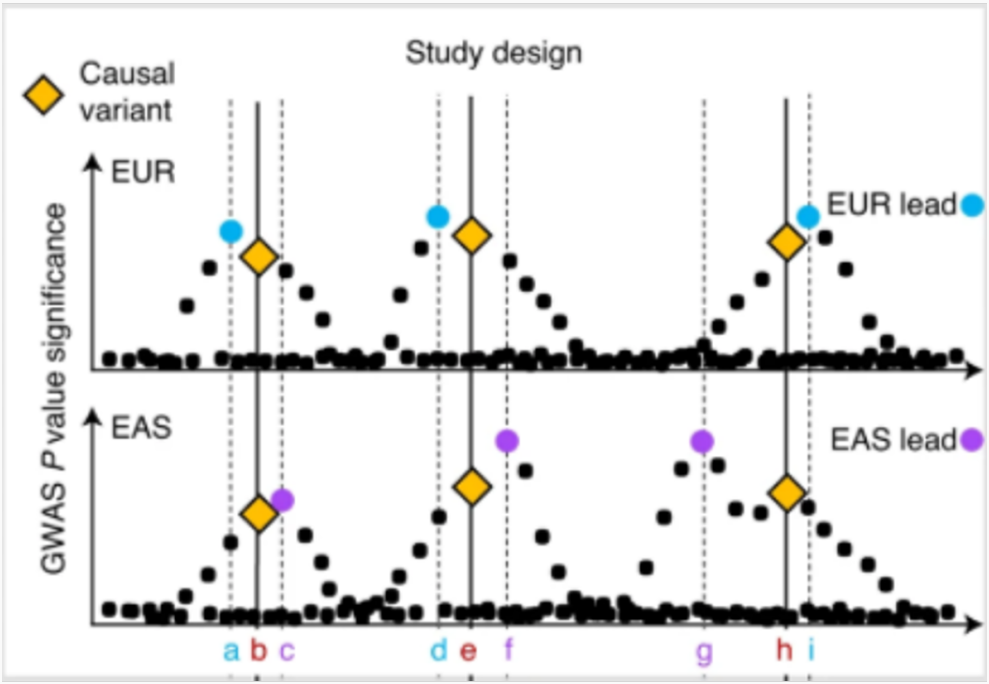
Clinical use of PRS could exacerbate race-based health disparities



Martin et al. (2019) Nature Genetics

Possible Explanations

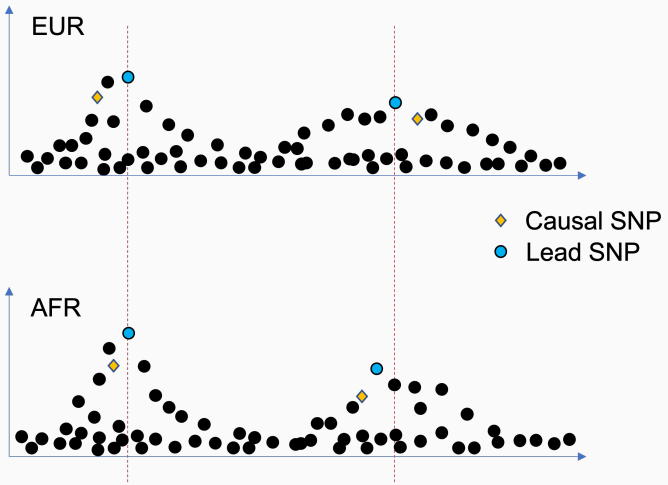
- The causal SNPs of the phenotype might differ, or the risk alleles may have different effect on disease risks across ancestries
- The methods do not select causal SNPs, but the SNPs in LD with causal variant.



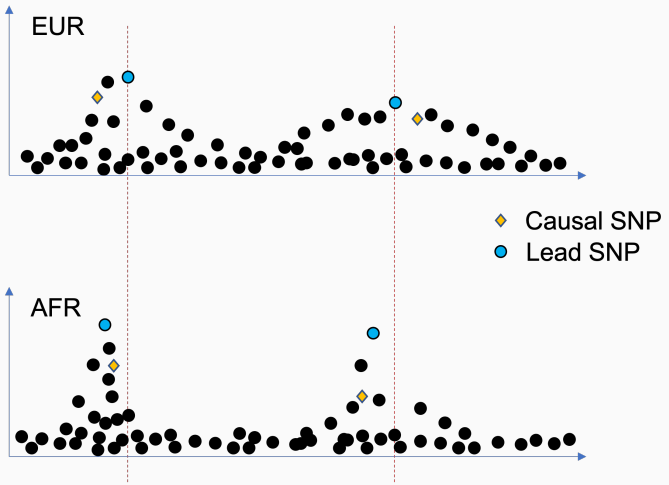
Amariuta et al. (2020) Nature Genetics

Sharpening the Scores

- PRS can work whether or not the causal SNPs are the same (if LD blocks are similar)
- PRS does not work if the LD blocks of the training data are much longer than those for the target population.
- Attenuate disparities by incorporating the genetic information from less represented population.



(a) Similar LD blocks



(b) Different LD blocks

The Available Data

- The available training data is in summary statistics. The correlation ($\boldsymbol{r}$) could be computed from the summary statistics.

CHROME	POS	ID	REF	ALT	OR	SE	P
1	17407	rs1570391677	G	A	1.034	0.034	0.337
1	54421	rs1570391629	A	G	1.012	0.034	0.826
1	59615	rs1639538192	C	T	1.051	0.034	0.365
1	108030	rs1639538207	G	T	0.975	0.034	0.507

- The LD blocks ($\boldsymbol{R}$) can be obtained from available full genotype libraries, such as 1000 Genome Project.

Lassosum: a penalized regression (LASSO) based method (Mak et al., 2017)

lassosum is the building block of the proposed method. It minimize the following objective function:

$$f(\beta) = \boldsymbol{y}^T\boldsymbol{y} + \beta^T\boldsymbol{R}\beta - 2\beta^T\boldsymbol{r} + 2\lambda\|\beta\|_1^1$$

- λ is the tuning parameter for LASSO penalty.
- $\boldsymbol{r}$ is a vector of correlations between SNPs and the phenotype, we only need summary statistics to get $\boldsymbol{r}$.
- $\boldsymbol{R} = \boldsymbol{X}^T\boldsymbol{X}$ is the correlations (LD) matrix.
- The data used to derive $\boldsymbol{R}$ and $\boldsymbol{r}$ are usually not the same.

The Proposed Method

$$
\begin{aligned}
f(\beta) =& \gamma \left(\boldsymbol{y}^T \boldsymbol{y} + (1-s)\beta^T \boldsymbol{R}\beta - 2\beta^T \boldsymbol{r} \right) \\
&+ (1-\gamma) \left(\boldsymbol{y'}^T \boldsymbol{y'} + (1-s)\beta^T \boldsymbol{R'}\beta - 2\beta^T \boldsymbol{r'} \right) \\
&+ s\beta^T\beta + 2\sum_j \lambda|\beta_j|
\end{aligned}
$$

The Proposed Method

$$
\begin{aligned}
f(\beta) =& \gamma \left(\boldsymbol{y}^T \boldsymbol{y} + (1-s)\beta^T \boldsymbol{R}\beta - 2\beta^T \boldsymbol{r} \right) \\
&+ (1-\gamma) \left(\boldsymbol{y'}^T \boldsymbol{y'} + (1-s)\beta^T \boldsymbol{R'}\beta - 2\beta^T \boldsymbol{r'} \right) \\
&+ s\beta^T\beta + 2\sum_j \lambda|\beta_j|
\end{aligned}
$$

- γ controls the balance between the two populations, λ is the penalty tuning parameter and s is a shrinkage parameter for ridge penalty.

The Proposed Method

$$
\begin{aligned}
f(\beta) =& \gamma \left(\boldsymbol{y}^T \boldsymbol{y} + (1-s)\beta^T \boldsymbol{R}\beta - 2\beta^T \boldsymbol{r} \right) \\
&+ (1-\gamma) \left(\boldsymbol{y'}^T \boldsymbol{y'} + (1-s)\beta^T \boldsymbol{R'}\beta - 2\beta^T \boldsymbol{r'} \right) \\
&+ s\beta^T\beta + 2\sum_j \lambda|\beta_j|
\end{aligned}
$$

- γ controls the balance between the two populations, λ is the penalty tuning parameter and s is a shrinkage parameter for ridge penalty.
- The method calls for GWAS summary statistics ($\boldsymbol{r}$ and $\boldsymbol{r}^{'}$) and LD ($\boldsymbol{R}$ and $\boldsymbol{R}^{'}$) from two populations, with one population much bigger than the other.

The Proposed Method

$$
\begin{aligned}
f(\beta) =& \gamma \left(\boldsymbol{y}^T \boldsymbol{y} + (1-s)\beta^T \boldsymbol{R}\beta - 2\beta^T \boldsymbol{r} \right) \\
&+ (1-\gamma) \left(\boldsymbol{y'}^T \boldsymbol{y'} + (1-s)\beta^T \boldsymbol{R'}\beta - 2\beta^T \boldsymbol{r'} \right) \\
&+ s\beta^T\beta + 2\sum_j \lambda|\beta_j|
\end{aligned}
$$

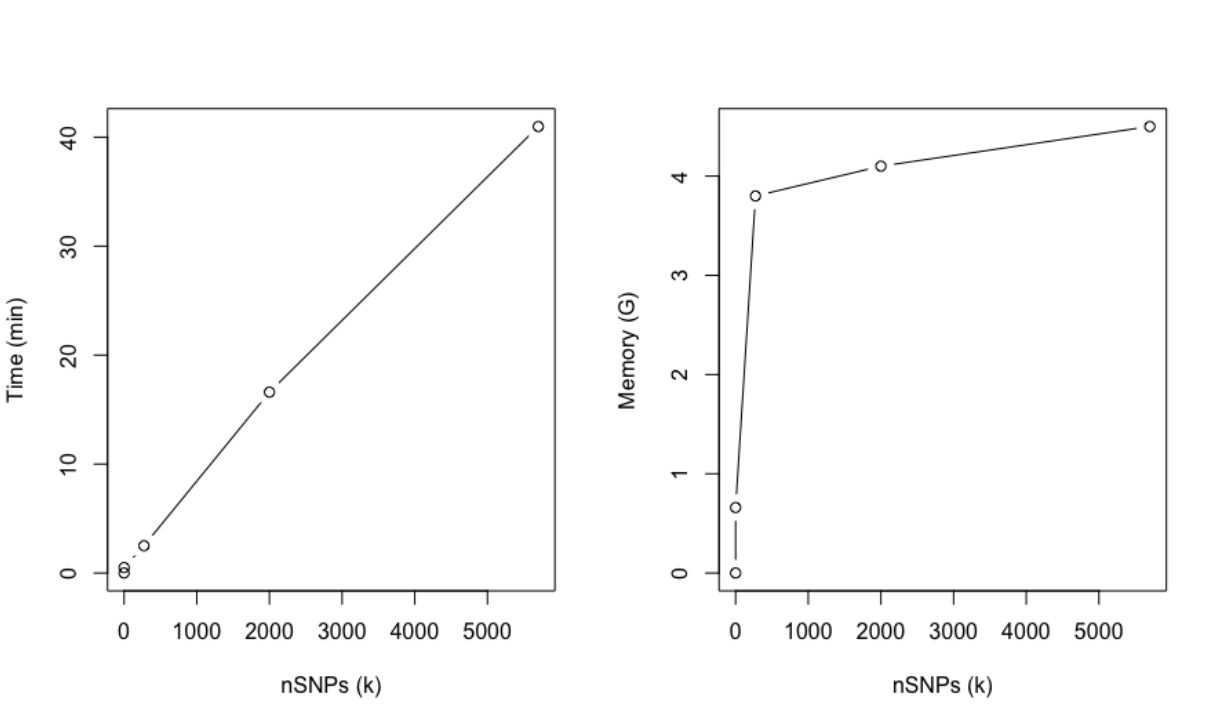
- γ controls the balance between the two populations, λ is the penalty tuning parameter and s is a shrinkage parameter for ridge penalty.
- The method calls for GWAS summary statistics ($\boldsymbol{r}$ and $\boldsymbol{r}^{'}$) and LD ($\boldsymbol{R}$ and $\boldsymbol{R}^{'}$) from two populations, with one population much bigger than the other.
- Leverage the genomics information from the bigger population, and inflate genomics information from the smaller population by γ

The Proposed Method

$$
\begin{aligned}
f(\beta) =& \gamma \left(\boldsymbol{y}^T \boldsymbol{y} + (1-s)\beta^T \boldsymbol{R}\beta - 2\beta^T \boldsymbol{r} \right) \\
&+ (1-\gamma) \left(\boldsymbol{y'}^T \boldsymbol{y'} + (1-s)\beta^T \boldsymbol{R'}\beta - 2\beta^T \boldsymbol{r'} \right) \\
&+ s\beta^T\beta + 2\sum_j \lambda|\beta_j|
\end{aligned}
$$

- γ controls the balance between the two populations, λ is the penalty tuning parameter and s is a shrinkage parameter for ridge penalty.
- The method calls for GWAS summary statistics ($\boldsymbol{r}$ and $\boldsymbol{r}^{'}$) and LD ($\boldsymbol{R}$ and $\boldsymbol{R}^{'}$) from two populations, with one population much bigger than the other.
- Leverage the genomics information from the bigger population, and inflate genomics information from the smaller population by γ
- The objective function can be solved by coordinate descent algorithm.

Computation Efficiency



- I/O and matrix algebra functions are programmed by Rcpp.
- Set a input data cap (default is 500Mb) to save memory.
- The program can extract SNPs, their neighbours (within a specified distance in kb) and samples specified by user.

Tuning Parameters

There are three tuning parameters (1) penalty parameter λ, (2) weighting parameter γ, and (3) shrinkage parameter s.

We would like to discuss model tuning in the following two scenarios

- No independent tuning data is available (worst case).
- Summary statistics for an independent tuning data are available.
    - Often the published results from a GWAS do not include individual level information. Instead, summary statistics are published.

When Only Summary Statistics are Available

We would like to minimize the distance between the predicted and true phenotype in testing data

$$\ell = ||\boldsymbol{y} - \hat{\boldsymbol{y}}||_2^2$$

where $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ are both normalized to unit norm.

When Only Summary Statistics are Available

We would like to minimize the distance between the predicted and true phenotype in testing data

$$\ell = ||\boldsymbol{y} - \hat{\boldsymbol{y}}||_2^2$$

where $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ are both normalized to unit norm.

After several steps of derivation, it can be approximated by

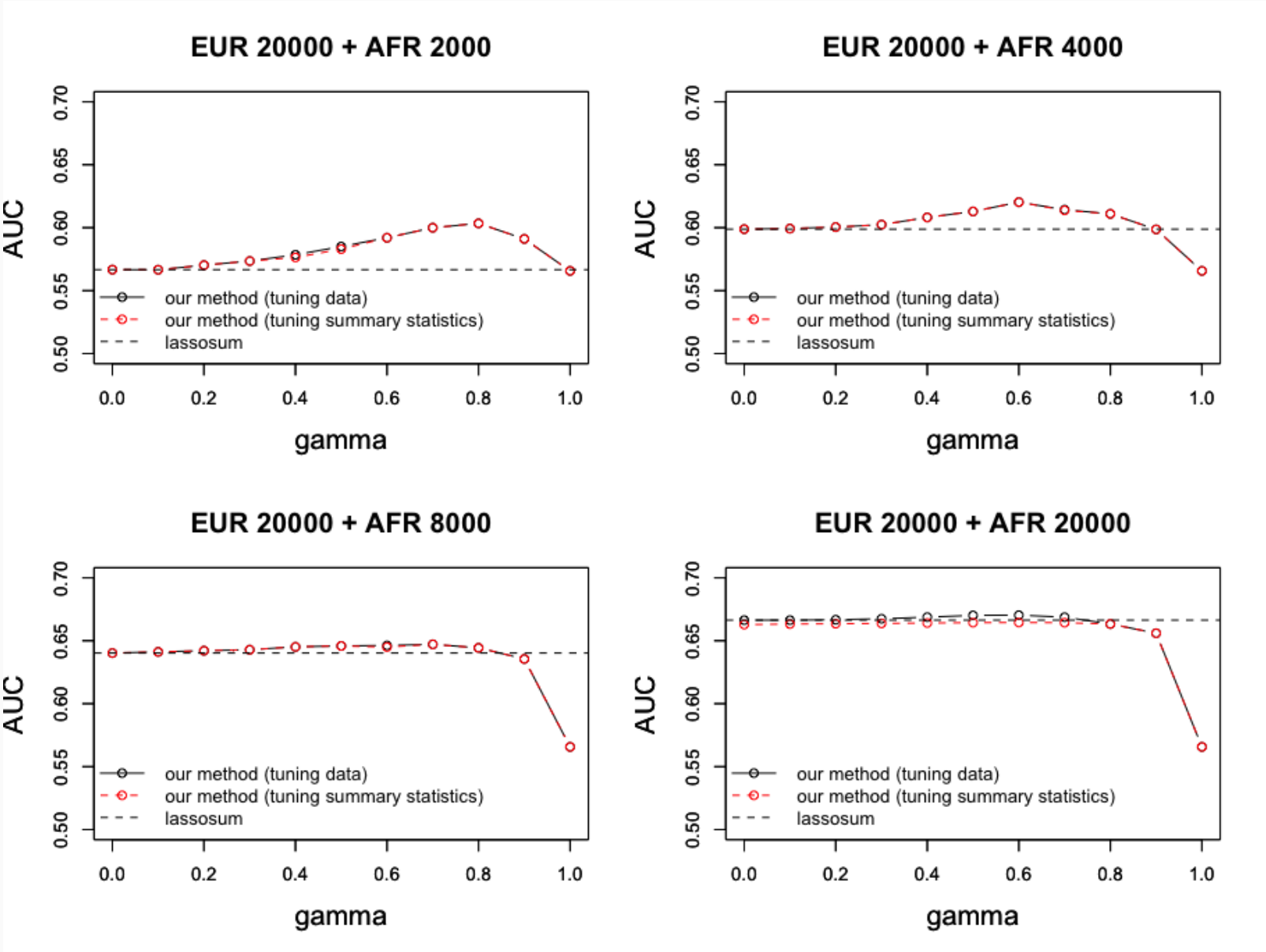
$$\ell = 1 - \frac{\hat{\boldsymbol{\beta}}^T \boldsymbol{r}}{\sqrt{\hat{\boldsymbol{\beta}}(\boldsymbol{R}/n)\hat{\boldsymbol{\beta}}}}$$

where $\boldsymbol{r}$ can be calculated from GWAS summary statistics, and $\boldsymbol{R}$ is the LD matrix.

Simulation Study I

- GWAS summary statistics: 20,000 Europeans (EUR), plus 2,000, 4,000 , 8,000 or 20,000 Africans (AFR)
- LD information: 2,000 EUR, 2,000 AFR
- Tuning data: 4,000 AFR
- Testing data: 4,000 AFR
- 5.7 million SNPs, 300 causal SNPs

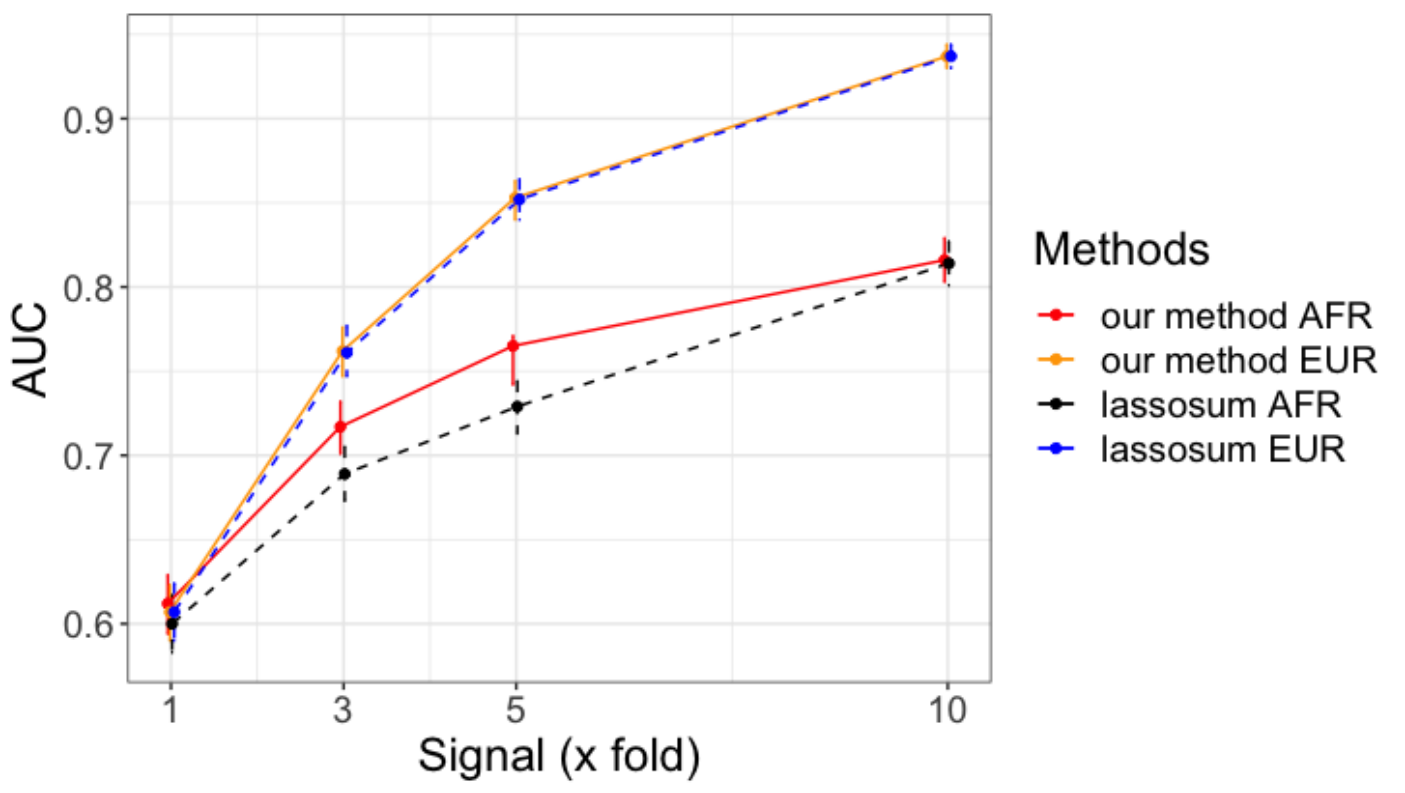
Parameter Tuning



Simulation Study II

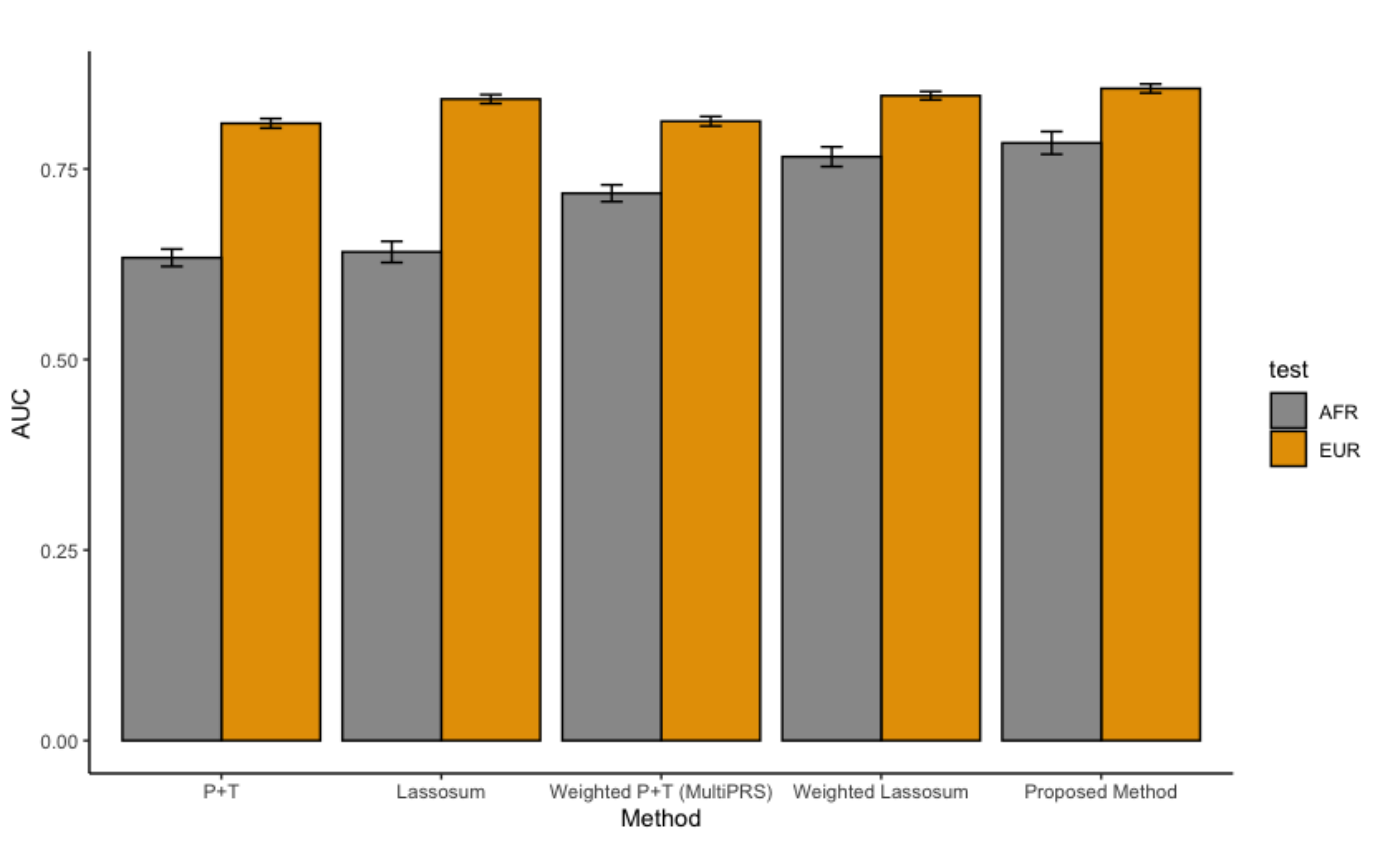
- Whole genome: 5.7 million SNPs, 4000 causal SNPs
- Simulated datasets for Europeans (EUR) and Africans (AFR)
  - GWAS summary statistics: 20,000 EUR, 4,000 AFR
  - A separate data for LD information
  - Tuning data
  - Testing data

Prediction: The Proposed Method vs Lassosum



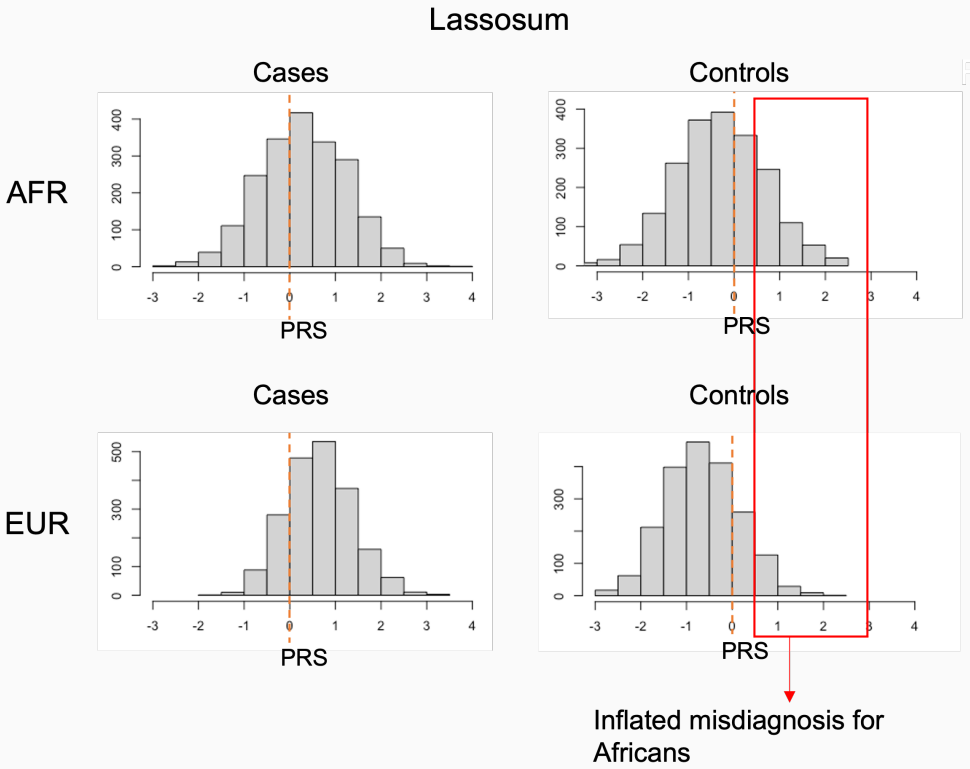
- The proposed method improves performance for both AFR and EUR testing data.
- The improvement depends on the level of causal SNP effects.

Prediction: Comparing Multiple Methods

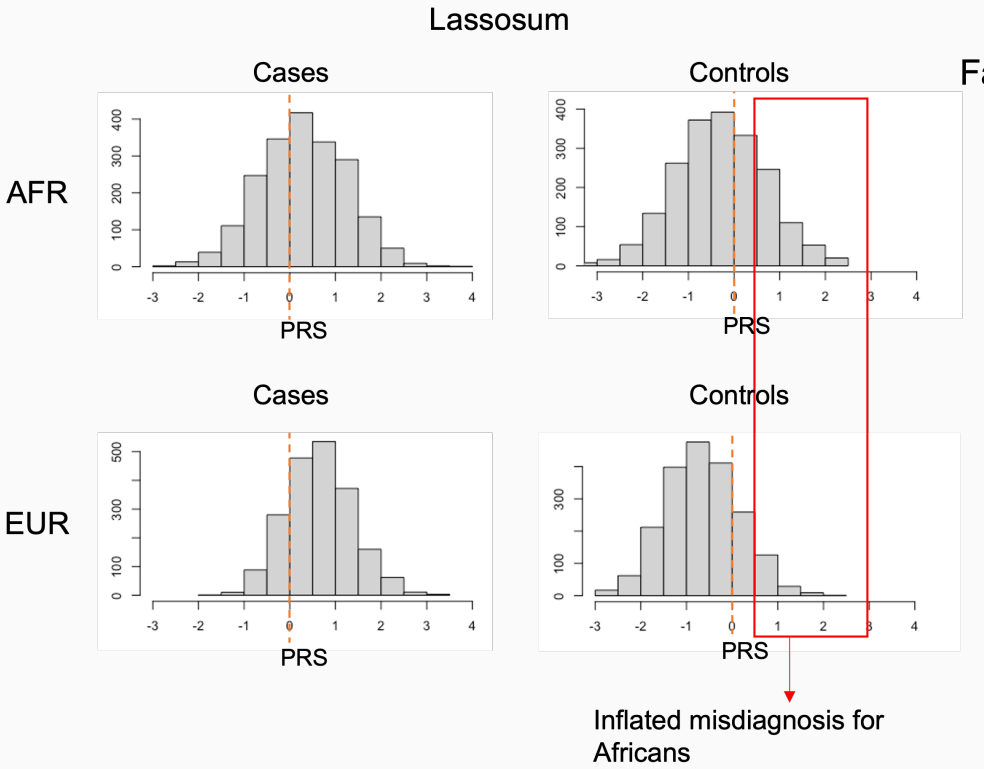


- MultiPRS (weighted P+T):
  $PRS_{combined} = \alpha PRS_{EUR} + (1 - \alpha) PRS_{AFR}$
- Weighted lassosum: same idea can be extended to lassosum

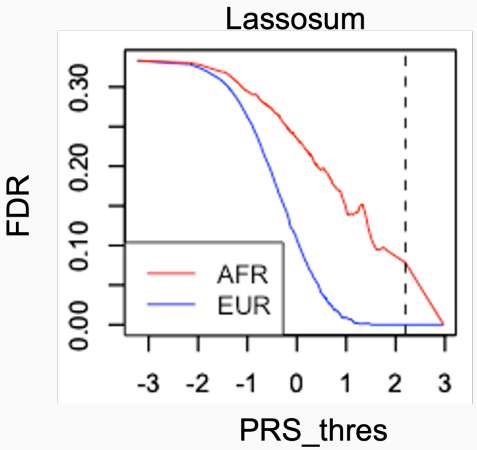
Disparities in Lassosum



Disparities in Lassosum



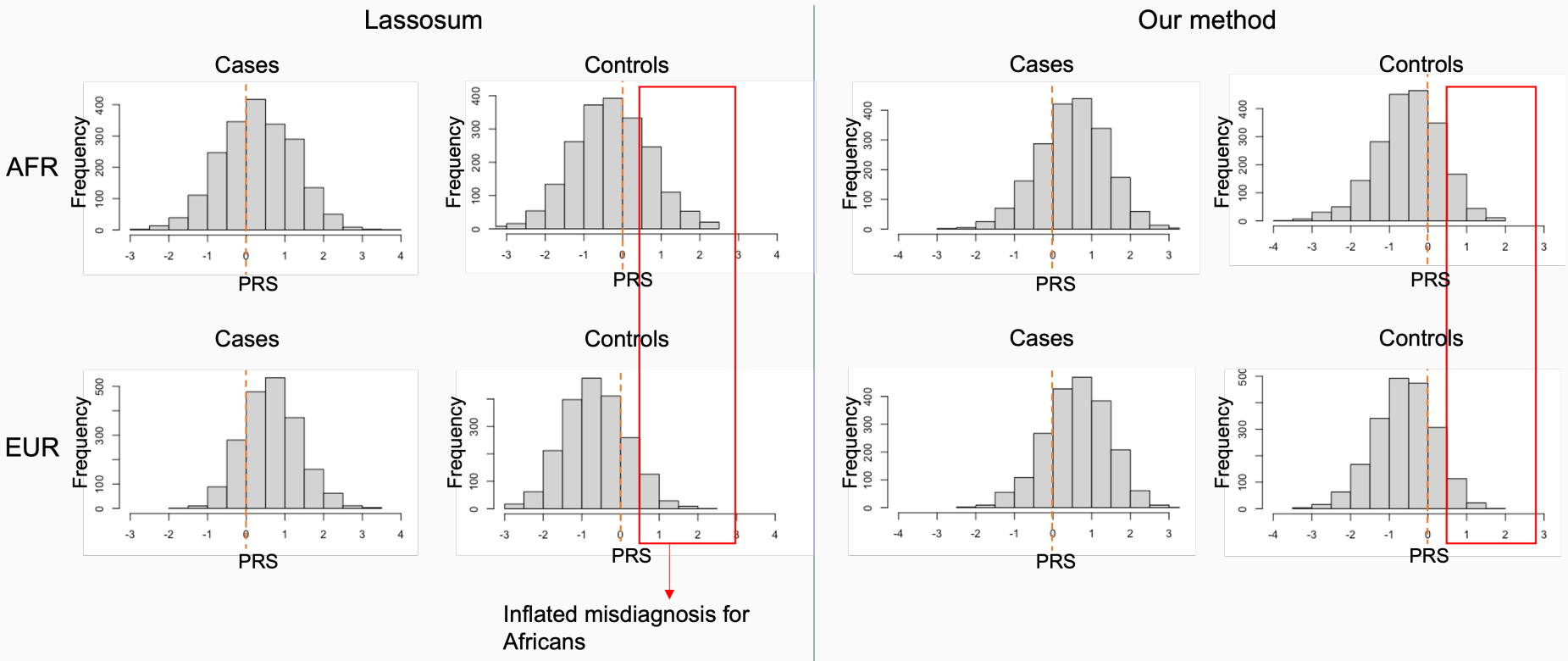
$$\text{False discovery rate (FDR)} = \frac{\#\text{ false positives}}{\text{total }\#\text{ positive predictions}}$$



Racial Disparities in Schizophrenia Diagnosis

-
- African-Americans are more likely to be misdiagnosed as having schizophrenia, which has been reported by numerous studies.
    - In a 2018 analysis of data from 52 different studies, Olbert et al. (2018) found that African-Americans are 2.4 times more likely to be diagnosed with schizophrenia.
    - Schwartz and Blankenship (2014) claim it's 3-4 times higher
    - Improved PRS might help with diagnosis bias.

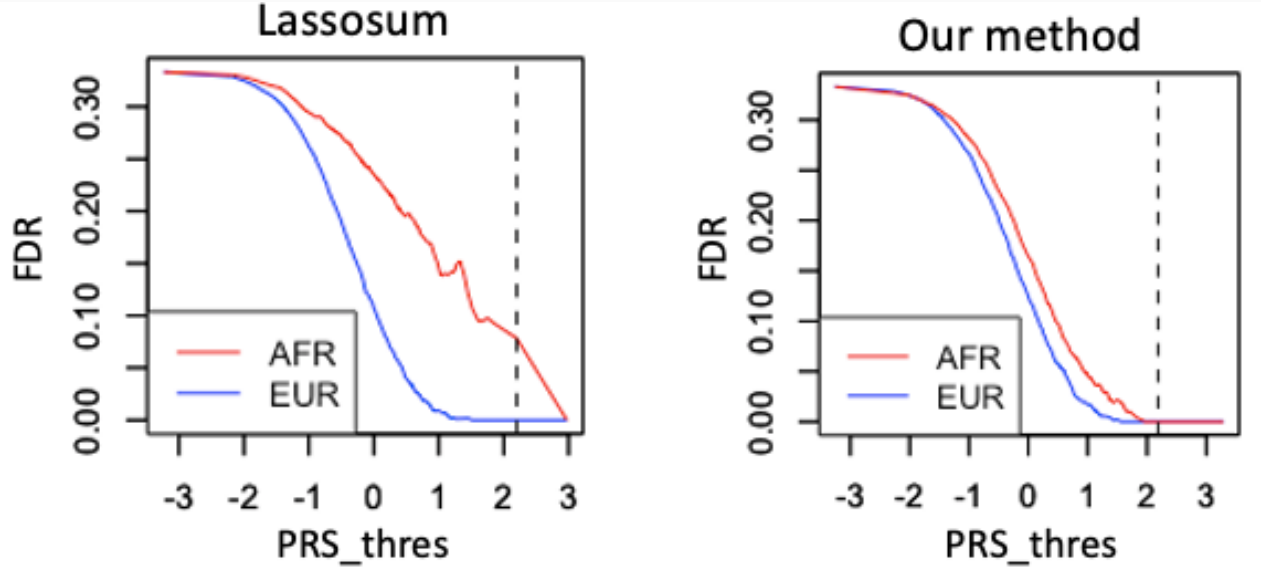
The Proposed Method Mitigates the Disparities



The proposed method mitigates the disparities in the number of false positives by lassosum in simulation studies.

The Proposed Method Mitigates the Disparities

False discovery rate (FDR): $\text{FDR} = \frac{\#\text{ false positives}}{\#\text{ positive predictions}}$



Summary

- We have developed a novel statistical method for PRS by incorporating genetic information from two ancestries.
- The proposed method jointly estimates the effect size and choose the SNPs that are predictive for both ancestries.
- The proposed method improves trans-ancestry portability of PRS.
- The proposed method could mitigate the disparities in the diagnosis of schizophrenia.

Future Direction

- Apply our method to admixed population.
- Can we borrow information from Autism to predict the patients with schizophrenia?
- How to integrate genetic and clinical risk?

Acknowledgement

- Kathryn
- Bernie
- Max
- Bert

References

Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X., and Sham, P. C. (2017). Polygenic scores via penalized regression on summary statistics. *Genetic epidemiology*, 41(6):469–480.

Olbert, C. M., Nagendra, A., and Buck, B. (2018). Meta-analysis of black vs. white racial disparity in schizophrenia diagnosis in the united states: Do structured assessments attenuate racial disparities? *Journal of Abnormal Psychology*, 127(1):104.

Schwartz, R. C. and Blankenship, D. M. (2014). Racial disparities in psychotic disorder diagnosis: A review of empirical literature. *World journal of Psychiatry*, 4(4):133.

Lassosum: a penalized regression (LASSO) based method (Mak et al., 2017)

Proof:
$\boldsymbol{R}_s = (1-s)\boldsymbol{X}_r^T\boldsymbol{X}_r + s\boldsymbol{I}$ is positive definite, and there is always exit $\boldsymbol{W}$ and $\boldsymbol{v}$ such that $\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{R}_s, \quad \boldsymbol{W}^T\boldsymbol{v} = \boldsymbol{r}$.
So the objective function becomes:

$$\begin{aligned} f(\beta) =& \boldsymbol{y}^T\boldsymbol{y} + (1-s)\beta^T\boldsymbol{X}_r^T\boldsymbol{X}_r\beta - 2\beta^T\boldsymbol{r} \\ &+ s\beta^T\beta + 2\lambda\|\beta\|_1^1, \end{aligned}$$

when $0 < s < 1$, the objective function could be solved using the method for elastic net problem by coordinate descent algorithms, avoiding computationally expensive LD matrix $\boldsymbol{X}_r^T\boldsymbol{X}_r$ calculation.

Coordinate Descent Algorithm

- It's reasonable to assume the SNPs from different LD blocks are not correlated. For computational efficiency, we would like to compute the gradient at β_j by its corresponding LD blocks.
- The coodinate-wise update for β_j is

$$\beta_j = \begin{cases} 0 & \text{otherwise} \\ \frac{\text{sign}(u_j)(|u_j|-\lambda)}{\gamma \tilde{\boldsymbol{x}}_j^T \tilde{\boldsymbol{x}}_j + (1-\gamma) \tilde{\boldsymbol{x}'}_j^T \tilde{\boldsymbol{x}'}_j + s} & \text{if } |u_j| > \lambda \end{cases} \tag{2}$$

where

$$u_j = \gamma \left(r_j - \tilde{\boldsymbol{x}}_j^T \tilde{\boldsymbol{X}}_{-j}^{[k]} \tilde{\boldsymbol{\beta}}_{-j}^{[k]} \right) + (1-\gamma) \left(r'_j - \tilde{\boldsymbol{x}'}_j^T \tilde{\boldsymbol{x}'}_{-j}^{[k']} \tilde{\boldsymbol{\beta}}_{-j}^{[k']} \right)$$

Coordinate Descent Algorithm

- It's reasonable to assume the SNPs from different LD blocks are not correlated. For computational efficiency, we would like to compute the gradient at β_j by its corresponding LD blocks.
- The coodinate-wise update for β_j is

$$\beta_j = \begin{cases} 0 & \text{otherwise} \\ \frac{\text{sign}(u_j)(|u_j| - \lambda)}{\gamma \tilde{\boldsymbol{x}}_j^T \tilde{\boldsymbol{x}}_j + (1-\gamma) \tilde{\boldsymbol{x}'}_j^T \tilde{\boldsymbol{x}'}_j + s} & \text{if } |u_j| > \lambda \end{cases} \tag{3}$$

s → shrinkage parameter

where

$$u_j = \gamma \left(r_j - \tilde{\boldsymbol{x}}_j^T \tilde{\boldsymbol{x}}_{-j}^{[k]} \tilde{\beta}_{-j}^{[k]} \right) + (1-\gamma) \left(r'_j - \tilde{\boldsymbol{x}'}_j^T \tilde{\boldsymbol{x}'}_{-j}^{[k']} \tilde{\beta}_{-j}^{[k']} \right)$$

Ancestry 1 Ancestry 2

Simulation Study III

- GWAS summary statistics: 20,000 Europeans (EUR), 4,000 Africans (AFR)
- LD information: 5,000 EUR, 5,000 AFR
- Testing data: ? AFR and ? EUR
- 5.7 million SNPs, 4000 causal SNPs