Polygenic Risk Prediction for IQ Across Continental Populations

Raehash Shah and Max Shushkovsky

Abstract

Understanding the genetic basis of complex traits, such as Intelligence Quotient (IQ), remains a central goal in population genetics. This study investigated the potential of polygenic risk scores (PRS) to predict IQ across diverse populations. PRS leverages genetic variants associated with heritability to assess an individual's predisposition to a trait. Even with the limited information on the heritability of IQ, a pipeline was created that used SNP data to estimate PRS for IQ within and across populations. The pipeline evaluated model performance in a single-population context and assessed the generalizability of the prediction across diverse populations. Additionally, the study explored the impact of training the PRS model on a diverse dataset compared to a single population. The results showed that the model had the best prediction when using a mixed population. However, in all schemas, the pipeline did not achieve performance close to the true heritability of IQ suggesting there are more factors other than genetics that determine the IQ of an individual. Further analysis on demographic specific populations rather than continental populations may lead to better estimations of the PRS of IQ.

Introduction

A. Project Formulation

The goal of this project is to investigate polygenic risk prediction (PRS) for Intelligence Quotient (IQ) across diverse populations. PRS is a method of assessing an individual's risk of developing a trait by analyzing their genetic variants and leveraging heritability. Recent research suggests that IQ exhibits some degree of heritability. Therefore this project developed a pipeline that uses heritability of IQ to perform PRS across individuals within a population. Once the model predicted risk for a single population, the study compared how generalizable it was in predicting risks between different populations and when trained on a diverse population. By doing this, this project contributed a comprehensive framework for generating PRS while offering insights into the genetic variants associated with IQ.

B. Motivation and Short Summary

Heritability of IQ is a divisive topic. On one hand, some scientists show that IQ is a trait that can be manipulated by human action, increased or lowered by environmental factors. On the other hand, other scientists proved that it is a fixed value determined by an individual's DNA. This debate suggested there may be a mixture of genetic and environmental factors that influence IQ and further research is needed to assess to what extent. Genetic testing, a method that tells an individual how likely they are to display a certain trait, has seen a recent boom where from 2012 to 2022, there has been a rise from 607 to 3,097 new genetic tests available in the USA[4]. This field uses the heritability of traits to "measure disease liability by summarizing disease risk across hundreds of thousands of genetic variants"[2] which is the hope of PRS. One immediate issue is the abundance of data for European-related populations. To ensure that PRS is valid for

individuals of all populations, there is "population-specific normalization" that allows for mathematically sound computation[2]. Since research has concluded that IQ is polygenic, IQ's heritability factor can be computed as an aggregation of individual gene effects in PRS since it will aggregate "those effects for trait prediction in independent samples"[3].

Regarding heritability of IQ, there are various schools of thought to its actual quantitative value. A popular study in the late 1990s concluded that narrow-sense heritability of IQ, "is about 34%"[1]. The authors argued that the original thought of IQ heritability being 60-85% from adoption studies is not as accurate as their values produced from adult monozygotic twin studies. However, discussion from a study published in 2017 concluded that statistical power was a major limiting factor in obtaining truly accurate values. These authors believed that "genome-wide polygenic scores will soon be available that can predict more than 10% of the variance in intelligence"[7]. Comparing the two studies across almost 20 years, there has been a decreasing shift in the perceived heritability of IQ.

C. Overview of Analysis and Data

This paper presents a comprehensive analysis of polygenic risk prediction for intelligence across three diverse continental populations using HapMap3 SNP data and published summary statistics associated with IQ to offer insights into the genetic determinants of IQ. The approach in this paper considered both LD pruning and p-value thresholding to maximize predictive accuracy. By adjusting the training and testing sets, a mixture model was analyzed to see how it impacts model performance and accuracy on our test set.

Results

A. Model Generation

a.

	Total Number of SNPs	Number of SNPs into PLINK	Causal SNPs after LD-Clumping
European	48708	4870	2702
Asian	48708	4870	2610
African	48708	4870	2873

b.

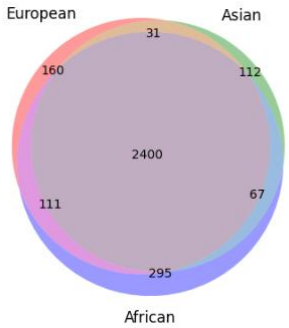
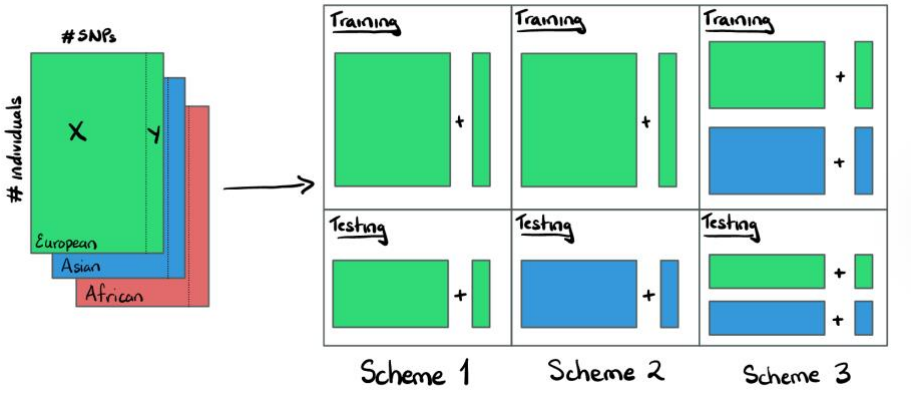


Figure 1: Identifying the Number of Causal SNPs. After identifying the number of intersecting SNPs in the HapMap3 project for chromosome 1 and the scDRS summary statistics, the goal was to identify the number of causal SNPs. Therefore, as seen in figure 1a, for each population, the top 10% of SNPs from the summary statistics were passed into PLINK to perform LD-clumping with hyperparameters of a 500kb window size, 2700 step size and 0.1 r^2value. Performing LD-clumping with each of the different datasets resulted in a different proportion of causal SNPs being identified since each continental population has a different

observed LD pattern. As seen in figure 1b, a venn-diagram is done to visualize the number of causal SNPs shared across the populations where the biggest set of size 2023 is the one shared among all of the populations.

B. Analysis of Model

a.



b.

		Testing Set						
Training set		European	Asian	African	European + Asian	European + African	Asian + African	European + Asian + African
	European	0.0103	0.0009	3.09e-5	0.0016	0.0016	0.0003	0.0004
	Asian	0.0205	0.0006	0.0060	0.0067	0.0109	0.0030	0.0063
	African	0.0303	**0.0081**	0.0447	0.0018	0.0379	0.0194	0.0203
	European + Asian	0.0132	0.0009	0.0050	0.0019	0.0079	0.0006	0.0028
	European + African	0.0040	0.0016	0.0509	0.0001	0.0329	0.0270	0.0203
	Asian + African	**0.0323**	6.46e-5	0.0546	**0.0075**	**0.0455**	**0.0297**	**0.0291**
	European + Asian + African	0.0094	0.0001	**0.0546**	0.0025	0.0373	0.0296	0.0237

Figure 2: Risk Prediction in Different Schemes. To analyze model performance, three different schemes as seen in Figure 2a were performed varying the training and testing set. In figure 2b, the results of all arrangements of training and testing can be seen with the values bolded as the best for the testing set. Looking at the risk prediction schema, it seems that training on Asian + African dataset had the highest r^2 when performing risk prediction on most of the other sets of testing data.

Results above in Figure 2 show that in most test splits, training on European data did not provide significant information in resulting PRS score. However, training on a mixed population that has Asian or African samples generally performed better than training on other mixed populations. This may be due to an increased number of individuals in that population cohort.

Discussions

Based on the risk prediction schema in Figure 2b, it was seen that using a mixed population, namely the Asian and African populations, led to the greatest PRS scores for other mixed populations. Single continental population training sets did not perform well on their respective testing sets. This leads to the major limitation of this investigation, which is that every population within a continental population has its intrinsic genetic differences. By mixing them together, regardless of shared continent, distinctual information is lost at the expense of larger training samples. As a result, there may have been overfitting that occurred. This is evidenced by Figure 3 in the appendix. There it is seen that, with greater SNP thresholding based on p-value, validation r^2 increases to values close to 0.9. High r^2 means that predicted phenotypes were very similar to true phenotypes. Further work would mean obtaining more samples from individual samples to use in single population analysis, which is supported by the earlier referenced literature regarding statistical power being the limiting factor in these explorations. Additionally, the specific PRS values are much smaller than the heritability value set to 0.2. This hints that there are strong non-genetic factors at play, such as the environment. One avenue to explore would be to modify the heritability value and amount of LD-pruning. More modern literature has suggested IQ heritability to be closer to 0.2, but various values could be analyzed to see if that impacts the PRS values. Additionally, the bottom 40% of the top 10% of causal SNPs were pruned out, as that was an estimate to the number of causal SNPs expected. However, further exploration could be considered in changing that estimate number and see how PRS values differed.

Methods

A. Data Acquisition

In this project, HapMap3[5] genotype for chromosome 1 and SNP data was used across 3 continent populations (European - CEU and TSI, Asian - CHB, JPT, and CHD, African - YRI, LWK, and MKK). In addition, single-cell disease relevance score (scDRS) summary statistics[12] related to IQ was used for the same populations to estimate causal effect size[9]. The intersection of the

datasets meant that this project worked with 48,708 SNPs for 200 European individuals, 255 Asian individuals and 346 African individuals.

B. Data Simulation

To simulate our phenotype data ($Y^{sim} \epsilon\ \mathbb{R}^{1\times N}$), this project regressed the genotype ($X\epsilon\ \{0,1,2\}^{M\times N}$) and the Z-scores ($\widehat{\beta}\ \epsilon\ \mathbb{R}^{1\times M}$) from scDRS summary statistics in the form of $Y^{sim} = \beta^{sim}X + \varepsilon$. Causal SNPs were identified from chromosome 1 using the top ~10% causal SNPs using Z-score. However, these SNPs may not all be causal. Therefore, since it has been shown that LD-pruning can improve PRS[10], LD-clumping was performed with PLINK[8] to identify a smaller set of causal SNPs. Once identified, the new causal effects were $\widehat{\beta_i}$ = Z-score if i is causal and 0 otherwise. Then these causal effects were normalized: $\beta^{sim} = \hat{\beta}\sqrt{\frac{h_g{}^2}{\sigma^2(X\hat{\beta})}}$ where Heritability $(h_g{}^2) = \frac{\sigma^2(X)}{\sigma^2(Y)} \leq 0.2$ and σ^2 is the variance. In addition, $\varepsilon \sim N(0, 1 - h_g{}^2)$ so $\sigma^2(Y^{sim}) = 1$.

C. Risk Prediction

Risk prediction involves estimating phenotypic data using genotype data. Therefore, this project split the data into training, validation and testing sets. To perform p-value thresholding, for a set of p-values, the number of SNPs were trimmed based on their Armitage Trend Test (ATT) value (see Figure 3). With this new genotype data, a linear regression was trained on the training set to get β_{model} that satisfies $Y_{train} = \beta_{model}X_{train}$ and passed onto the validation set to evaluate the model. After performing five-fold cross validation, the best β_{model} from the maximum prediction accuracy[11] in the validation set was used to estimate $\hat{Y}_{test} = \beta_{model}X_{test}$. Final PRS scores were computed using $r^2(\hat{Y}_{test}, Y_{test}^{sim})$.

D. Evaluation of Pipeline

Upon choosing the model that best captures a non-infinitesimal architecture, an important question is how well this model generalizes to different populations. Multi-ethnic PRS have been shown to improve prediction accuracy in a non-European population[6]. Therefore this project answered the following questions to identify the generalizability of a single population model compared to a mixed population model:

(1) How does risk prediction for IQ change across different populations when trained and tested on the same population?
(2) How does risk prediction for IQ change when trained on one population and tested on other single and mixed populations?
(3) How does risk prediction for IQ change when trained with a mixed population and tested on other single and mixed populations?

References:

[1] Devlin, B., et al. “The heritability of IQ.” *Nature*, vol. 388, no. 6641, 31 July 1997, pp. 468–471, https://doi.org/10.1038/41319.
[2] Folkersen, Lasse, et al. “Impute.me: An open-source, non-profit tool for using data from direct-to-consumer genetic testing to calculate and interpret polygenic risk scores.” *Frontiers in Genetics*, vol. 11, 30 June 2020, https://doi.org/10.3389/fgene.2020.00578.
[3] Genç, Erhan, et al. “Polygenic scores for cognitive abilities and their association with different aspects of General Intelligence—a deep phenotyping approach.” *Molecular Neurobiology*, vol. 58, no. 8, 5 May 2021, pp. 4145–4156, https://doi.org/10.1007/s12035-021-02398-7.
[4] Halbisen, Alyssa L., and Christine Y. Lu. “Trends in availability of genetic tests in the United States, 2012–2022.” *Journal of Personalized Medicine*, vol. 13, no. 4, 6 Apr. 2023, p. 638, https://doi.org/10.3390/jpm13040638.
[5] “Integrating common and rare genetic variation in diverse human populations.” *Nature*, vol. 467, no. 7311, Sept. 2010, pp. 52–58, https://doi.org/10.1038/nature09298.
[6] Márquez-Luna, Carla et al. “Multiethnic polygenic risk scores improve risk prediction in diverse populations.” *Genetic epidemiology* vol. 41,8 (2017): 811-823. doi:10.1002/gepi.22083
[7] Plomin, Robert, and Sophie Von Stumm. “The new genetics of Intelligence.” *Nature Reviews Genetics*, vol. 19, no. 3, 8 Jan. 2018, pp. 148–159, https://doi.org/10.1038/nrg.2017.104.
[8] Purcell, Shaun, et al. “PLINK: A tool set for whole-genome association and population-based linkage analyses.” *The American Journal of Human Genetics*, vol. 81, no. 3, Sept. 2007, pp. 559–575, https://doi.org/10.1086/519795.
[9] Savage, Jeanne E., et al. “Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence.” *Nature Genetics*, vol. 50, no. 7, 25 June 2018, pp. 912–919, https://doi.org/10.1038/s41588-018-0152-6. WAS 4
[10] Stahl, Eli A et al. “Genome-wide association study identifies 30 loci associated with bipolar disorder.” *Nature genetics* vol. 51,5 (2019): 793-803. doi:10.1038/s41588-019-0397-8
[11] Vilhjálmsson, Bjarni J et al. “Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores.” *American journal of human genetics* vol. 97,4 (2015): 576-92. doi:10.1016/j.ajhg.2015.09.001
[12] Zhang, Martin Jinye; Hou, Kangcheng (2022). scDRS_data_release_030122. figshare. Dataset. https://doi.org/10.6084/m9.figshare.19312583.v1

Appendix

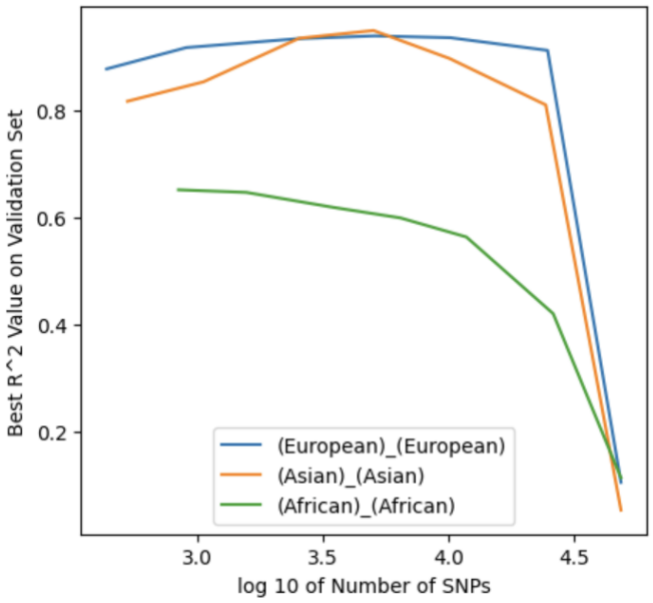


Figure 3: R^2 Values on Validation Sets with P-Value Thresholding. Here, the r^2 value on validation sets were plotted to determine the best model for the testing set of the same continental data. There were 7 p-value thresholds [1.0, 0.5, 0.2, 0.1, 0.05, 0.02 and 0.01] to limit the number of SNPs to use. It is seen that with increasing number of SNPs, the r^2 decreases below the initial heritability value of 0.2