

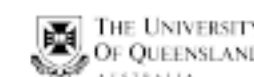
# Integrating the genetic and environmental information to improve the phenotype prediction for body mass index

Huanwei Wang<sup>1</sup>, Angli Xue<sup>1</sup>, Longda Jiang<sup>1</sup>, Julia Sidorenko<sup>1</sup>, Peter M. Visscher<sup>1</sup>, Naomi R. Wray<sup>1,2</sup>, Allan F McRae<sup>1</sup>, Jian Zeng<sup>1</sup>, Jian Yang<sup>1,3</sup>

1 Institute for Molecular Bioscience, The University of Queensland, Brisbane, Queensland, Australia

2 Queensland Brain Institute, The University of Queensland, Brisbane, Queensland, Australia

3 School of Life Sciences, Westlake University, Hangzhou, Zhejiang 310024, China



## Background

- Prediction analysis is valuable in medicine and most predictors, apart from family history and monogenic mutations, are environmental/non-genetic factors (environmental risk score, ERS);
- Genetic/polygenic risk score (GRS) is promising due to its increasing prediction accuracy and low cost;
- This project is aiming to explore different methods to combine GRS and ERS (genetic and environmental risk score, GERS) for body mass index (BMI).

## Methods

### Data



- Target phenotype: BMI
- Environmental factors: 8 lifestyle factors
- Genotype: 1,317,930 HapMap3 SNPs
- Individuals: 348,501 unrelated European individuals
- Training/tuning/test datasets: randomly (rrr), based on age (yoo or yyo), or based on the availability of second measurements of BMI (ffs)

### Risk scores

- $GRS = w_1 G_1 + w_2 G_2 + \dots + w_m G_m$ , by SBayesR<sup>1</sup>
- $GERS = u_0 GRS + u_1 E_1 + \dots + u_k E_k$ 
  - GERS<sub>rp</sub>:  $u_0=1$ ;  $u_1, \dots, u_k$ =phenotypic correlation ( $r_p$ )
  - GERS<sub>mlr</sub>:  $u_0, \dots, u_k$  estimated by multiple linear regression (mlr) in the tuning dataset
  - GERS<sub>mr</sub>:  $u_0=1$ ;  $u_1, \dots, u_k$ =causal effect size  $b_{xy}$  inferred by GSMR<sup>2</sup>
- $MGRS = v_0 GRS_0 + v_1 GRS_1 + \dots + v_k GRS_k$

## Results

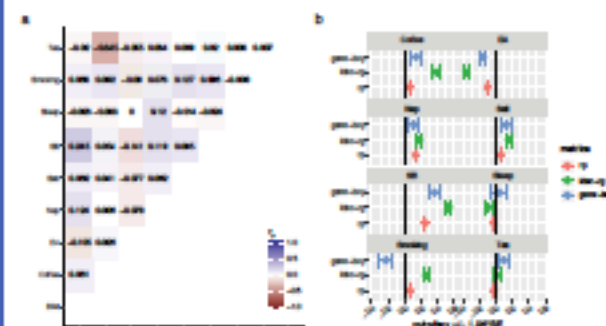


Figure 1 Phenotypic correlation ( $r_p$ ), genetic correlation ( $r_g$ ), and causal effect size ( $b_{xy}$ ) for BMI and eight environmental factors. SB=Sedentary behaviour, EA=Education attainment

We found all GERSs with 8 environmental factors (Fig. 1) built by three different methods performed better than the GRS across all four data splitting strategies (Fig. 2), where GERS<sub>mlr</sub>, which accounted for both the correlation between GRS and environmental factors and the correlation within different environmental factors, performed the best.

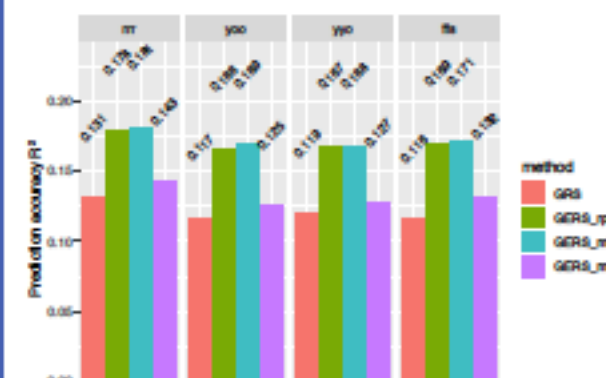


Figure 2 Prediction accuracy of GRS and GERSs with eight environmental factors in different data splitting strategies.

We found that the prediction accuracy of GERS with SB was the highest, followed by Nap, and GERS with Coffee, Sleep and Tea almost cannot improve prediction accuracy (Fig. 3).

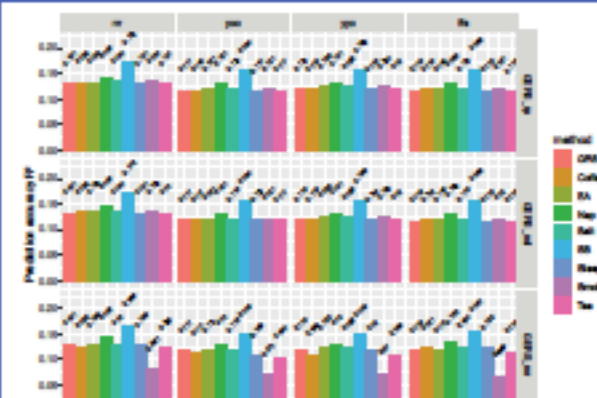


Figure 3 Prediction accuracy of GRS and GERSs with each one of eight environmental factors in different data splitting strategies.

There are other risk scores integrating only the genetic part of environmental factors, which we called MGRS here, including multi-trait prediction (e.g. MTAG, SMTpred), multiple polygenic risk scores (MPS), and metaGRS. We found that the MGRS combining the GRSs of BMI and eight environmental factors did not improve the prediction accuracy in comparison with BMI GRS only (Fig. 4)

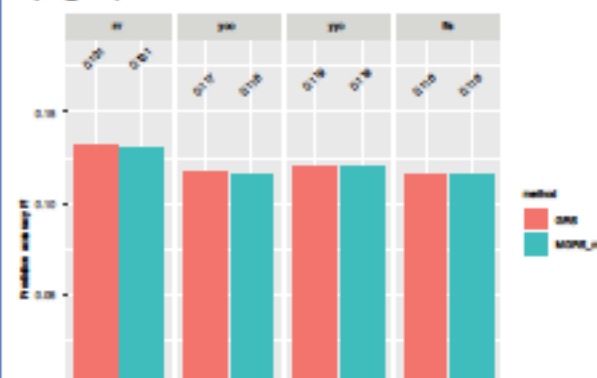


Figure 4 Prediction accuracy of GRS of BMI and MGRS of BMI and eight environmental factors in comparison with GRS in different data splitting strategies.

Our simulation study showed the prediction accuracy of MGRS cannot be improved if all GRSs were based on same training dataset. (Fig. 5)

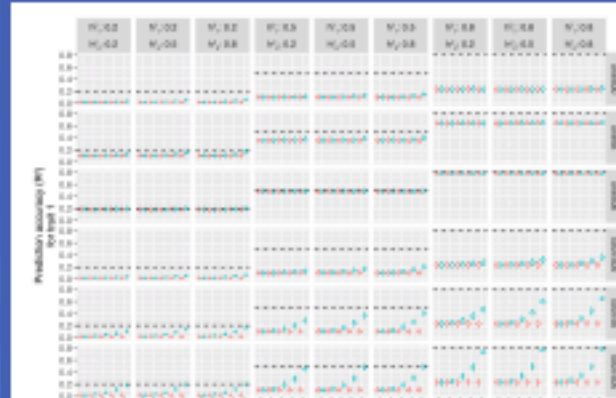


Figure 5 Simulation study evaluating the prediction accuracy for trait 1 of MGRS in comparison to GRS.

## Summary

- Incorporating environmental information can improve the prediction accuracy of GRS for BMI
- A multiple linear regression (GERS<sub>mlr</sub>) showed the best performance comparing with phenotypic correlation (GERS<sub>rp</sub>) and mendelian randomization estimates (GERS<sub>mr</sub>) for GERS construction
- Sedentary behavior contributed most of the improvement of GERS vs GRS for BMI prediction
- MGRS method did not work if the GWAS summary data were from the same dataset

## Reference

1. Lloyd-Jones et al., Nature Communications, 2019
2. Zhu et al., Nature Communications, 2018

This research has been conducted using the UK Biobank Resource under Application Number 12505