Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations

Ying Wang, Jing Guo, Guiyan Ni, Jian Yang, Peter M. Visscher, Loic Yengo

Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD 4072, Australia

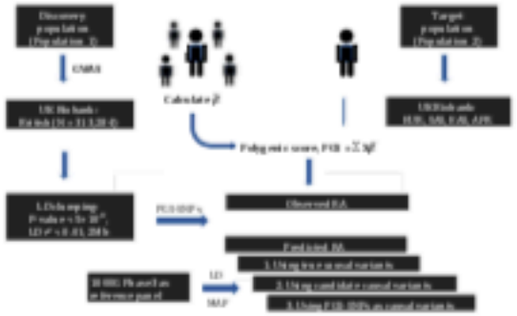
Research Question

1) Eurocentric PGS studies relative to worldwide populations[ref]

2) Disparity in PGS predictive performance between ancestries[ref]

Q: How to quantify the relative contribution of key factors (such as differences in LD and MAF between ancestries) to that loss of prediction accuracy?

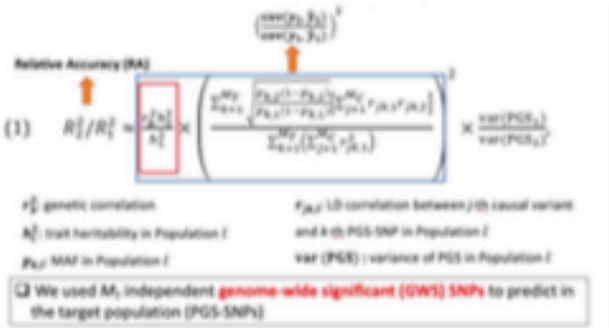
Method



Candidate causal variants: any SNP in LD (r^2 > 0.45) with a GWS SNP and located within 100 kb of the latter

Results

1) Expected RA of PGS in ancestry divergent populations



2) Performance of the method on simulated data

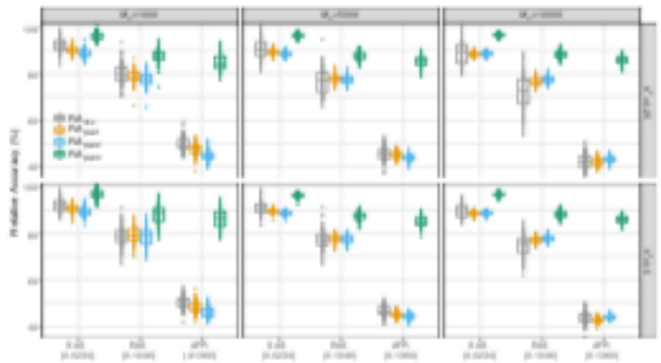


Fig 1. Trans-ancestry relative prediction accuracy of PGS in different simulation scenarios. RA_{obs} refers to the observed RA calculated. The predicted RA labelled as RA_{pred1} is estimated based on parameters calculated from SNP pairs of PGS-SNPs and known causal variants within 100 kb; RA_{pred2} refers to RA calculated using SNP pairs of PGS-SNPs and candidate causal variants. RA_{pred3} refers to the naive predicted RA when assuming that PGS-SNPs are the causal variants. The numbers under the ancestry labels in *x*-axis denoted the pairwise F_{ST} calculated using HapMap3 SNPs between discovery population and target population. Boxes represent the first and third quantiles and whiskers are 1.5-fold the interquartile range. The points represent the RA for 100 replicates. The median estimates are shown as the horizontal line in the boxes.

3) Impact of negative selection

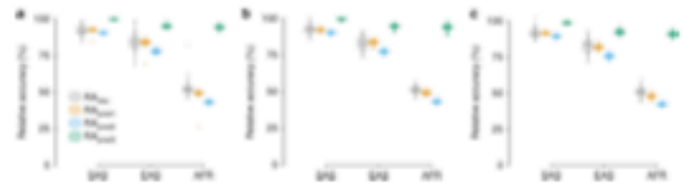


Fig 2. Impact of negative selection on PGS trans-ancestry relative accuracies. Traits were simulated with a heritability $h^2 = 0.5$ and assuming $M_C = 5000$ causal variants. Negative selection was modelled using a parameter S such that smaller values of S indicate stronger strength of selection. Values of S are denoted S_1 and S_2 in the discovery population and target populations, respectively. We considered three scenarios: a $S_1 = S_2 = -0.5$; b $S_1 = -0.5$, $S_2 = -0.75$; and c $S_1 = -0.75$, $S_2 = -0.5$. RA_{obs}, RA_{pred1}, RA_{pred2} and RA_{pred3} labels are defined as in the legend of Fig 1.

4) Application to real data

Table 1. The number of GWS SNPs for traits and diseases studied in the UK Biobank

Trait	Abb.	GWS SNPs
Standing height	Height	1,182
Body mass index	BMI	338
LDL cholesterol	LDL	179
HDL cholesterol	HDL	271
Triglycerides	TG	178
Asthma	Asthma	71
Type 2 Diabetes	T2D	44
Hypertension	HTN	74

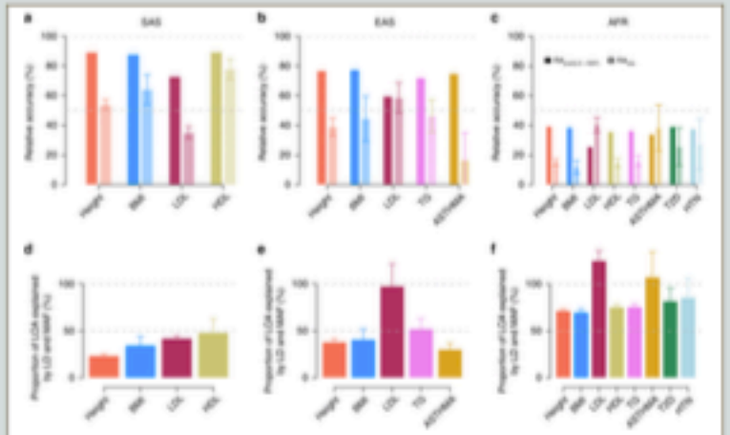


Fig 3. Trans-ancestry relative prediction accuracy of PGS of 5 quantitative traits and three common diseases. a–c Relative accuracies (RA) are calculated as the ratio of the squared correlation between PGS and traits/diseases in UKB participants of non-European ancestry over the same squared correlation estimated in ~20,000 independent UKB participants of European ancestry. We report here only ancestry-traits/disease pairs, with a significant reduction in RA (Wald test, p-value <0.05). $RA_{pred(LD+MAF)}$ refers to the RA predicted only using information from LD and MAF differences between ancestries. RA_{obs} refers to observed RA calculated using independent genome-wide significant trait-associated SNPs. Panels d–f show the proportion of the loss of accuracy (LOA) explained by LD and MAF calculated as $100\% \times (1 - RA_{pred(LD+MAF)})/(1 - RA_{obs})$. The grey dashed lines are $y = 100\%$ and $y = 50\%$. Error bars in the figure represent the standard errors of observed RA or proportion of LOA explained by LD and MAF in each ancestry-traits/disease pair.

Summary

1. Our theory can predict the relative accuracy attributable to LD and MAF differences between ancestries with little bias.
2. When heritability is constant and effect sizes of causal variants are perfectly correlated between ancestries, differences in strengths of selection between ancestries might have a negligible impact on the RA of PGS.
3. Over 2/3rd of LOA in AFR ancestry is expected because of LD and MAF differences between ancestries for traits like T2D, BMI and height.

Reference

1. Duncan, L. et al. Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. Commun.* **10**, 1–9 (2019).

*This poster is based on the published paper ***Wang, Y., Guo, J., Ni, G. et al. Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. Nat Commun 11, 3865 (2020).***

IMB Institute for Molecular Bioscience

Program in Complex Trait Genomics