Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2017 doi:10.20944/preprints201701.0127.v1

Polygenic Selection, Polygenic Scores, Spatial Autocorrelation and Correlated Allele Frequencies. Can We Model Polygenic Selection on Intellectual Abilities?

Davide Piffer

Department of Psychology, Ben Gurion University of the Negev, Beer Sheva, Israel; Email: piffer@post.bgu.ac.il

Abstract: The majority of polygenic selection signal of educational attainment GWAS hits is confined to a handful of SNPs within genomic regions replicated across GWAS publications. A polygenic score comprising 9 SNPs predicts population IQ (r=0.9), outperforming 99.9% of the polygenic scores obtained from sets of random SNPs. Its predictive power remains unaffected after controlling for spatial autocorrelation. Even random polygenic scores are moderate predictors of population IQ, and their predictive power increases logarithmically with the number of SNPs, indicating an exponential reduction in noise.Thus, the predictive power of polygenic scores has to be scaled in proportion to the number of SNPs composing them.

Keywords: GWAS; educational attainment; polygenic selection

Introduction

Piffer [1] identified 9 genomic loci that were replicated across the three largest GWAS of educational attainment published to date [2-4]. The 9 loci contain GWAS significant alleles that were found to be in strong LD (r>0.8). One locus was replicated across three GWAS [2-4] and the same SNP (rs9320913) was found in two of them[2,4]. The population frequencies of the 9 pairs (one member belonging to each GWAS publication) of alleles were highly correlated (r=0.919), hence the SNPs published in [4] were used. Thus, this set of 9 SNPs was considered the best candidate for analysis of natural selection on educational attainment and related phenotypes (e.g. general cognitive ability or gca).

Another set of 7 SNPs that reached significance in the UK Biobank and another database was identified by [4]. Population IQ estimates were obtained from [1].

This paper has several aims: to test the presence of correlated frequencies among GWAS hits and the predictive power of polygenic scores (average frequencies of GWAS alleles with positive effect), independently of spatial autocorrelation. A null model will be built using a large set of random SNPs and the polygenic selection model will be tested against it.

Methods and Results

A simulation was performed using a random dataset (a large sample (N=7369) of random unlinked (minor alleles, downloaded from 1000 Genomes, phase 3) matched SNPs frequencies (r<0.1 among EUR). Matching was carried out using SNPSNAP[5], by feeding the 9 SNPs and setting LD r^2 <0.1. A correlation between all the variables (i.e. SNPs

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2017 doi:10.20944/preprints201701.0127.v1

frequencies) was run on the entire dataset. This produced a very large correlation matrix (N=27,147,396 for the lower triangle).

The average correlation coefficient was 0.058 (SD=0.537). The slightly positive value is likely due to the differential representation of minor alleles among populations. As the SD value for the smaller samples tended to be lower (0.45-0.49), the larger SD for the random set (0.537) was used to compute corrected Z scores.

The same analysis was applied to the educational attainment GWAS hits (table 1).

It is clear that there is very little signal in the GWAS hits and it seems to be concentrated within a small subset of SNPs, possibly the 9 replicated loci and the 7 cross-replicated hits (r=0.278 and 0.125, respectively). Similar null results (table 1) were obtained for the 600+ SNPs from the largest GWAS of human height [6].

Since there was some LD between the 9 quasi-replicated SNPs, only one SNP per chromosome was retained, yielding 6 unlinked SNPs. This gave a "pure" (LD-free) measure of correlation. The average correlation was slightly higher than for the 9 SNPs (r=0.343), implying that LD did not produce the correlation among the full set of 9 hits.

	Average r	**SD**	**SE**	**Percentile**
9 replicated loci	0.278	0.458	0.076	
6 replicated loci (linkage pruned)	0.343	0.342	0.138	
7 cross-replicated SNPs	0.125	0.455	0.117	
74 EduYears SNPs (Okbay et al., 2016)	0.003	0.496	0.011	
161 SNPs, Pooled meta-analysis (Okbay et al., 2016)	0.004	0.495	0.004	
691 Height SNPs (Woods et al, 2014)	0.006	0.514	0.006	

Correlation between polygenic scores and population IQ

The polygenic score computed using the 9 SNPs was highly correlated (r=0.9) to an estimate [6] of average population IQ (fig. 1). An empirical simulation was run using 819 PS computed from groups of 9 SNPs taken from the random dataset. The average correlation between population IQ and the random polygenic scores was 0.22 (N=819). The slightly positive correlation can be interpreted as an effect of spatial/phylogenetic autocorrelation [1]. Indeed, the correlation between population IQ and the polygenic score of all the random SNPs (N=7369) was r=0.425, suggesting again the presence of phylogenetic autocorrelation. The increase in the correlation coefficients moving up from low (9)to high SNPs number (7k+) is due to the reduction in the noise associated with each SNP. Because the correlation coefficients were not normally distributed (fig. 2), z-score computation was not appropriate. Hence, the percentile corresponding to a correlation coefficient r=0.9 was found to be 99.9% (using the 819 random polygenic scores), implying that the result is highly significant (produced only 1 out of 1000 times using random sets of SNPs).

Figure 1. Correlation between population IQ and polygenic score.

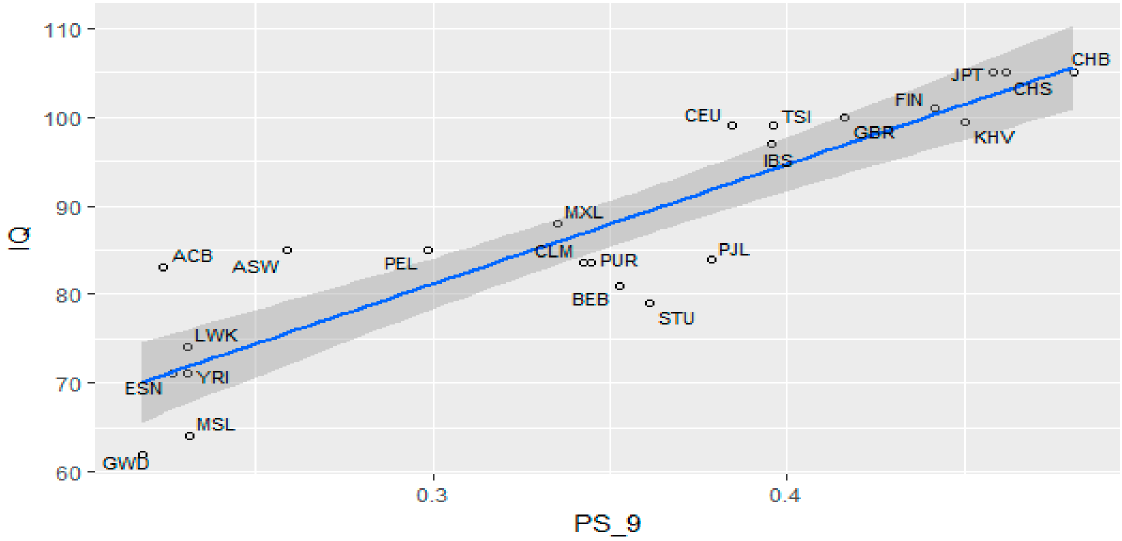
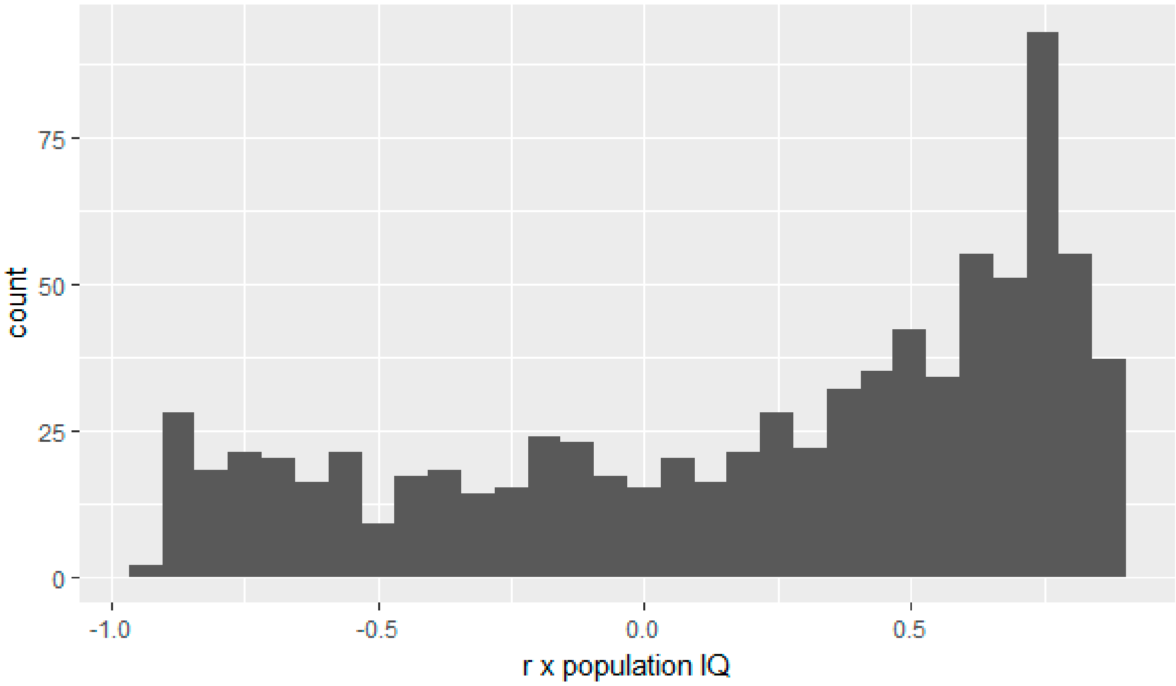


Figure 2. Distribution of correlation coefficients (r population IQ x sets of nine random SNPs).



Partialling out spatial autocorrelation using multiple regression

Population IQ was regressed on the "random PS" (computed using the 7k+ random SNPs) and the 9 GWAS hits PS. The model was significant (F=43.06, p=5.649e-08, Adj R^2=0.793. The random PS had no predictive power (B=0.037),whereas the 9 GWAS hits PS had strong predictive power (Beta=0.884).

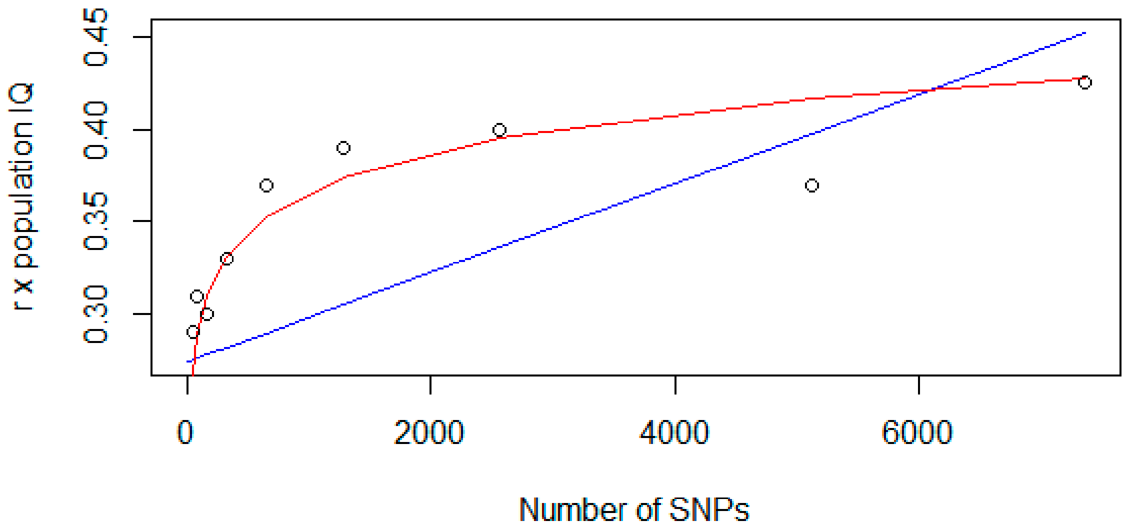
Table 2. Multiple regression

Variable	Beta	t	sig	VIF
PS 9 GWAS hits	0.883	8.182	8.21e-08	1.237
Random PS	0.038	0.351	0.729	1.237

How does the number of SNPs affect the correlation between average population phenotypic value (i.e. IQ) and polygenic score?

A polygenic score computed from a higher number of SNPs should reduce the noise in the data. In order to test this model, polygenic scores were created using different number of SNPs over the random SNPs dataset. The correlation of each polygenic score with population IQ was computed. The data followed a logarithmic function (figure 3), and the log regression model was compared to a linear model: the former had a much better fit to the data (Adjusted R-squared: 0.9388, F-statistic: 185.2 on 1 and 11 DF, p-value: 3.16e-08) compared to the latter (F-statistic: 7.568 on 1 and 11 DF, p-value: 0.019.

Figure 3. Relationship between number of SNPs and predictive power.



Discussion

The method of correlating allele frequencies seems to have low power to detect signals of polygenic selection and picks up the signal only for the most powerful genetic loci. Simply computing polygenic scores (average of allele frequencies with positive GWAS beta) seems a more powerful method to detect polygenic adaptation. The polygenic score obtained from 9 quasi-replicated SNPs looks like a good candidate for estimating selection strength on educational attainment, as shown by it outperforming 99.9% of the polygenic scores obtained from polygenic scores composed of (nine) random SNPs and its robustness to tests controlling for spatial autocorrelation.

The reliability of polygenic scores increases as a logarithmic function of the number of SNPs, even with random SNPs, hence independently of selection signal and this is due to the reduction in noise. The correlation asymptotically approaches a value around 0.45 (which can be interpreted as the degree of spatial autocorrelation in the dataset). Thus, the predictive power of polygenic scores has to be scaled in proportion to the number of SNPs composing them (i.e. simulations need to be performed using the same number of SNPs).

References:

1.Piffer, D. Evidence for Recent Polygenic Selection on Educational Attainment Inferred from GWAS Hits. Preprints 2016, 2016110047 (doi: 10.20944/preprints201611.0047.v1).

2.Rietveld, C.A., Medland, S.E., Derringer, J., Yang, J., Esko, T., Martin, N.W., et al. (2013). GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science,* 340, 1467-1471. doi: http://doi.org/10.1126/science.1235488

3.Davies, G., Armstrong, N., Bis, J. C., Bressler, J., Chourake, V., Giddaluru, S., et al. (2015). Genetic contributions to variation in general cognitive function: a meta-analysis of genome-wide association studies in the CHARGE consortium (N = 53949). Molecular Psychiatry, 20, 183–192. http://dx.doi.org/10.1038/mp.2014.188.

4.Okbay, A., Beauchamp, J.P., Fontana, M.A., Lee, J., Pers, T.H., et al. (2016). Genome-wide association study identifies 74 loci associated with educational attainment. *Nature*, doi:10.1038/nature17671

5.Tune H. Pers, Pascal Timshel, Joel N. Hirschhorn; SNPsnap: a Web-based tool for identification and annotation of matched SNPs. *Bioinformatics* 2014; 31 (3): 418-420. doi: 10.1093/bioinformatics/btu655

6.Wood AR, Esko T, Yang J, *et al.*: Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet.* 2014; **46**(11): 1173–86

© 2017 by the author; licensee *Preprints*, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons by Attribution (CC-BY) license (http://creativecommons.org/licenses/by/4.0/).