Predicting key GWAS outcomes under a point-normal polygenic model

Tian Wu[1], Zipeng Liu[2], Timothy Shin Heng Mak[2], Yan Dora Zhang [2,3], Pak Chung Sham [1,2,4]

1. Department of Psychiatry, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China 2. Centre for PanorOmic Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China 3. Department of Statistics & Actuarial Science, Faculty of Science, The University of Hong Kong Hong Kong SAR, China 4. State Key Laboratory of Brain and Cognitive Sciences, The University of Hong Kong, Hong Kong SAR, China

INTRODUCTION

It has been observed that for complex phenotypes, associated SNPs become detectable as the sample size reaches a certain minimum, above which the number of significantly associated SNPs increases with increasing sample size in an approximately linear fashion [1]. However, a detailed modelling of the entire relationship between the number of significant SNPs and sample size has not been performed. Here, we consider a point-normal polygenic model, to predict the number of independent significant SNPs given genome-wide association study (GWAS) sample size, key parameters of genetic architecture of the phenotype and disease. Under the same assumptions, we also predicted several important outcome indices of the GWAS, including the overall phenotypic variance explained by significant SNPs, and the predictive accuracy of polygenic scores that weight SNPs by the effect size estimates from GWAS, after shrinkage by various methods. Calculations were performed from analytically derived formulae, and were validated by simulations. We compared the predictions of our method to the observed results of GWAS on height and BMI, and found that they were in agreement. Thus, our method could be a useful tool for the design of GWAS and for predicting the future behavior of GWAS as sample sizes increase further.

MODEL DESCRIPTION

1. Effect size distribution

The overall effect size across all SNPs follows a point-normal distribution.

$$\beta \begin{cases} 0 & \text{proportion of } \pi_0 \\ \sim N\left(0, \frac{h^2}{m(1-\pi_0)}\right) & \text{proportion of } 1-\pi_0 \end{cases}$$

- h^2 is the SNP heritability or the heritability on the liability scale if the trait is binary.
- m is total number of nearly **independent** SNPs that may contribute to the phenotypic variance.
- π_0 is the proportion of SNPs that do not contributing to the variance of phenotype.

Accordingly, the distribution of effect size estimates across all SNPs is a nor

$$\hat{\beta} \begin{cases} \sim N\left(0, \frac{1}{n}\right) & \text{proportion of } \pi_0 \\ \sim N\left(\beta, \frac{h^2}{m(1-\pi_0)} + \frac{1}{n}\right) & \text{proportion of } 1-\pi_0 \end{cases}$$

- n is sample size.

When phenotype is binary, we approximate the log odds ratios calculated by logistic regression model to effect size on the liability scale [2]. K is disease prevalence.

2. Polygenic model

To predict genetic risk, we adopt the polygenic model $y_i = \sum_{j=1}^{m_s} \beta_j x_{ij} + \varepsilon_i$. Under this model, the true and estimated polygenic scores are $G_i = \sum_{j=1}^{m_s} \beta_j x_{ij}$ and $\hat{G}_i = \sum_{j=1}^{m_s} \hat{\beta}_j x_{ij}$. $corr^2(G, \hat{G})$ is used to measure the prediction accuracy.

METHODS

- To obtain the expectation and variance of statistical power, we first partitioned a wide range of probable effect sizes into a large number of small, equal and non-overlapping intervals.

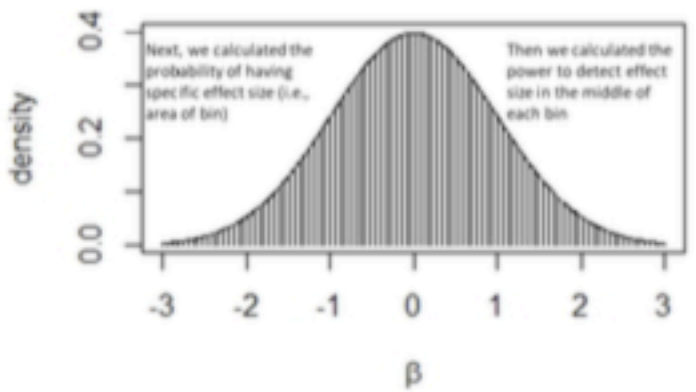

Fig. 1 Calculation of the expectation of statistical power: a demonstration

- Based on the definition of variance, we calculated the variance of power.
- The statistical powers of the intervals, weighted by their probabilities, are used to obtain a kernel density estimate of the distribution of statistical power.
- Based on the expectation and variance of statistical power, we derived the formula to calculate the expected number of significant SNPs, E(X), and variance explained by the significant SNPs, Var(X).
- We validated the derived formulae by simulating 100 GWASs with given parameters, counting the number of significant SNPs and calculating the variance explained by these SNPs. Comparison between the theoretical (orange and green dots) and empirical (violin plots) results is shown in Fig. 2.
- We also predict the polygenic score (PGS) prediction accuracy based on derived formulae.

APPLICATION

We Compared results predicted by our model to the reported GWASs on height, BMI, and major depressive disorder and found agreement.

Phenotype	Study sample size	Number of significant SNPs		Variance explained by significant SNPs	
		Predicted	Reported	Predicted	Reported
Height	~700,000	3194 [3137.51, 3249.56]	3290	16.03%	19.7%[6]
BMI	~700,000	574 [538.58, 610.32]	716	8.5%	5.0%[6]
Major depressive disorder	~800,000	308 [279.47, 336.99]	102	2.16%	1.5%~3.2%[7]

Table 1. Theoretical versus reported number of independent SNPs and variance explained by these SNPs

REFERENCES

1. Visscher, P. M., Wray N.R., Zhang Q., Sklar, P., McCarthy M. I., Brown, M. A., & Yang, J. (2017). 10 years of GWAS discovery: Biology, function, and translation. Am J Hum Genet, 101(1), 5-22. doi:10.1016/j.ajhg.2017.06.005. 2. Wu, T, & Sham, PC. (2019). On the transformation of genetic effect size from logit to liability scale (submitted). 3. Mak, T. S. H., Kwan, J. S., Campbell, D. D., & Sham, P. C. (2016). Local True Discovery Rate Weighted Polygenic Scores Using GWAS Summary Data. Behav Genet, 46(4), 573-582. doi:10.1007/s10519-015-9770-2. 4. Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society Series B (Methodological), 58(1), 267-288. 5. Euesden, J., Lewis, C. M., & O'Reilly P.F. (2015). PRSice: Polygenic Risk Score software. Bioinformatics, 31(9), 1466-1468. doi:10.1093/bioinformatics/btu848. 6. Yengo, L., Sidorenko, J., Kemper, K. E., Zheng, Z., Wood, A. R., Weedon, M. N., ... Consortium, G. (2018). Meta-analysis of genome-wide association studies for height and body mass index in approximately 700000 individuals of European ancestry. Hum Mol Genet, 27(20), 3641-3649. doi:10.1093/hmg/ddy271. 7. Howard, D. M., Adams, M. J., Clarke, T. K., Hafferty, J. D., Gibson, J., Shirali, M., ... McIntosh, A. M. (2019). Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. Nat Neurosci, 22(3), 343-352. doi:10.1038/s41593-018-0326-7. 8. Bigdeli, T.B., Lee, D., Webb, B. T., Riley, B. P., Vladimirov V. I., Fanous, A. H., ... Bacanu, S. A. (2016). A simple yet accurate correction for winner's curse can predict signals discovered in much larger genome scans. Bioinformatics, 32(17), 2598-2603. doi:10.1093/bioinformatics/btw303. 9. Efron, B. (2010). Large-scale inference: empirical Bayes methods for estimation, testing, and prediction. Cambridge: Cambridge University Press.

RESULTS

1. Distribution of statistical power to detect causal SNPs becomes a bi-model with the increase of sample size.

2. To Predict the expectation and variance of the number of independent significant SNPs

- $E(X) = m\pi_0\alpha + m(1-\pi_0)E(p)$
- $Var(X) = m\pi_0\alpha(1-\alpha) + m(1-\pi_0)[E(p)(1-E(p)) - var(p)]$, p is statistical power.

Theoretical values or their 95% confidence intervals overlap with the empirical results, which validated our derived formulae (Fig. 2).

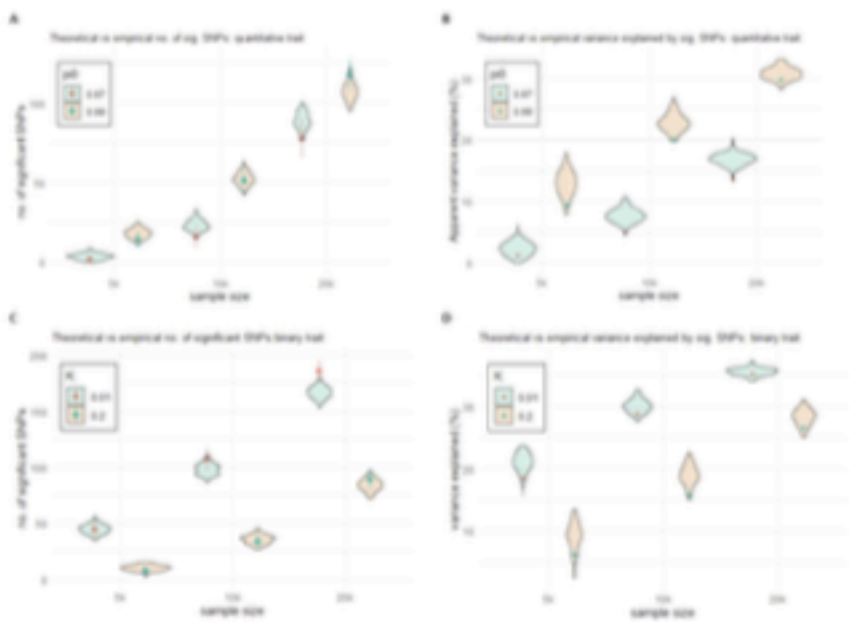

Fig. 2 Theoretical versus empirical number of independent significant SNPs and variance explained by these SNPs ($m = 50k$, $\pi_0 = 0.99$, $h^2 = 0.4$; case ratio is 0.5 for binary trait)

3. To predict the polygenic score (PGS) prediction accuracy

- Among the methods to construct PGS, the conditional expectation $E(\beta|\hat{\beta})$ or the local TDR [3] method, is always the best in terms of the prediction accuracy.
- With moderate sample size, LASSO [4] outperforms p-value thresholding [5] and ordinary least square estimate. However, when sample size is large enough, p-value thresholding is also able to give accurate polygenic prediction.

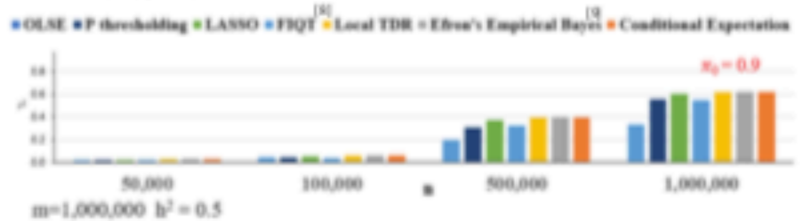

Fig. 3 Predicting prediction accuracy of PGS constructed by various of shrinkage method