Evaluating polygenic score methods in admixed populations

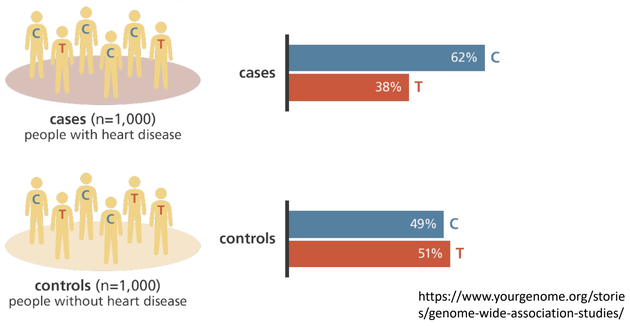
Ziqi Xu[1,2], Kangcheng Hou[3], Yi Ding[3], Bogdan Pasaniuc[3,4,5,6]

1. Dept. of Computer Science, UCLA; 2. BIG Summer Program, Institute of Quantitative and Computational Biosciences, UCLA;
3.Bioinformatics Interdepartmental Program, UCLA; 4. Dept. of Pathology & Laboratory Medicine, UCLA;
5. Dept. of Human Genetics, UCLA; 6. Dept. of Computational Medicine, UCLA

GWAS, Admixed Populations, and PGS

Genome-wide association study

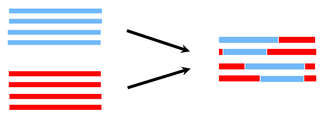
- genome-wide association test to investigate if any variant is associated with a trait
- Participants 95.86% are European, 0.19% admixed population for all studies to date. (https://gwasdiversitymonitor.com/)



https://www.yourgenome.org/stories/genome-wide-association-studies/

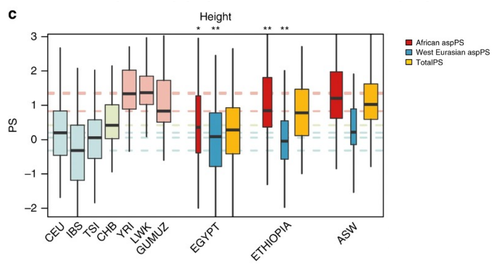
Admixed populations

- Mosaic ancestry segments originating from multiple continental ancestral populations
- Limited sample size



PGS $\text{PGS} = \mathbf{x}_n^T\boldsymbol{\beta} + \epsilon_n$

- Standard linear model to predict phenotype using trained effect sizes
- Genetic Value is the $\mathbf{x}_n^T\boldsymbol{\beta}$ component of PGS
- PGS accuracy decreases on population with ancestry far way from European ancestry (Martin et al. NG 2021)
- According to Marnetto et al 2019, partial PGS based on ancestry segments is a useful way to apply PGS on admixed populations.



Motivation

- investigate the best practices of applying PGS on admixed populations

Mathematics of LDpred2

LDpred2 extension (Ding et al. NG 2020)

- Approximating the posterior distribution of genetic values
- Prior distribution of effect sizes

$$\beta_j \sim \begin{cases} \mathcal{N}\left(0, \frac{h_g^2}{Mp_{\text{causal}}}\right), & \text{with probability } p_{\text{causal}} \\ 0, & \text{with probability } 1 - p_{\text{causal}} \end{cases}$$

- For each training dataset, derive a posterior distribution

$$p(\beta|\hat{\beta}_{\text{GWAS}}, \hat{\mathbf{R}}, h_g^2, p_{\text{causal}})$$

- Use MCMC sampling to generate posterior samples from causal effect posterior distribution

$$p\left(\tilde{\boldsymbol{\beta}}\right):\tilde{\boldsymbol{\beta}}^{(1)}, \tilde{\boldsymbol{\beta}}^{(2)}, \ldots, \tilde{\boldsymbol{\beta}}^{(B)}$$

- For each individual, generate point estimate of PGS

$$\widehat{\text{PRS}}_i = \mathbf{x}_i^\top \mathbb{E}\left[\tilde{\boldsymbol{\beta}}\right] \text{ where } \mathbb{E}\left[\tilde{\boldsymbol{\beta}}\right] = \int \tilde{\boldsymbol{\beta}} p\left(\tilde{\boldsymbol{\beta}}\right) d\tilde{\boldsymbol{\beta}}$$

Data

- 500k individuals of various ancestries from UK Biobank
  - Training: 370k individuals of European white British ancestry
  - Testing: 4323 admixed individuals with 40% European and 60% African ancestry in UK Biobank

Method

We first simulated 100 set of genetic value and phenotype for all 500k individuals in UK Biobank. For each simulation, we trained PGS models with LDpred2 auto mode on 270k individuals with European ancestries to obtain weights for computing PGS. We compute the PGS for each individual by three following methods:

Total PGS

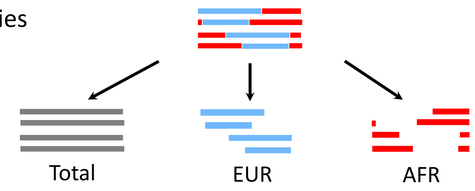
- PGS weights were applied to genotypes of admixed individuals regardless of local ancestries

Partial PGS – EUR

- PGS weights were applied to only genotypes of European local ancestries

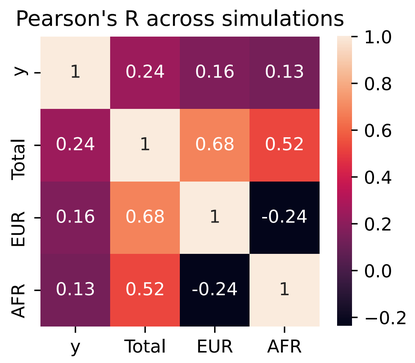
Partial PGS - AFR

- PGS weights were applied to only genotypes of European local ancestries

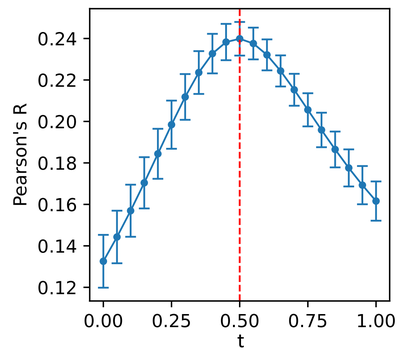


Results

Among these three methods, Total PGS has the best performance.

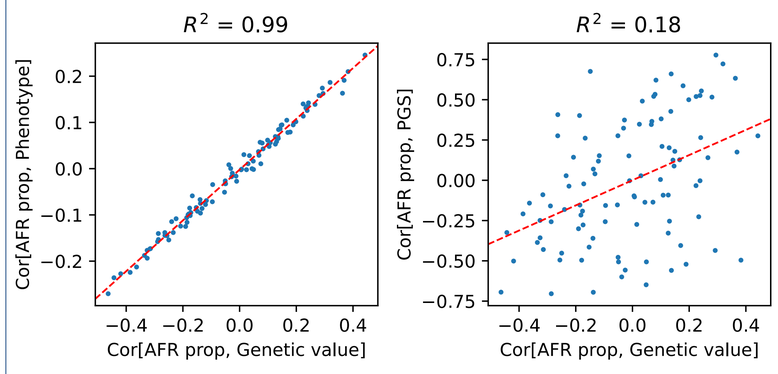


- The pair-wise average correlation heat-map shows that compared with EUR and AFR partial PGS, Total PGS has the highest prediction accuracy of phenotype.
- For African American population, EUR partial PGS surpasses AFR partial PGS.
- Define weighted PGS as: weighted PGS = t × EUR + (1 − t) × AFR
- In this equation, t is the partial proportion weight between two ancestries.
- Only when t equals to 0.5, denoting that partial PGS is equivalent to total PGS, the prediction accuracy is the highest.



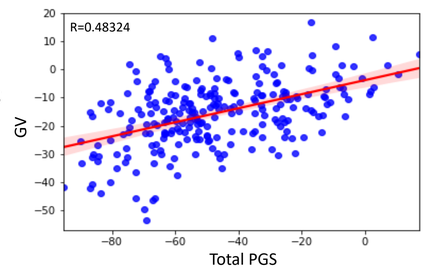
PGS cannot track well with the true genetic value (or phenotype) as a function of ancestry proportion.

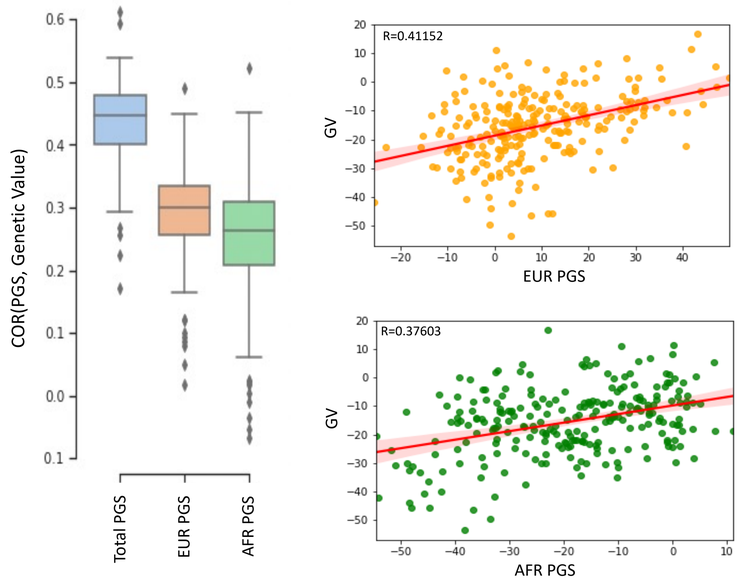
- The relation between genetic value and African ancestry proportion is not consistent with the relation between PGS and African ancestry proportion



Total PGS outperforms partial PGS when predicting genetic value.

- Correlation between different PGS methods and Genetic Value shows that Total PGS has the highest prediction accuracy and lowest prediction uncertainty.





Discussion

Limitations

- For partial PGS, we apply homogeneous weights (European trained effect sizes) to different ancestry subsets, which generates lower accuracy than total PGS.
- To improve the partial PGS method, we could train ancestry-specific effect sizes and apply to genotype subset of corresponding ancestry.

Acknowledgments

- The UK Biobank Resource (33297, 33127)

Contact

Ziqi Xu
ziqixu091@g.ucla.edu

UCLA Bioinformatics

References

1. Ding et al. Nature Genet 2020
2. Marnetto et al. Nature Comm 2019
3. Martin et al. Nature Genet 2021
4. "GWAS Diversity Monitor." *Home - GWAS Diversity Monitor*, https://gwasdiversitymonitor.com/.
5. Bycroft et al. Nature 2018
6. "Genome-wide association studies." Your Genome, https://www.yourgenome.org/stories/genome-wide-association-studies/

UCLA B.I.G Summer Bruins in Genomics