Thresholding for population-level polygenic scores to maximize predictive accuracy: IQ and Educational abilities.

Davide Piffer
Ulster Institute for Social Research, UK
Email: pifferdavide@gmail.com

Abstract

Polygenic scores (PGS) are being used to predict group-level traits across time and space, hence proving useful to detect recent selection. The aim of this study is to investigate the impact of different p value thresholds on signal to noise ratio and predictive validity of polygenic scores, using the largest GWAS of educational attainment and cognition to date.

Signal to noise ratio linearly decreases with p value, but this phenomenon is limited to Eurasians. There is a linear degradation of validity and population differentiation in allele frequencies with higher p values. However, compounded polygenic scores have a quadratic relationship with p value. A thresholding at or below the conventional GWAS significance ($p<5*10^{-8}$) seems to maximize validity, corroborating earlier results.

The highest correlation with population IQ is achieved by the Highest Math Class completed and the EDU MTAG PGS (r= 0.90 and 0.89, respectively). Using random SNPs, it is shown that correlations of this magnitude occur only once out of 46k trials. A table provides an empirical estimate of the rarity of the correlation coefficients and it is shown that they are a function of PGS size.

Finally, an analogy between the noise contained in polygenic scores and physical instruments is put forward.

Introduction

Systematic differences in allele frequencies between populations have been studied in relation to phenotypes and environmental variables. However, a main driver of allele frequency differences between populations is genetic drift, which is a random process due to sampling error. In the absence of strong selective forces, drift adds considerable noise to population-level polygenic scores.

Recently, I showed that the predictive validity of polygenic scores for educational attainment is a function of both number of SNPs and significance value, so that increasing the number of SNPs improves the predictive validity only up to a certain threshold, and inclusion of low-significance SNPs adds too much noise to the data, resulting in lower predictive validity (Piffer, 2019).

Predictive validity was measured as the correlation between population polygenic scores and average IQ. The aim of this study is to directly quantify the signal and noise in the data by adopting methods commonly used in engineering (i.e. electrical currents).

That is, by representing allele frequencies as signals, one can apply techniques used in engineering to detect patterns in the data, and to measure signal to noise ratio.

A traditional way to quantify noise in the data is to calculate the standard deviation of the samples of the signal; the mean is the average value of the signal and the signal to noise ratio is equal to the mean divided by the standard deviation (Smith, 1999).

A prediction of the polygenic model of the evolution of intelligence is that the signal-to-noise ratio (SNR) at the population level will be stronger for the SNPs that have higher GWAS significance, that is to say that are more likely to be predictive of cognitive abilities at the individual level. In fact, the standard deviation in a population of the frequencies of higher significance (i.e. lower p value) SNPs should have a lower relative value (mean/sd) compared to SNPs with lower significance. This is due to selective pressure on a phenotype acting homogeneously on the SNPs that influence it. The average frequency of thousands of alleles contains more signal than the average of a handful, which mostly consists of noise.

This effect should also be stronger for the GWAS-reference population, and decay with genetic distance, and it can be quantified as the interaction term of population and SNP significance regressed on SNR.

In turn, the correlation between SNPs and the degree of population differentiation in allele frequencies should increase with higher GWAS significance and SNR.

Methods

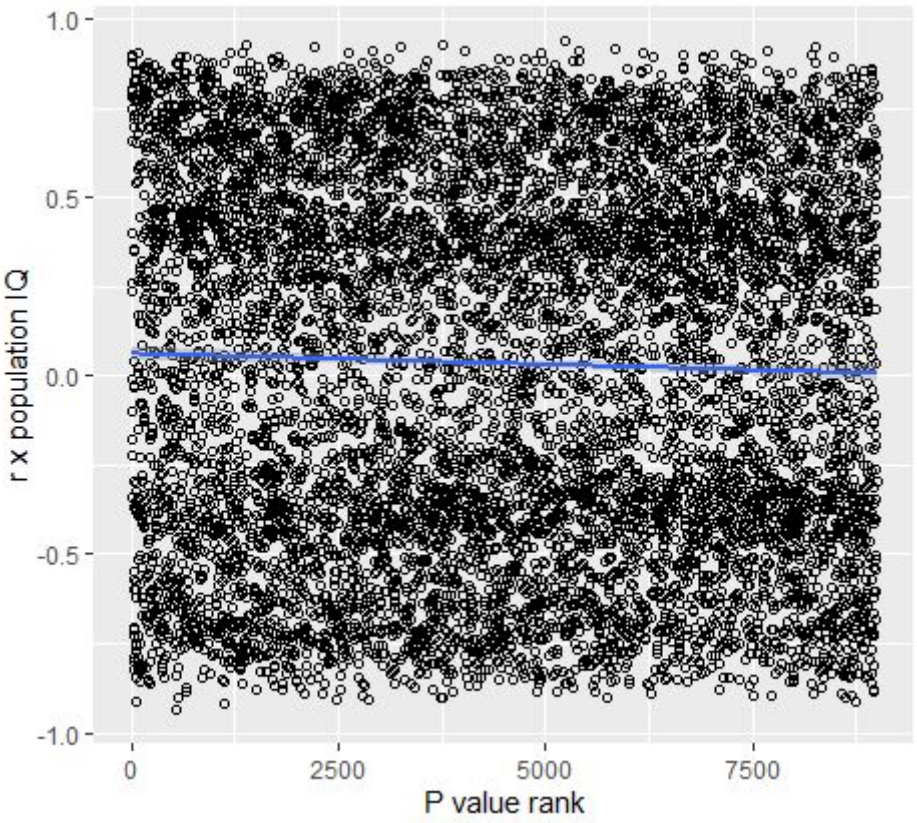
The lead SNPs (N=9034) from the multi-trait analysis of Edu Years, cognitive performance, self-reported math ability and highest math class taken (EA MTAG) were used to perform the analysis because of the high predictive validity (Lee et al., 2018) and the high correlation with population IQ (r= 0.887). In total, 9001 SNPs were found in the 1000 Genomes database for the 26 populations.

One-way ANOVA will be used to measure the amount of differentiation in allele frequencies between populations. Four non-admixed populations (BEB= Bengali in Bangladesh, CEU= Northwestern Europeans from Utah, YRI= Yorubans in Nigeria, CHS= Southern Chinese).

Results

Since p values were extremely right-skewed, they were rank-transformed. The average correlation between SNP and population IQ was r= 0.0343. The corresponding Z score is 6.28 (mean(r_xIQ)/(sd(r_xIQ)/sqrt(9001)). The Spearman-rank correlation between SNP p value and the trait-increasing allele frequency was r= -0.03 (N= 9001, p= 0.0052), implying that the SNPs with lower p value (higher GWAS significance) have higher correlation with population IQ (figure 1). This replicates a previous finding using a different set of SNPs (Piffer, 2019).

Figure 1. SNP's correlation with population IQ by GWAS p-value.

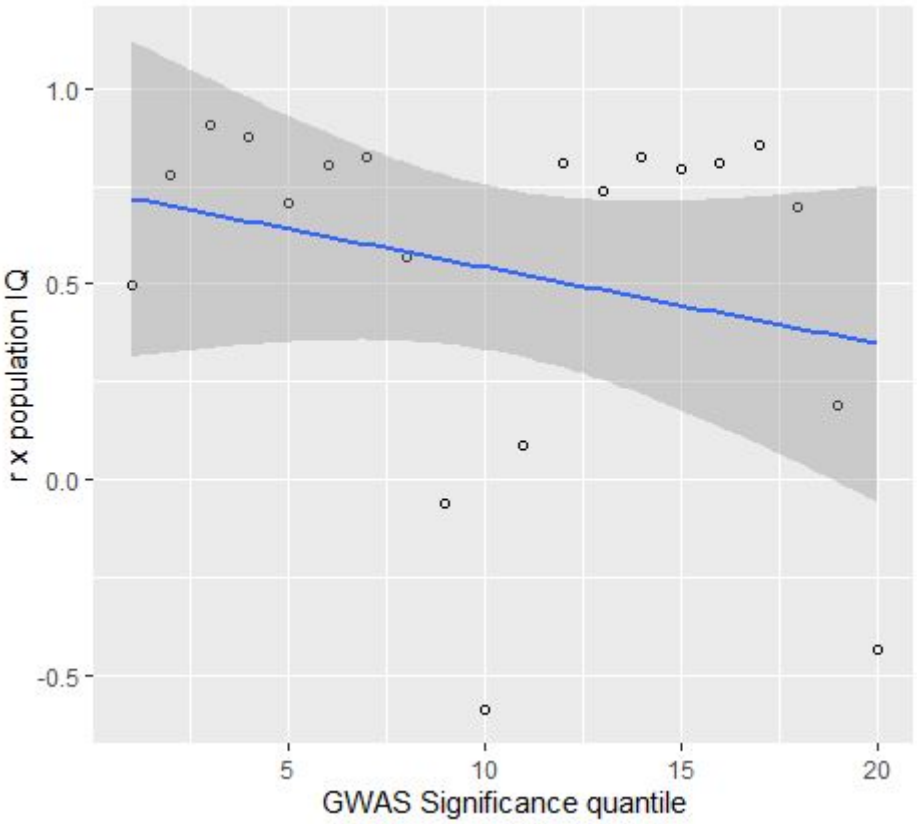


Polygenic scores divided into discrete quantiles

The SNPs were divided into 20 quantiles according to p value. The correlation between quantile rank and population IQ was slightly negative but not significant (Spearman's r= -0.2, p= 0.39).

Only a minority of quantiles (3/20) had a negative correlation with IQ (figure 2). The average correlation coefficient was r= 0.533.

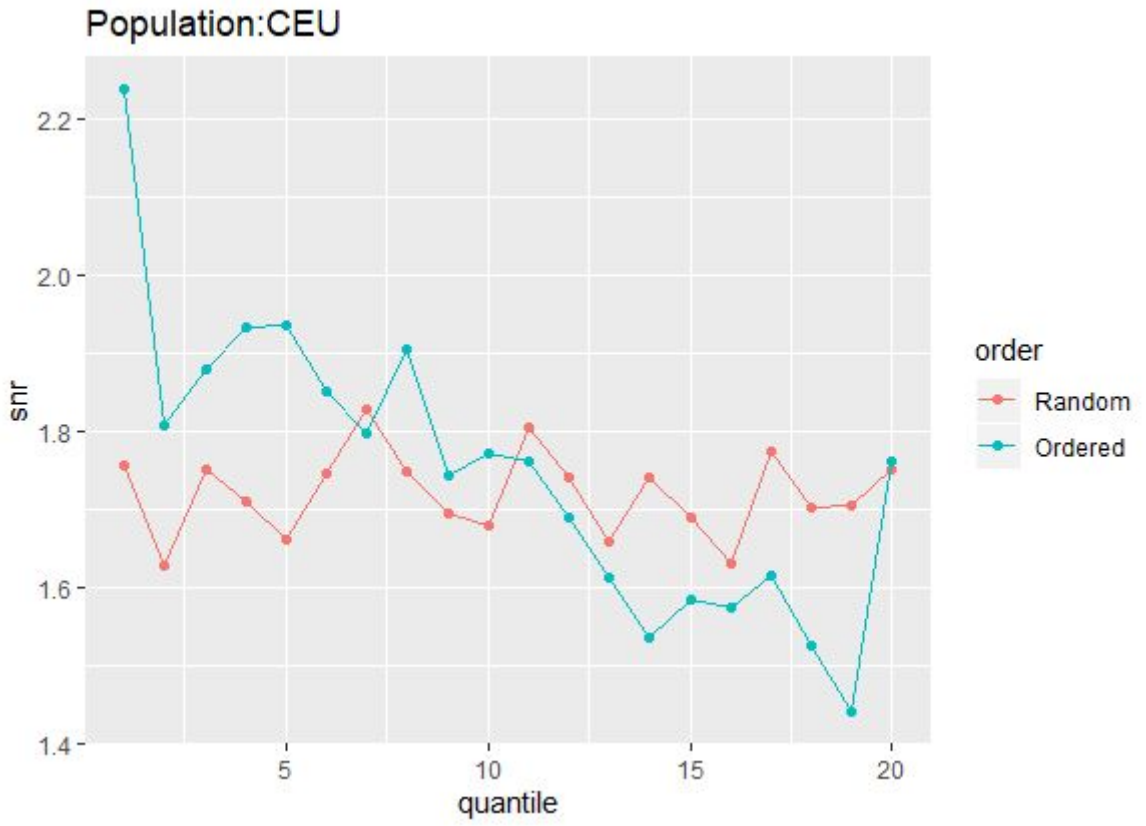
Figure 2. Relationship between p-value quantile and population IQ.



Signal to noise ratio

The average frequency and SD of each quantile were calculated for CEU, YRI and CHS, and the SNR was computed (SNR= μ /σ). As a control, the SNPs were also randomly shuffled so as to create quantiles with randomly distributed p-values.
SNR and quantile were negatively correlated in CEU (r= -0.85, p<0.001, N=20). In the random group, the correlation with quantile was not significant (r= -0.14, p= 0.56, N= 20).
This interaction can be visualized in figure 3.

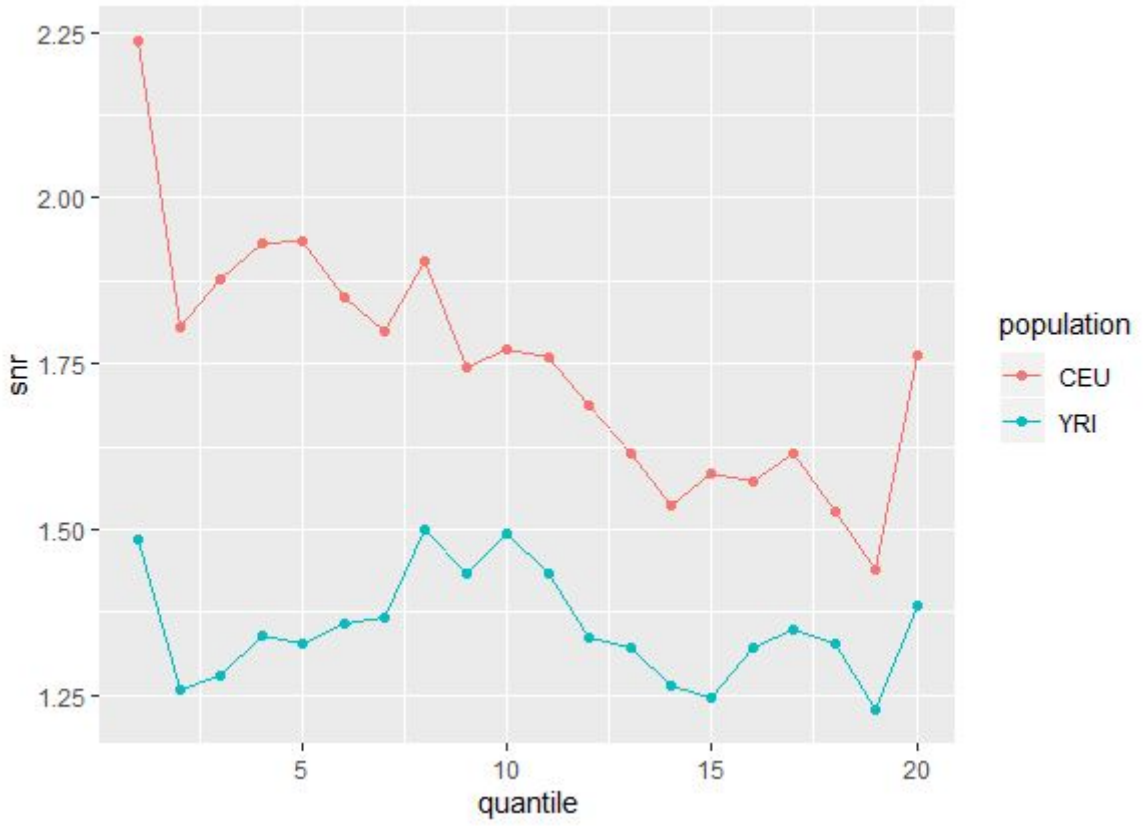
Figure 3. Signal to noise ratio by quantile for random quantiles and p-value quantiles.



Multiple linear regression was used to compute the significance of this interaction by regressing SNR on quantile and group (Ordered by p value or random). A significant model emerged (F= 22.29; Adj. R-squared= 0.621, p= 2.478e-08). The interaction term (quantile: order) was significant (p= 7.75e-08).
However, the correlation between quantile and SNR was significant for CHS too (r= -0.8, p= 0, N=20) but it was not significant for YRI (r= -0.24, p= 0.31). In fact, the quantile: order interaction in the regression model was not significant for YRI (p= 0.666).

Another regression was run using the p-value ranked SNPs with SNR as the dependent variable and population (YRI, CEU) + quantile as the predictors. A significant model emerged (F= 74.7; Adj. R-squared: 0.8501; p= 1.581e-15). The interaction term was significant (p= 0.0001). This interaction can be visualized in figure 4.

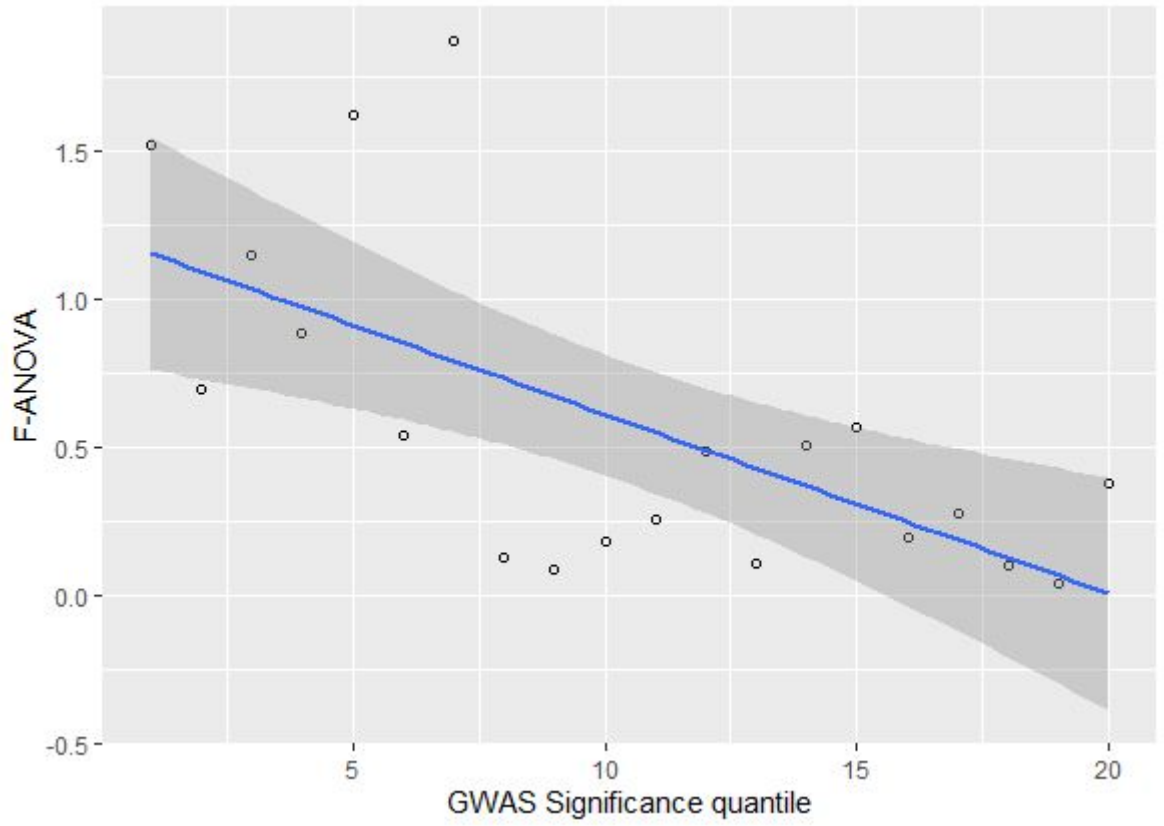
Figure 4. Signal to noise ratio by p-value quantile for YRI and CEU



A similar analysis (not reported here) was carried out with East Asians and revealed a pattern closer to that of the CEU population.

Another measure of selection pressure is population differentiation in allele frequencies. ANOVA was used to compute the difference in allele frequencies between four populations for each quantile.
There was a negative correlation between quantile and F value (r= -0.64, p= 0.0026) (fig. 5).

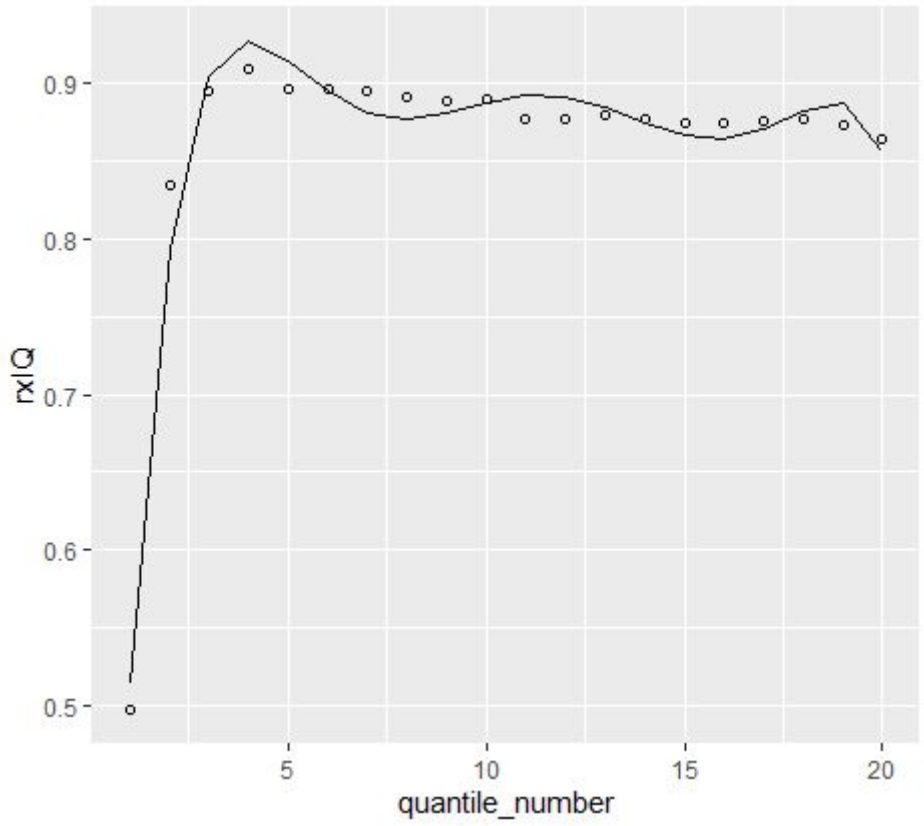
Figure 5. Correlation between F value and p-value quantile.



Compounded Polygenic scores

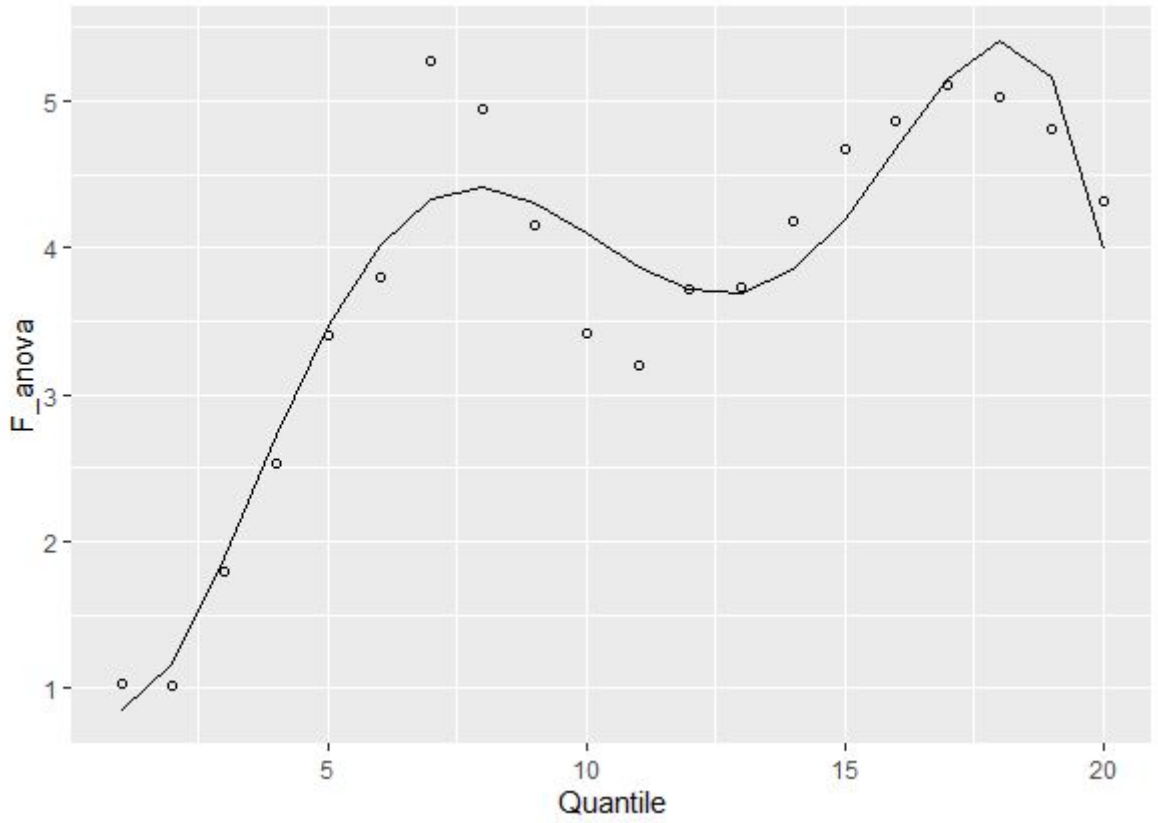
Polygenic scores were created starting from the lowest quantile (highest significance) and incrementally adding SNPs from the higher quantiles. The correlation with population IQ was tracked across quantiles, showing a quadratic pattern (figure 6). A polynomial regression model was fit to the data

Figure 6. Relationship between p-value quantile and correlation with population IQ.



The correlation with IQ peaks around the 4th quantile ($p < 2.20E-10$) to level off and then decrease again. This relationship was best explained by a 6th degree polynomial ($F = 72.29$, Adjusted R-squared: 0.971, $p = 3.137e-09$).
Similarly, the relationship between quantile and F value in ANOVA best fit a 5th degree polynomial ($F = 26.9$, Adjusted R-squared: 0.8721, $p = 1.035e-06$) (figure 7).

Figure 7. Relationship between p-value quantile and F value.



Correlation between different polygenic scores for 1KG populations

Figure 8 shows the correlation matrix ordered with hierarchical clusters of the different polygenic scores of cognitive abilities from the Lee et al. 2018 GWAS, weighted and unweighted, for whole set and GWAS significant SNPs ("sig8"). The PGS computed from the GWAS significant SNPs using Beta weights tend to be more similar to each other, suggesting higher reliability. Excluding phenotypic IQ, the PGS putatively causal SNPs are the most closely related to the others. However, including IQ, the Highest Math class and the EA MTAG (GWAS significant) PGS have the highest reliability (figure 9).

Figure 8. Correlation matrix (Polygenic scores).

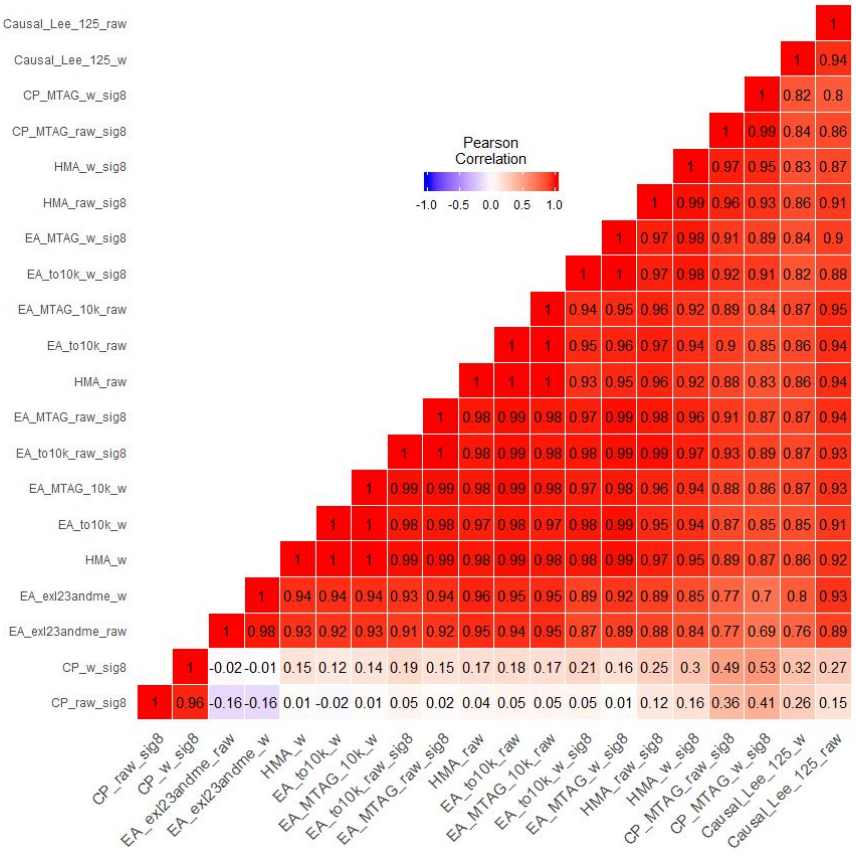
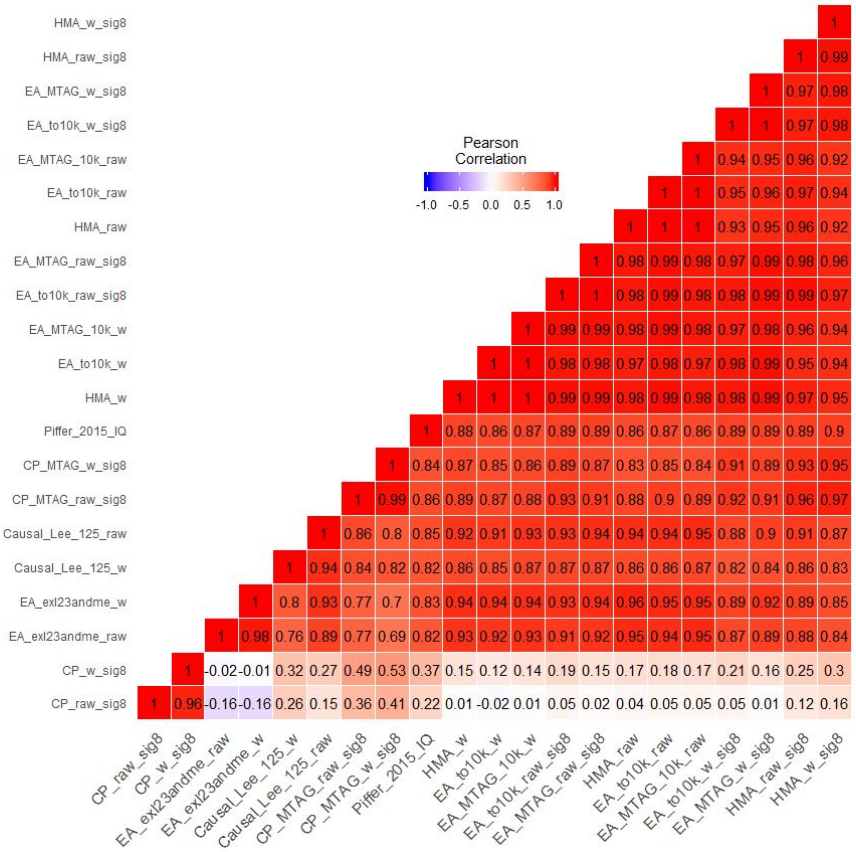


Figure 9. Correlation matrix (Polygenic scores + IQ).



The polygenic scores for the phenotypes with the highest reliability are reported in figure 10 and 11.

Figure 10. Polygenic scores (weighted) for EA MTAG ($p<5*10^{-8}$).

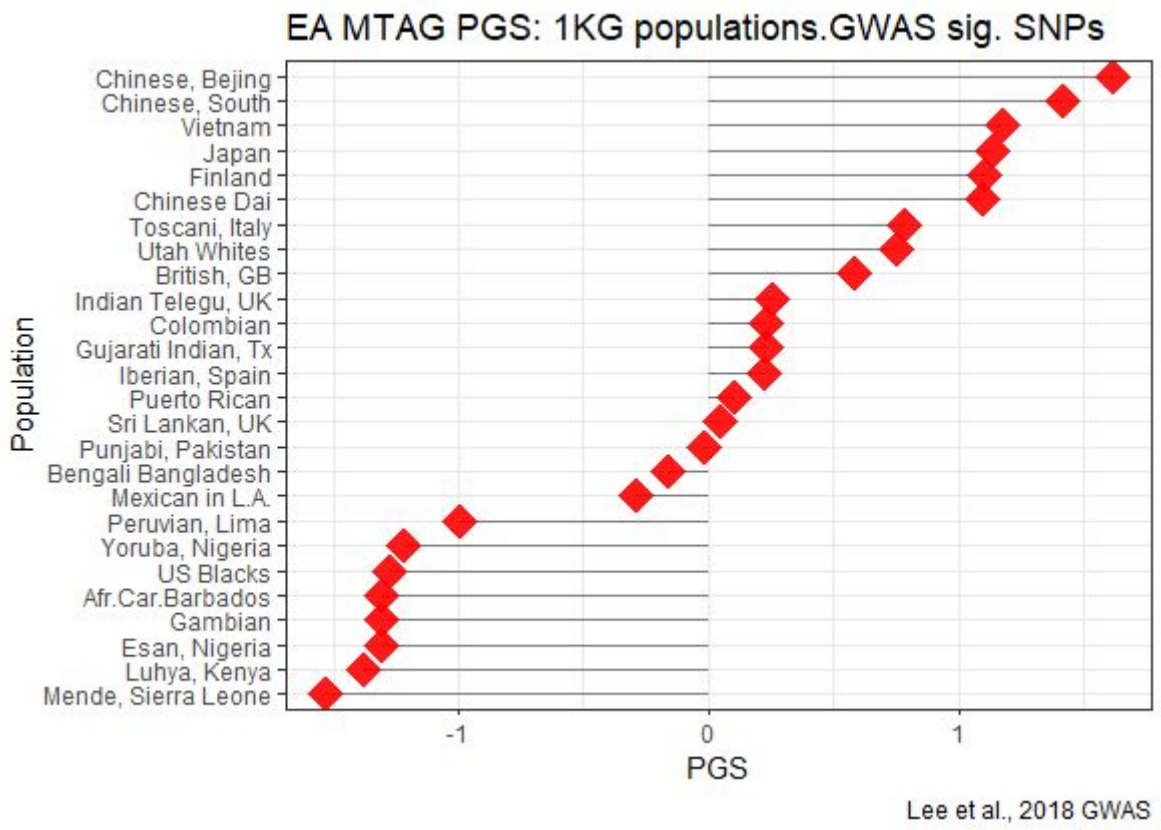
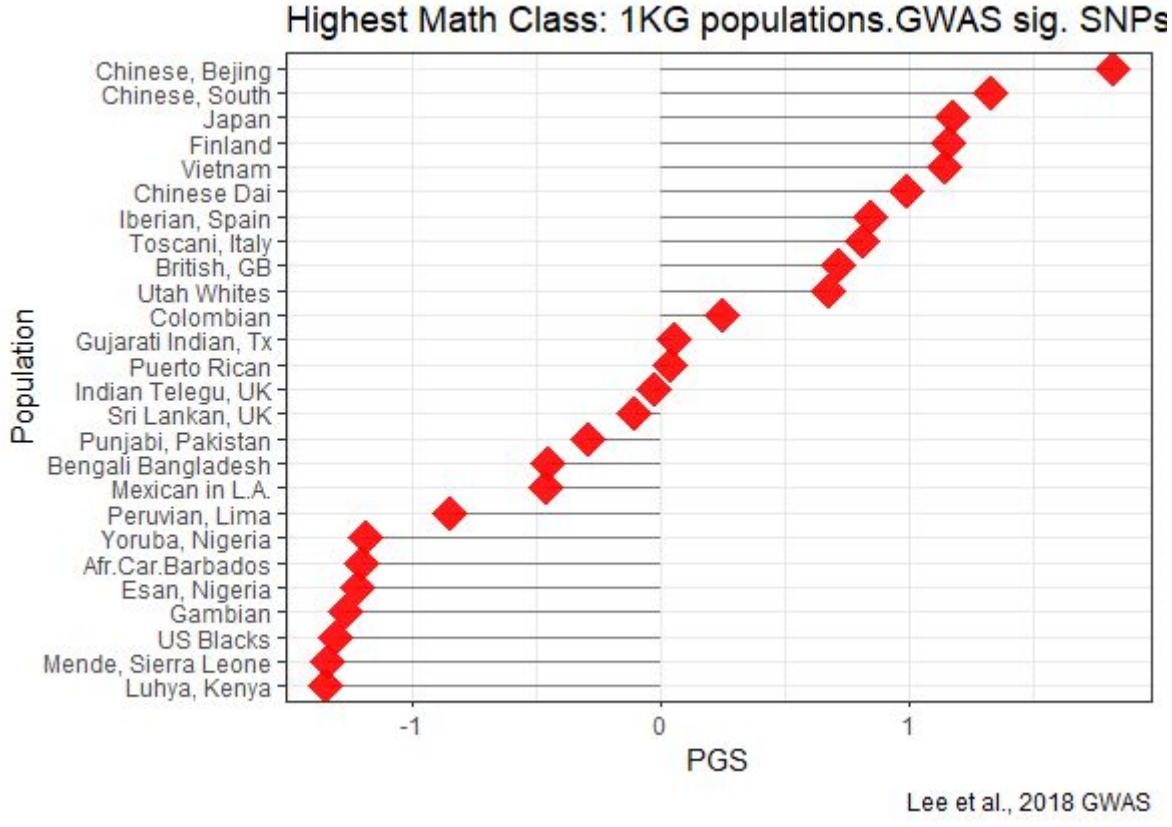


Figure 11. Polygenic scores (weighted) for Highest Math Class completed ($p<5*10^{-8}$).



With the exception of the CP PGS, the correlations among the PGS and of those with IQ are mostly in excess of r= 0.8.

Number of SNPs and significance of correlation.

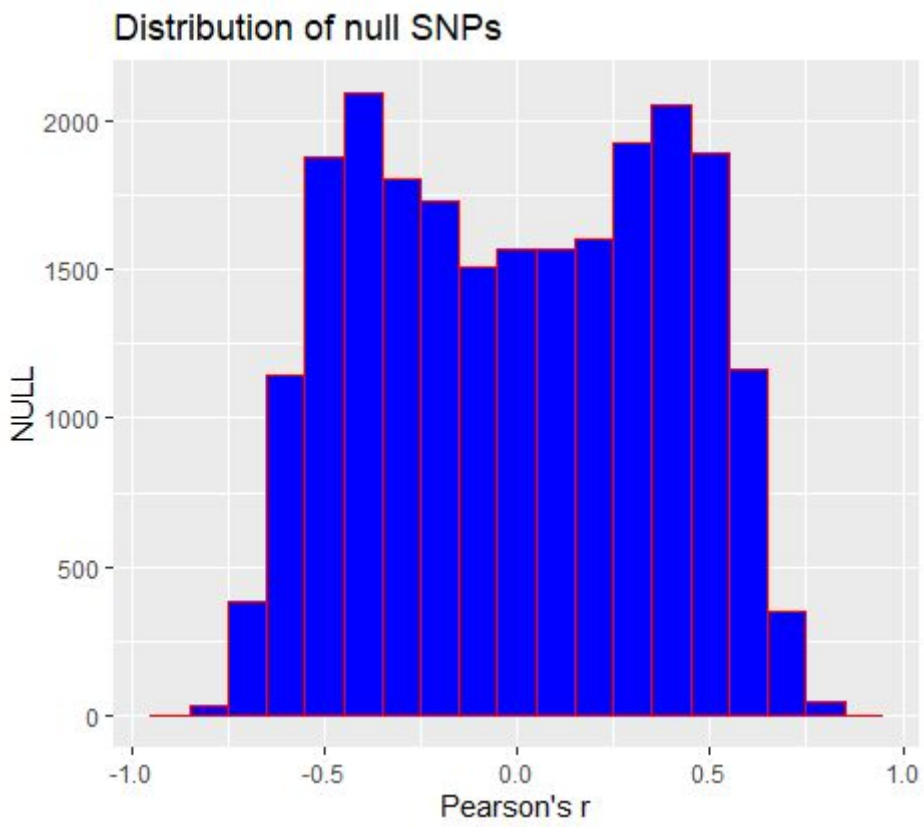
A simulation with 2.2 million randomly matched SNPs was run in order to compute the rarity of the correlation coefficients as a function of number of SNPs. Since polygenic scores consisting of fewer SNPs have higher sample noise, higher correlation coefficients occur more frequently as a result of chance. This can be seen in table 1. Sample noise goes down with increasing number of SNPs in a quadratic fashion.

Table 1. Rarity of correlation coefficients (r x IQ) for random SNPs as a function of PGS size.

N SNPs	r>0.6	r>0.65	r>0.7	r>0.75	r>0.8	r>0.85	r>0.9	N PGS
50	1520	751	321	95	25	2	1	46386
100	897	402	168	51	14	2	0	23193
500	182	67	19	4	0	0	0	4639
1000	82	25	19	2	1	0	0	2320
1500	46	18	3	2	0	0	0	1547
2000	40	10	2	1	0	0	0	1160
2500	31	11	3	1	0	0	0	928
3000	25	7	1	0	0	0	0	774
3500	21	6	1	1	0	0	0	663
5000	13	4	2	0	0	0	0	464
10000	8	1	0	0	0	0	0	232

The distribution of the random PGS's correlation with population IQ is shown in figure 12.

Figure 12. Correlation coefficients (with population IQ) for random PGS (size= 100).



Discussion

Dividing the SNPs into significance quantiles revealed two phenomena: 1) There is internal consistency, with the majority of polygenic scores (17/20) having positive correlations with population IQ (figure 2). The average correlation coefficient was 0.53, much smaller than the correlation between the polygenic score using the full set (r= 0.86) or the GWAS significant hits (r= 0.89): This is due to sample noise being higher for smaller sample sizes.

There is a degradation of signal to noise ratio in allele frequencies with higher p value. This manifests across a range of metrics: 1) A traditional SNR metric, based on the mean/standard deviation ratio: higher p-value PGS have lower SNR, although this phenomenon is restricted to the European (reference) population and to East Asians, but not Africans; 2) The correlation of population IQ to individual SNP (rank ordered) p-value is significantly negative, and to the p-value quantile of polygenic scores (not significant for EA_MTAG but significant for EA (Piffer, 2019); 3) The allele frequency differences between populations are higher for low p-value SNPs (fig. 4).

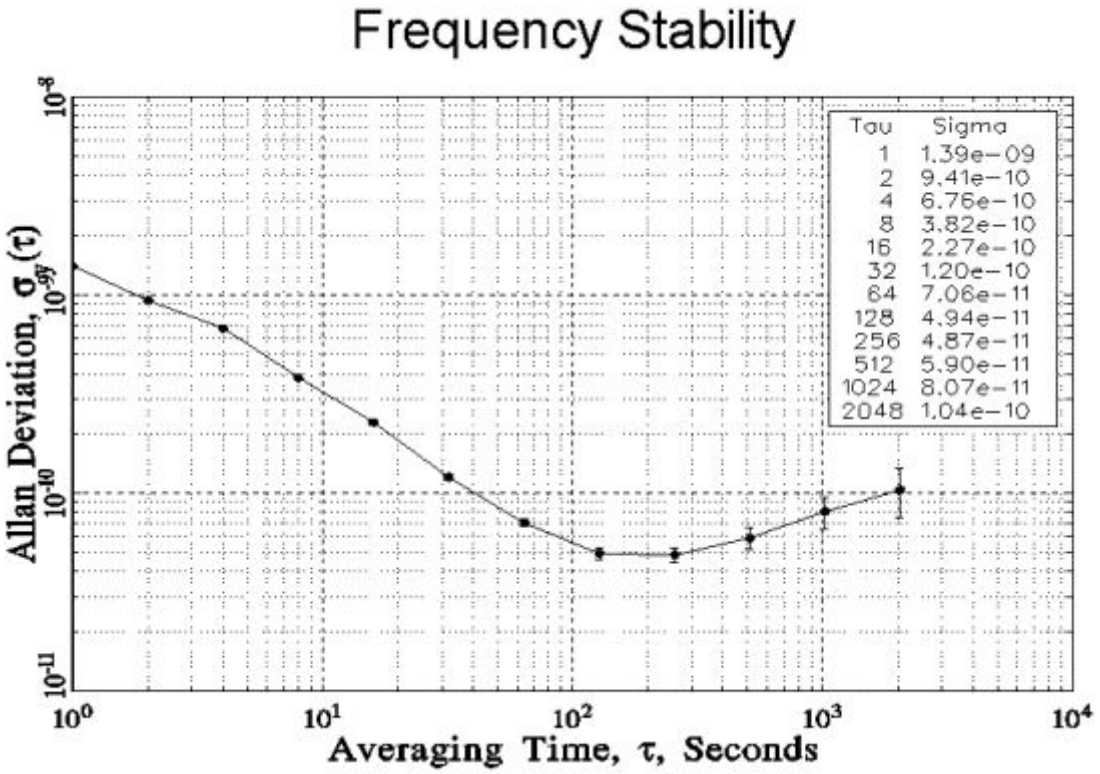
Applying different thresholds increases predictive accuracy up to the fourth quantile (fig. 5), to level-off and then decrease ($p<2.2*10^{\wedge-10}$). It is difficult to establish the optimal threshold solely on this criterion because the average IQ estimates are far from perfect, but the optimal threshold seems to lie not far off from the conventional GWAS significance

threshold ($p<5*10^{\wedge-8}$). This corroborates previous studies which used PGS built with this threshold (Piffer, 2013, 2015, 2019).

This process is explainable in terms of noise reduction with higher number of SNPs, but after some point adding more low p-value SNPs introduces too much noise and not enough signal.

The pattern in noise across SNP number is similar to a phenomenon observed in clocks, oscillators and amplifiers, known as Allan variance. At very short observation time τ, the Allan deviation is high due to noise. After some time, it decreases because the noise averages out. At some point (called the "noise floor") however, the Allan deviation starts increasing again due to nonstationary processess such as aging or random walk (NIST, 2016). This can be seen in the example provided in figure 11.

Figure 11. Example plot of the Allan Deviation of a clock (http://www.nist.gov/pml/div688/grp40/glossary.cfm)



The reduction in sample noise is evident also for random SNPs (table 1). These provide a benchmark for the performance of GWAS SNPs. It can be seen that "smaller" (with lower N) polygenic scores tend to produce higher correlations due to noise alone. Hence, the same correlation coefficient has a higher statistical significance for "larger" polygenic scores. A correlation with population IQ equal to or higher than that yielded by the GWAS polygenic scores computed from the Lee et al. 2018 GWAS (r=> 0.85) occurs very rarely by chance alone. Specifically, using random PGS of size 50 (a conservative approach), a correlation coefficient equal to or higher than the EA MTAG GWAS significant PGS (r= 0.887) occurs once in 46,386 trials, corresponding to a Monte Carlo-corrected p= 4.31e-05.

Supplementary files: https://osf.io/jhqc3/

References

Lee, J.J.; Wedow, R.; Okbay, A.; Kong, E.; Maghzian, O.; Zacher, M.; Nguyen-Viet, T.A.; Bowers, P.; Sidorenko, J.; Karlsson Linnér, R.; et al. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nat. Genet., 50, 1112–1121.

NIST Time and Frequency Services Group, 2016. https://www.nist.gov/pml/time-and-frequency-division/popular-links/time-frequency-z

Piffer, D. (2013). Factor analysis of population allele frequencies as a simple, novel method of detecting signals of recent polygenic selection: The example of educational attainment and IQ. Mankind Q., 54, 168–200.

Piffer, D. (2015). A review of intelligence GWAS hits: Their relationship to country IQ and the issue of spatial autocorrelation. Intelligence, 53, 43–50.

Piffer, (2019). Evidence for Recent Polygenic Selection on Educational Attainment and Intelligence Inferred from Gwas Hits: A Replication of Previous Findings Using Recent Data. Psych, 1, 55-75; https://doi.org/10.3390/psych1010005.

Smith, S.W. (1999). *The Scientist and Engineer's Guide to Digital Signal Processing*. Ch.2, Second Edition, California Technical Publishing.