JOHNS HOPKINS
UNIVERSITY

An ensemble penalized regression method for multi-ancestry polygenic risk prediction

Jingning Zhang[*], Jianan Zhan, Jin Jin, Cheng Ma, Ruzhang Zhao, Jared O' Connell,

Yunxuan Jiang, 23andMe Research Team, Bertram L. Koelsch, Haoyu Zhang, **Nilanjan Chatterjee**[*]

[*]Correspondence to: **Jingning Zhang (jzhan218@jhu.edu)** and **Nilanjan Chatterjee (nilanjan@jhu.edu)**

Penalized regression for PRS construction

- We propose to solve the PRS coefficients in a penalized regression
- If we have individual-level data, estimate of $\boldsymbol{\beta}$ can be obtained by minimizing

$$\frac{1}{N}\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2 + \text{penalty}(\boldsymbol{\beta})$$

- However, we usually don't have individual-level data
  - By approximating $\boldsymbol{X^T X}$ by LD, $\boldsymbol{D}$, in a separate reference data, and $\boldsymbol{X^T y}$ by summary statistics, $\widehat{\boldsymbol{\beta}}$, it is equivalent to minimizing

$$\boldsymbol{\beta^T D\beta} - 2\boldsymbol{\beta^T}\widehat{\boldsymbol{\beta}} + \text{penalty}(\boldsymbol{\beta})$$

Polygenic Risk scOres based on enSemble of PEnalized Regression models (PROSPER)

- Solution of PRS coefficient for K populations, $\boldsymbol{\beta}_i$, $i = 1, \dots, K$, can be obtained using **coordinate decent algorithm** by minimizing

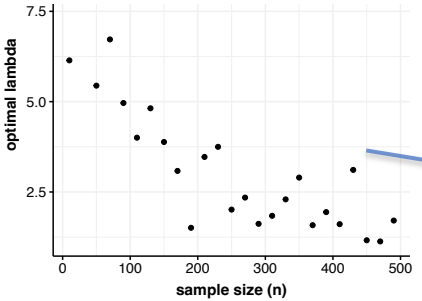
$$\sum_{i=1}^{K} \boldsymbol{\beta}_i^T \boldsymbol{D}_i \boldsymbol{\beta}_i - 2\boldsymbol{\beta}_i^T \widehat{\boldsymbol{\beta}}_i + \sum_{i=1}^{K} \lambda_i |\boldsymbol{\beta}_i| + \sum_{1 \le i_1 < i_2 \le K} c_{i_1 i_2} \left(\boldsymbol{\beta}_{i_1} - \boldsymbol{\beta}_{i_2}\right)^2$$

LD matrix

Summary statistics

Tuning parameter for sparsity

Tuning parameter for genetic similarity

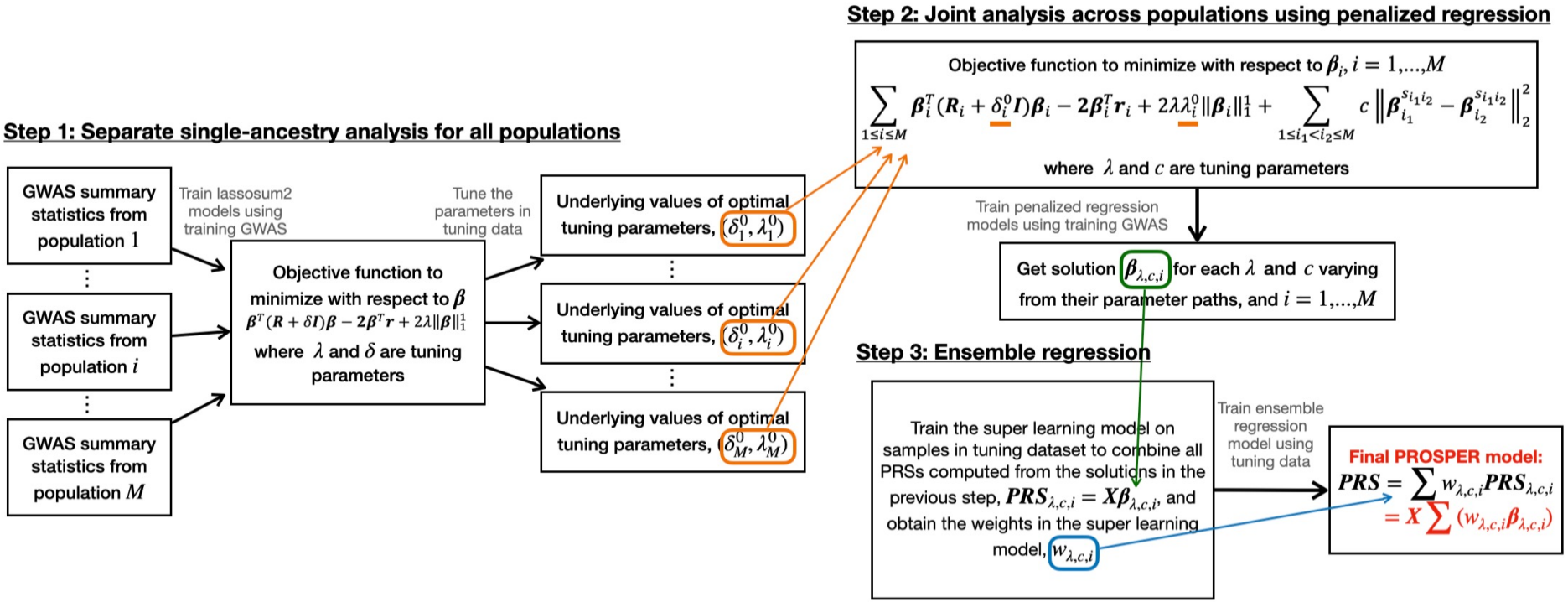


- Tuning parameter
  - λ_i, $i = 1, \dots, K$
  - $c_{i_1 i_2}$, $i_1 < i_2 = 1, \dots, K$
- **How to reduce the number of grid search when K is large?**
  - Let $\lambda_i = \lambda\, \lambda_{i_0}$ where λ_{i_0} is the value used in lassosum (single-ancestry)
  - Let $c_{i_1 i_2} = c$
- **Super learning**
  - Weighted combine PRS generated from all tuning parameter settings for all ancestries
  - # PRS combined

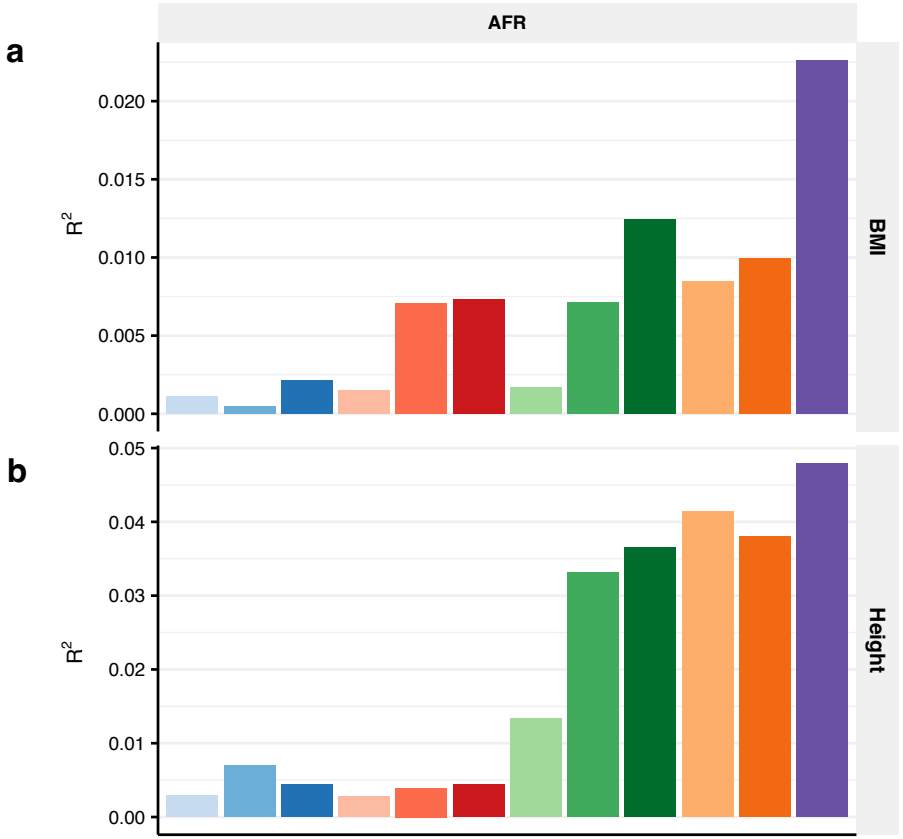
$$(\#\ candidate\ values)^{(\#\ tuning\ parameters)} \times (\#\ ancestries)$$

- E.g. 5-ancestry analysis
  - 2 tuning parameters: λ and c; 5 candidate values each
  - # PRS combined by super learning: $5^2 \times 5 = 125$

Flowchart


Step 1: Separate single-ancestry analysis for all populations
GWAS summary statistics from population 1
GWAS summary statistics from population i
GWAS summary statistics from population M
Train lassosum2 models using training GWAS
Objective function to minimize with respect to β
β^T(R + δI)β − 2β^T r + 2λ‖β‖_1^1
where λ and δ are tuning parameters
Tune the parameters in tuning data
Underlying values of optimal tuning parameters, (δ_1^0, λ_1^0)
Underlying values of optimal tuning parameters, (δ_i^0, λ_i^0)
Underlying values of optimal tuning parameters, (δ_M^0, λ_M^0)
Step 2: Joint analysis across populations using penalized regression
Objective function to minimize with respect to β_i, i = 1,...,M
Σ_{1≤i≤M} β_i^T(R_i + δ_i^0 I)β_i − 2β_i^T r_i + 2λλ_i^0‖β_i‖_1^1 + Σ_{1≤i_1<i_2≤M} c ‖β_{i_1}^{s_{i_1 i_2}} − β_{i_2}^{s_{i_1 i_2}}‖_2^2
where λ and c are tuning parameters
Train penalized regression models using training GWAS
Get solution β_{λ,c,i} for each λ and c varying from their parameter paths, and i = 1,...,M
Step 3: Ensemble regression
Train the super learning model on samples in tuning dataset to combine all PRSs computed from the solutions in the previous step, PRS_{λ,c,i} = Xβ_{λ,c,i}, and obtain the weights in the super learning model, w_{λ,c,i}
Train ensemble regression model using tuning data
Final PROSPER model:
PRS = Σ w_{λ,c,i} PRS_{λ,c,i}
= X Σ (w_{λ,c,i} β_{λ,c,i})

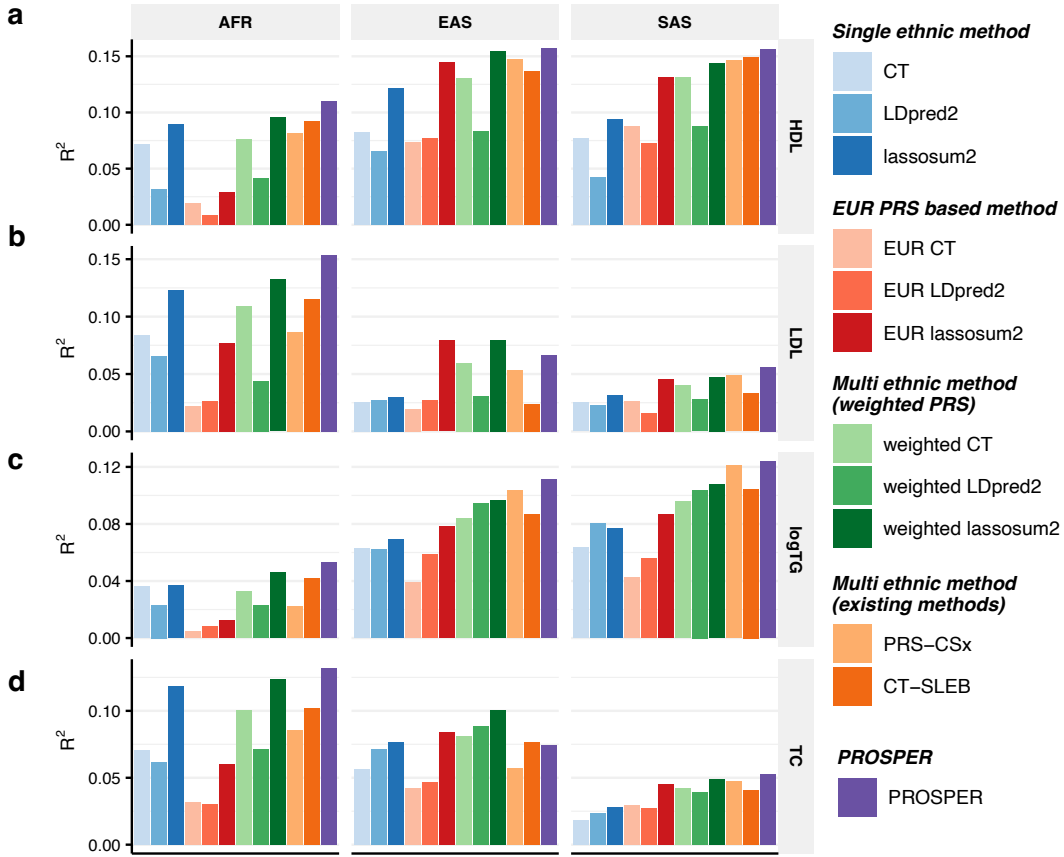
Results on data from All of US (AoU)



- Averaged sample size
  - EUR (N = 48K)
  - AFR (N = 22K)
- Improvement PROSPER over alternative methods (average across traits and ancestries) in R^2
  - 91.3% over PRS-CSx
  - 76.5% over CT-SLEB
  - 56.8% over weighted lassosum2

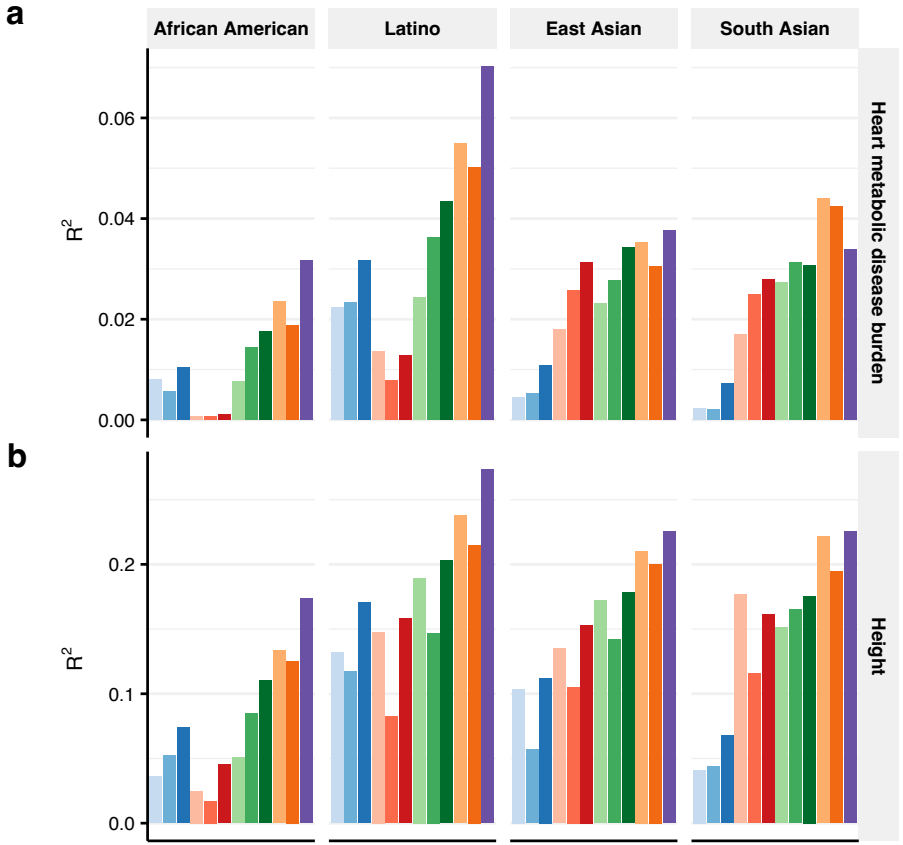
SNP set:
HM3 for PRS-CSx
HM3+MEGA for all other methods

Results on data from Global Lipids Genetics Consortium (GLGC)



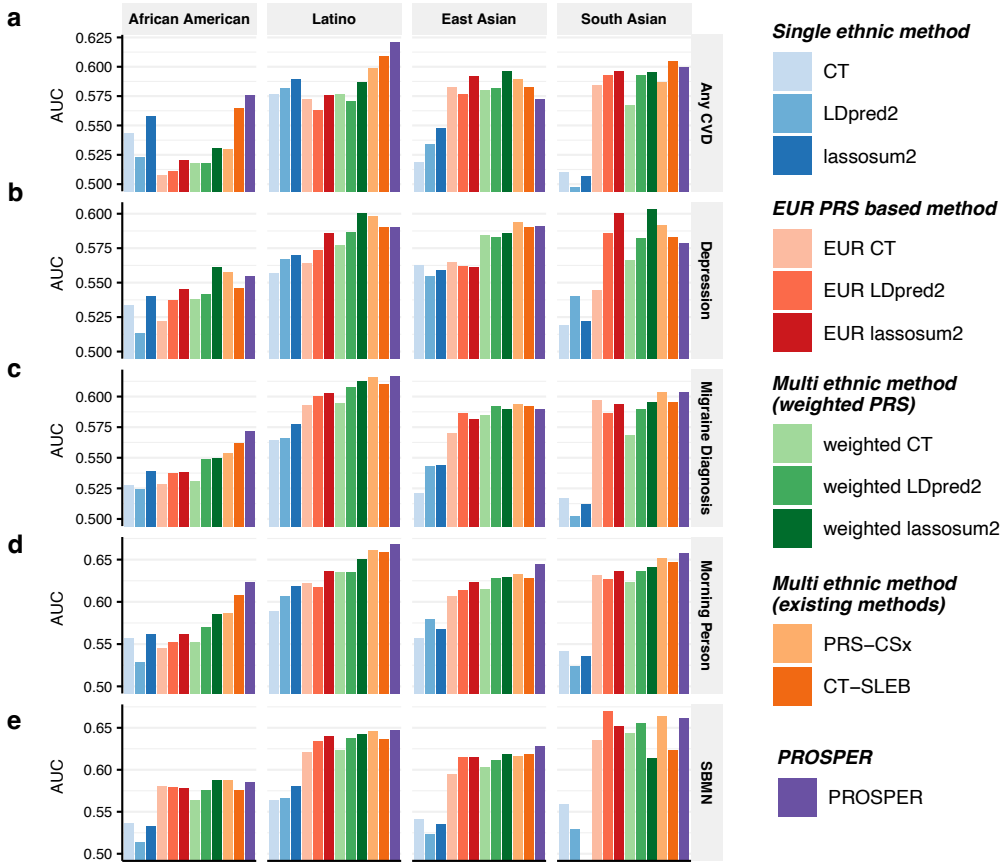
- Averaged sample size
  - EUR (N = 931K)
  - AFR (N = 93K)
  - EAS (N = 146K)
  - SAS (N= 34K)
- Improvement of PROSPER over alternative methods (average across traits and ancestries) in R^2
  - 34.2% over PRS-CSx
  - 37.7% over CT-SLEB
  - 13.1% over weighted lassosum2 in AFR
  - 12.3% over weighted lassosum2 in SAS

Results on data from 23andMe Inc. (23andMe) continuous traits



- Averaged sample size
  - EUR (N = 2700K)
  - African American (N = 136K)
  - Latino (N = 442K)
  - East Asian (N = 116K)
  - South Asian (N = 31K)
- Improvement of PROSPER over alternative methods (average across traits) in R^2
  - 32.4% over PRS-CSx in African American
  - 21.5% over PRS-CSx in Latino

Results on data from 23andMe Inc. (23andMe) binary traits



- Averaged sample size
  - EUR (N = 2370K)
  - African American (N = 109K)
  - Latino (N = 401K)
  - East Asian (N = 86K)
  - South Asian (N = 24K)
- Improvement of PROSPER over alternative methods (average across traits and ancestries) in logit-scale variance
  - 12.3% over PRS-CSx
  - 7.8% over CT-SLEB

Computational and memory usage

- Chromosome 22
- two ancestries: AFR, and EUR
  five ancestries: AFR, AMR, EAS, EUR, and SAS
- Computational time and memory usage

Method	Computational time (minutes)	Memory (Gb)
PROSPER (two ancestries)	3.0	2.24
PROSPER (five ancestries)	6.8	2.35
PRS-CSx (two ancestries)	111.1	0.78
PRS-CSx (five ancestries)	595.8	0.84

Conclusion

- PROSPER has substantial improvement over alternative methods and can deal with traits with a variety of genetic architectures

- PROSPER is not only more powerful for complex traits **with high polygenicity**, but also **robust to biomarker traits with large-effect loci.** In addition, PROSPER developed based on penalized regression is **an order of magnitude faster** compared to alternative Bayesian methods.