**Recent polygenic selection for educational attainment**

**Davide Piffer**

**Email: pifferdavide@gmail.com**

**Abstract**

The genetic variants identified by two large GWAS of educational attainment were used to test a polygenic selection model.

Average frequencies of alleles with positive (Beta) effect on the phenotype (polygenic scores) were compared across populations and racial groups using data from 1000 Genomes and ALFRED. Strong correlations between polygenic scores and population IQ were found (r>0.8). Moreover, the polygenic score obtained from the two independent GWAS exhibited a strong correlation (r= 0.83), even after pruning for linkage disequilibrium.

Factor analysis revealed that most alleles loaded on a single factor, which in turn was strongly correlated to population IQ.

Polygenic and factor scores survived control for phylogenetic autocorrelation, although the latter’s net effect on population was stronger (Betas= 0.361 and 0.861, respectively).

Results obtained from ALFRED data were similar and revealed a peak in polygenic and factor scores among East Asians (60.8% and 1.06, respectively) and a nadir among Africans and Native Americans (44.9%,44.1% and -1.47, -0.493, respectively).

Geographic distance from Eastern Africa (assuming an origin of modern humans there) was only weakly predictive of factor and polygenic scores (r= 0.21-0.29).

**Introduction**

The aim of this study is to replicate the studies by Piffer (2015, 2013) that educational attainment and cognition GWAS hits have different frequencies across populations and thus, were subject to different selection pressures. To this end, the hits from the latest GWAS on educational attainment (Davies et al., 2016) will be used in the analysis. This GWAS was carried out using the UK Biobank sample (N=100K+). Over a thousand SNPs reached genome-wide significance (P< 5 x 10-8), but after controlling for linkage disequilibrium (Genotypes were LD pruned using clumping to obtain SNPs in linkage equilibrium with an r2<0.25 within a 200 bp window), a few independent signals were identified (Davies et al., 2016).

It is important that only the alleles with a positive effect on the phenotype are chosen for analysis, otherwise only a signal of population structure will be identified.

**Methods**

*1000 Genomes*

Davies et al. (2016) reported 1115 SNPs reaching GWAS significance, of which 15 were independent signals for educational attainment. 942 SNPs were found on 1000 Genomes. Among the 15 independent signals, one (2:48696432\_G\_A) was missing. Frequencies were calculated from VCF files belonging to the phase 3 data: ftp://[ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20130502/](http://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20130502/)

The polygenic score with the 14 SNPs (independent signals) was checked for LD overlap with a previously reported polygenic score (Piffer, 2015). Two SNPs were found to be in LD (r>0.1) with two previously reported sites (rs55686445 and rs1906252, see Piffer (2015)). A new LD-free polygenic score was created.

Finally, a polygenic score was created by combining all the independent signals (N=16) for educational attainment known to date (3 from Rietveld et al. and 13 from Davies, 2016: rs1906252 was eliminated because in LD with rs9320913 from Rietveld et al., 2013).

Population IQs for the 1000 Genomes populations were obtained from Piffer (2015).

*ALFRED*

Most of the independent signals were not present in ALFRED, thus all the GWAS significant SNPs (Davies et al, 2016: supplementary table 1) were searched. When more than 1 SNP was found in the same region (within 500 Kb), they were lumped together to form an average score. The same procedure was applied to all the GWAS significant hits reported by Rietveld et al. (2013) for college and years of education.

Geographic distances for the ALFRED populations were obtained from an analysis of genetic variation using the HGDP-CEPH data, which showed that 77% of the genetic variance (Fst) could be explained by geographic distance between populations (Lawson Handley, Manica, Goudet and Balloux, 2007). There were 54 and 53 populations in the HGDP and ALFRED datasets, respectively. Of these, 49 were overlapping and were used for calculation of PS distances.

**Results**

The average frequencies for the populations (N=26) from 1000 Genomes are reported in table 1. The correlation of I.S. PS (independent signals polygenic score) with ethnic IQ was r= 0.822. The new polygenic score was highly correlated to the polygenic score (Piffer\_2015) calculated by Piffer (2015) using 9 GWAS hits, r= 0.91. This polygenic score was also correlated (r= 0.88) to the average frequencies of the top 3 hits (Rietveld\_2013) for educational attainment by Rietveld et al. (2013).

The LD-free polygenic score had similar (albeit slightly lower) correlations: r x IQ= 0.774; r x Piffer\_2015= 0.879; r x Rietveld\_2013= 0.83.

The polygenic score created by combining all the independent hits (N=16) had a correlation r=0.843 with population IQ.

The polygenic score obtained from all the hits (N=942) had these correlations: x population IQ (r= 0.705); x Rietveld et al. (2013)’s 3 hits (r= 0.758); x Piffer\_2015(r= 0.87).

**Table 1. Polygenic scores (independent signals) and ethnic IQ**.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Population** | **Ed.att. P.S, I.S.(Davies et al., 2016). N=14** | **All\_Ed\_Att\_2016. N=942** | **PS All Ind. (N=16)** | **Factor All Ind.** | **IQ** |
| Afr.Car.Barbados | 0.419 | 0.411 | 0.361 | -1.385726617 | 83 |
| US Blacks | 0.447 | 0.428 | 0.387 | -1.040795929 | 85 |
| Bengali Bangladesh | 0.516 | 0.566 | 0.461 | -0.009509494 | 81 |
| Chinese Dai | 0.610 | 0.652 | 0.564 | 1.229103674 |  |
| Utah Whites | 0.493 | 0.467 | 0.461 | 0.385060759 | 99 |
| Chinese, Bejing | 0.671 | 0.682 | 0.636 | 1.614207102 | 105 |
| Chinese, South | 0.648 | 0.674 | 0.606 | 1.399490512 | 105 |
| Colombian | 0.500 | 0.512 | 0.462 | 0.155020855 | 83.5 |
| Esan, Nigeria | 0.416 | 0.417 | 0.362 | -1.517446756 | 71 |
| Finland | 0.560 | 0.560 | 0.524 | 0.873423777 | 101 |
| British, GB | 0.526 | 0.494 | 0.499 | 0.568096086 | 100 |
| Gujarati Indian, Tx | 0.498 | 0.550 | 0.457 | 0.064904594 |  |
| Gambian | 0.438 | 0.398 | 0.381 | -1.3799435 | 62 |
| Iberian, Spain | 0.512 | 0.488 | 0.481 | 0.4310518 | 97 |
| Indian Telegu, UK | 0.510 | 0.583 | 0.457 | 0.030182344 |  |
| Japan | 0.652 | 0.679 | 0.625 | 1.422186914 | 105 |
| Vietnam | 0.618 | 0.642 | 0.579 | 1.25233893 | 99.4 |
| Luhya, Kenya | 0.425 | 0.428 | 0.372 | -1.438642624 | 74 |
| Mende, Sierra Leone | 0.416 | 0.421 | 0.364 | -1.40422492 | 64 |
| Mexican in L.A. | 0.499 | 0.555 | 0.455 | 0.01771732 | 88 |
| Peruvian, Lima | 0.477 | 0.559 | 0.430 | -0.00789958 | 85 |
| Punjabi, Pakistan | 0.511 | 0.564 | 0.475 | -0.049972973 | 84 |
| Puerto Rican | 0.489 | 0.480 | 0.451 | 0.026710407 | 83.5 |
| Sri Lankan, UK | 0.506 | 0.564 | 0.454 | 0.070996352 | 79 |
| Toscani, Italy | 0.501 | 0.486 | 0.458 | 0.265676967 | 99 |
| Yoruba, Nigeria | 0.421 | 0.417 | 0.372 | -1.572005998 | 71 |

P.S. = polygenic score.

ANOVA

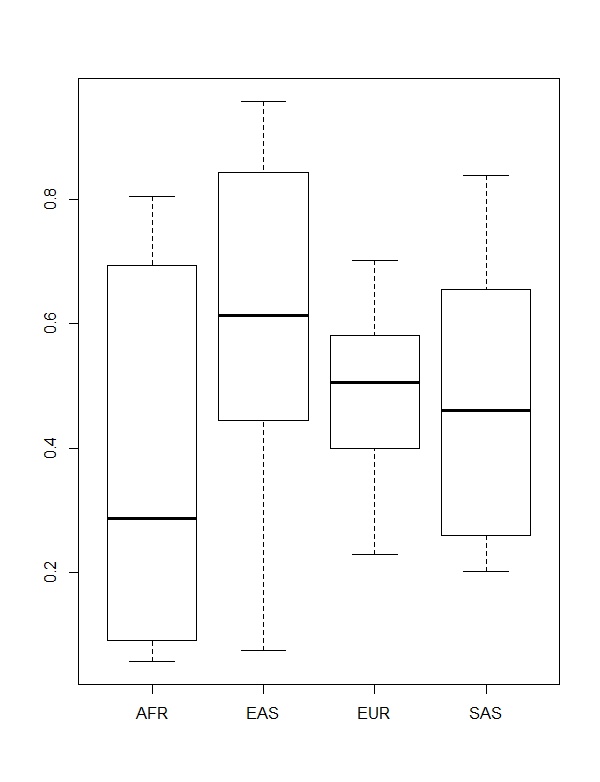
The polygenic score created by combining all the independent hits (N=16) was calculated for the five 1000 Genomes continental groups (African, American, East Asian, European, South Asian).

**Table 2. Frequencies of independent hits by racial group.**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | rs13086611\_T | rs11130222\_A | rs12553324\_G | rs55686445\_C | rs9393692\_G | rs3847225\_C | rs4799950\_G | rs4318611\_A | rs112374913\_A | rs12042107\_C | rs11210887\_A | rs482507\_T | rs7701440\_T | rs9320913\_A | rs11584700\_G | rs4851266\_T | **PS** |
| AFR | 0.403 | 0.793 | 0.156 | 0.220 | 0.804 | 0.354 | 0.582 | 0.763 | 0.076 | 0.797 | 0.094 | 0.626 | 0.087 | 0.180 | 0.057 | 0.064 | **0.379** |
| AMR | 0.186 | 0.705 | 0.363 | 0.380 | 0.718 | 0.380 | 0.552 | 0.771 | 0.421 | 0.539 | 0.476 | 0.576 | 0.454 | 0.340 | 0.094 | 0.297 | **0.453** |
| EAS | 0.075 | 0.877 | 0.465 | 0.595 | 0.894 | 0.457 | 0.810 | 0.958 | 0.432 | 0.632 | 0.687 | 0.877 | 0.786 | 0.415 | 0.317 | 0.554 | **0.614** |
| EUR | 0.336 | 0.585 | 0.430 | 0.366 | 0.577 | 0.413 | 0.653 | 0.702 | 0.425 | 0.502 | 0.664 | 0.567 | 0.526 | 0.509 | 0.229 | 0.387 | **0.492** |
| SAS | 0.202 | 0.839 | 0.458 | 0.331 | 0.761 | 0.263 | 0.625 | 0.817 | 0.332 | 0.602 | 0.463 | 0.687 | 0.476 | 0.247 | 0.255 | 0.247 | **0.475** |

To test the hypothesis that the group means were significantly different, a one-way Anova was run using 4 groups. The American cluster was excluded because it comprises highly mixed populations (e.g. Mexicans in L.A.; Puerto Ricans, etc.).

**Figure 1. Boxplot (1000 Genomes data).**



ANOVA showed that the effect of racial group on allele frequencies was significant, F= 2.77,

p=0.049.

However, Bartlett’s test showed that the assumption of equality of variances was violated (Bartlett’s K-squared= 9.249, df=3, p= 0.026). Since the aov function in R assumes equality of variances, the function “oneway.test” was used, which applies Welch’s correction for non-homogeneity. This time, the result was not significant according to the conventional alpha level (F= 2.015, p= 0.132).

Kruskal-Wallis rank sum test (“kruskal.test”) produced a marginally significant effect (Kruskal-Wallis chi-squared = 7.0598, p-value = 0.070).

**Table 3. Tukey post-hoc test**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Δ** | **C.I. 95%** | **P adj.** |
| EAS-AFR | 0.236 | 0.018 - 0.453 | 0.028 |
| EUR-AFR | 0.113 | -0.104 - 0.331 | 0.517 |
| SAS-AFR | 0.097 | -0.121 - 0.314 | 0.644 |
| EUR-EAS | -0.122 | -0.339 - 0.095 | 0.451 |
| SAS-EAS | -0.139 | -0.357 - 0.078 | 0.337 |
| SAS-EUR | -0.017 | -0.234 - 0.201 | 0.997 |

*Factor Analysis of allele frequencies*

A factor analysis (function “fa”, package “psych”) was carried out using Ordinary Least Squares to find the minimum residual solution. The proportion of variance explained was 0.54. Factor scores are reported in table 1 (column 5). This factor was correlated to population IQ (r= 0.89).

**Table 4. Factor loadings (structure matrix)**

|  |  |
| --- | --- |
| **SNP** | **Factor Loading** |
| rs13086611\_T | -0.76 |
| rs11130222\_A | 0.03 |
| rs12553324\_G | 0.87 |
| rs55686445\_C | 0.88 |
| rs9393692\_G | 0.03 |
| rs3847225\_C | 0.43 |
| rs4799950\_G | 0.7 |
| rs4318611\_A | 0.52 |
| rs112374913\_A | 0.89 |
| rs12042107\_C | -0.67 |
| rs11210887\_A | 0.93 |
| rs482507\_T | 0.53 |
| rs7701440\_T | 0.98 |
| rs9320903\_A | 0.75 |
| rs11584700\_G | 0.85 |
| rs4851266\_T | 0.95 |

ALFRED

9 independent SNPs were found in the Davies et al. (2016) dataset, and 3 independent SNPs from the Rietveld et al. (2013) dataset. The average frequency was calculated for the 53 populations separately and then together for the two sets of SNPs (table 5).

The Rietveld et al. (2013) and Davies et al. (2016) polygenic scores were moderately correlated (r= 0.616).

**Table 5. Polygenic and factor scores: ALFRED populations.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Population** | **Davies 2016** | **Rietveld 2013** | **Total PS** | **Factor Tot** |
| Bantu SW | 0.487 | 0.293 | 0.409 | -1.842 |
| Bantu Kenya | 0.471 | 0.237 | 0.377 | -1.907 |
| Biaka | 0.566 | 0.307 | 0.462 | -1.402 |
| Mandenka | 0.507 | 0.313 | 0.430 | -1.249 |
| Mbuti | 0.522 | 0.320 | 0.441 | -1.637 |
| Mozabite | 0.523 | 0.310 | 0.438 | -0.469 |
| San | 0.456 | 0.250 | 0.373 | -1.886 |
| Yoruba | 0.518 | 0.207 | 0.393 | -1.405 |
| Balochi | 0.468 | 0.367 | 0.428 | -0.541 |
| Balochi | 0.539 | 0.393 | 0.481 | 0.144 |
| Bedouin | 0.444 | 0.360 | 0.410 | -0.634 |
| Brahui | 0.501 | 0.267 | 0.407 | -0.448 |
| Burusho | 0.515 | 0.360 | 0.453 | -0.266 |
| Druze | 0.484 | 0.343 | 0.428 | -0.394 |
| Hazara | 0.592 | 0.417 | 0.522 | 0.578 |
| Kalash | 0.473 | 0.330 | 0.416 | -0.121 |
| Mongolian | 0.587 | 0.633 | 0.605 | 1.068 |
| Oroqen | 0.621 | 0.433 | 0.546 | 0.789 |
| Palestinian | 0.508 | 0.367 | 0.452 | -0.418 |
| Pashtun | 0.498 | 0.270 | 0.407 | -0.611 |
| Sindhi | 0.527 | 0.300 | 0.436 | -0.533 |
| Cambodians, Khmer | 0.550 | 0.380 | 0.482 | 0.116 |
| Dai | 0.653 | 0.500 | 0.592 | 1.276 |
| Daur | 0.686 | 0.573 | 0.641 | 1.277 |
| Han | 0.645 | 0.477 | 0.577 | 1.171 |
| Hezhe | 0.612 | 0.593 | 0.605 | 1.262 |
| Japanese | 0.668 | 0.453 | 0.582 | 1.036 |
| Koreans | 0.672 | 0.475 | 0.593 | 1.102 |
| Lahu | 0.589 | 0.517 | 0.560 | 0.862 |
| Miao | 0.704 | 0.500 | 0.623 | 1.368 |
| Naxi | 0.594 | 0.483 | 0.550 | 0.102 |
| She | 0.674 | 0.433 | 0.578 | 1.199 |
| Tu | 0.654 | 0.500 | 0.593 | 1.078 |
| Tujia | 0.621 | 0.583 | 0.606 | 1.359 |
| Uyghur | 0.582 | 0.450 | 0.529 | 0.455 |
| Xibe | 0.672 | 0.440 | 0.579 | 1.531 |
| Yi | 0.623 | 0.517 | 0.581 | 0.980 |
| Adygei | 0.494 | 0.333 | 0.430 | -0.096 |
| Basque | 0.504 | 0.353 | 0.444 | -0.270 |
| French | 0.468 | 0.350 | 0.421 | -0.093 |
| Italians T | 0.499 | 0.477 | 0.490 | 0.244 |
| Italians N | 0.523 | 0.353 | 0.455 | -0.307 |
| Orcadian | 0.565 | 0.447 | 0.517 | 0.598 |
| Russians | 0.491 | 0.307 | 0.417 | -0.360 |
| Sardinian | 0.473 | 0.370 | 0.432 | -0.459 |
| Maya, Yucatan | 0.526 | 0.193 | 0.393 | -0.567 |
| Pima, Mexico | 0.538 | 0.200 | 0.403 | -0.406 |
| Melanesian, Nasioi | 0.565 | 0.473 | 0.529 | 0.027 |
| Papuan New Guinean | 0.641 | 0.353 | 0.526 | -0.509 |
| Yakut | 0.579 | 0.487 | 0.542 | 0.700 |
| Amerindians | 0.521 | 0.080 | 0.345 | -0.909 |
| Karitiana | 0.498 | 0.223 | 0.388 | -0.086 |
| Surui | 0.582 | 0.143 | 0.407 | -0.496 |

**Table 6. Polygenic and factor scores by racial group. ALFRED data.**

|  |  |  |
| --- | --- | --- |
| **Continent** | **Frequency** | **Factor** |
| African | 0.449 | -1.475 |
| Middle East | 0.448 | -0.482 |
| South Asian | 0.471 | -0.259 |
| East Asian | 0.608 | 1.091 |
| European | 0.472 | -0.092 |
| Native American | 0.441 | -0.493 |
| Oceanian | 0.555 | -0.241 |

ANOVA showed that the effect of racial group on allele frequencies was not significant, F= 0.941, p=0.471.

*Factor Analysis of allele frequencies*

A factor analysis (function “fa”, package “psych”) was carried out using Ordinary Least Squares to find the minimum residual solution. The proportion of variance explained was 0.25.

Factor scores are reported in table 5 (col. 5).

**Table 7. Factor loadings (structure matrix)**

|  |  |
| --- | --- |
| **SNP** | **Factor loading** |
| rs2883059.T | 0.39 |
| rs9824301.C | 0.64 |
| rs9393692.G | 0.10 |
| rs3847223.T | 0.07 |
| rs1906252.A | 0.46 |
| rs1941955.C | 0.65 |
| rs7666007.G | 0.56 |
| rs684498.C | -0.11 |
| rs3001723.G | 0.46 |
| rs7309.G | -0.06 |
| rs11588857.A | 0.69 |
| rs11686372.A | 0.88 |

*Spatial Autocorrelation (SAC)*

1000 Genomes

Spatial (phylogenetic) correlation was calculated using the procedure illustrated in a previous paper (Piffer, 2015), which was based (then unknown to the author) on Mantel test (Mantel, 1967). Regression analysis applied to Mantel test enables estimation of polygenic selection pressures (Piffer, 2015).

Pairwise Fst distances were correlated to pairwise score distances (absolute value of the difference in polygenic scores).

**Table 8. SAC control: Betas.**

|  |  |  |
| --- | --- | --- |
| **Source** | **Fst** | **PS** |
| Davies 2016. Β= | 0.385 | 0.294 |
| Davies 2016 + Rietveld 2013. Β= | 0.329 | 0.361 |
| Factor | -0.162 | 0.861 |

ALFRED

The correlation between PS distances and pairwise geographic distances was r= 0.164 (N=1176). The correlation between factor and pairwise geographic distances was r= 0.216.

As no IQ data were available, the effect on population IQ independent of SAC could not be estimated.

*Correction for derived and ancestral allele status*

At a theoretical level, an ancestral allele is the allele that was carried by the last common ancestor between humans and other primates whereas an allele is derived when it arose in the human lineage after the split from other primates. In practice, this allele is usually ascertained via comparison with chimpanzees. One limitation of this procedure is that if a mutation arose in chimpanzees after the split from humans, then the ancestral allele is not the chimp allele. Thus, 1000 Genomes infers ancestral alleles via alignment with 6 primate species (Ensembl, 2015).

Substantial DAF differences across populations have been found, largely due to random drift and population bottlenecks but in part also by different selection pressures (Henn et al., 2015). Non-African populations tend to have higher frequencies of derived alleles, and DAF is positively correlated to distance from Africa (Henn et al., 2015).

Since some populations have higher baseline frequencies of derived alleles, if GWAS hits tend to be derived, polygenic scores will be biased in favor (i.e. higher) of populations with higher derived allele frequencies. To correct for this, polygenic scores were calculated for derived and ancestral alleles separately. Then, the two scores were averaged to give equal weights to the two allele classes (table 9). This calibration was applied only to 1000 Genomes frequencies, as the set of independent SNPs for ALFRED is smaller (hence not giving a sufficiently big number of alleles belonging to each class), and the population samples are smaller too, thus providing less reliable estimates.

Among the 16 independent hits, there were 9 and 6 derived and ancestral alleles, respectively (the ancestral status of a SNP was missing from 1000 Genomes). We can see that African populations tend to have lower frequencies of derived alleles, and higher frequencies of ancestral alleles, confirming previous findings.

The corrected polygenic score was highly correlated to the uncorrected polygenic score (r=0.965) and to the factor (r=0.999).

**Table 9. Frequencies of alleles by ancestral vs derived allele status.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Population | Derived PS | Ancestral PS | Corrected PS |
|  | Afr.Car.Barbados | 0.238 | 0.629 | 0.434 |
|  | US Blacks | 0.278 | 0.628 | 0.453 |
|  | Bengali Bangladesh | 0.417 | 0.592 | 0.505 |
|  | Chinese Dai | 0.587 | 0.615 | 0.601 |
|  | Utah Whites | 0.472 | 0.484 | 0.478 |
|  | Chinese, Bejing | 0.641 | 0.671 | 0.656 |
|  | Chinese, South | 0.615 | 0.652 | 0.634 |
|  | Colombian | 0.409 | 0.545 | 0.477 |
|  | Esan, Nigeria | 0.218 | 0.656 | 0.437 |
|  | Finland | 0.525 | 0.548 | 0.537 |
|  | British, GB | 0.495 | 0.523 | 0.509 |
|  | Gujarati Indian, Tx | 0.433 | 0.562 | 0.497 |
|  | Gambian | 0.233 | 0.663 | 0.448 |
|  | Iberian, Spain | 0.466 | 0.523 | 0.494 |
|  | Indian Telegu, UK | 0.425 | 0.573 | 0.499 |
|  | Japan | 0.601 | 0.684 | 0.642 |
|  | Vietnam | 0.612 | 0.618 | 0.615 |
|  | Luhya, Kenya | 0.227 | 0.657 | 0.442 |
|  | Mende, Sierra Leone | 0.221 | 0.642 | 0.432 |
|  | Mexican in L.A. | 0.393 | 0.568 | 0.481 |
|  | Peruvian, Lima | 0.358 | 0.556 | 0.457 |
|  | Punjabi, Pakistan | 0.406 | 0.620 | 0.513 |
|  | Puerto Rican | 0.405 | 0.538 | 0.472 |
|  | Sri Lankan, UK | 0.424 | 0.561 | 0.493 |
|  | Toscani, Italy | 0.452 | 0.517 | 0.484 |
|  | Yoruba, Nigeria | 0.211 | 0.677 | 0.444 |

*Does distance from the Ancestral Environment predict polygenic scores?*

Lawson-Handley et al. (2007) calculated geographic distance from Addis Ababa (AA), regarded as the centre of the human expansion, assuming an origin of modern humans in Eastern Africa. For the populations for which ALFRED had a corresponding entry, the correlation between geographic distance from AA x Total PS, x Factor were r= 0.206 and r= 0.288, respectively (N=49).

**Discussion**

The analysis of independent signals from two different GWAS revealed a significant overlap across two genomic datasets. Using ALFRED and 1000 Genomes, the Rietveld et al. (2013) and Davies et al. (2016) polygenic scores were strongly correlated (r= 0.62 and 0.83, respectively). Both sets of GWAS hits were strong predictors of population IQ. The polygenic score (N=14) computed from the new independent hits (Davies et al., 2016) had a strong correlation to population IQ (r= 0.82). Similar correlation was observed for the polygenic score created by combining all the independent hits (free of LD) from the two publications (N=16): r=0.843 with population IQ.

Factor analysis produced a factor that even more strongly correlated to population IQ (r= 0.89) and survived control for spatial autocorrelation. Indeed, the predictive value of this factor was not affected by partialling out Fst distances. The high Beta value (B=0.82) and the null effect of Fst distances (B= -0.16) are suggestive of polygenic selection on these SNPs, independent of noise due to migrations or drift.

Comparisons of mean frequencies across racial groups via one-way ANOVA produced either non significant or marginally significant results, but the addition of new GWAS hits is needed to provide a definitive picture.

A remarkable finding is the high divergence in factor and polygenic scores between East Asians and Native Americans (table 6), which contrasts with their relatively recent (about 15Kya) split from the Asian lineage, after the first settlers crossed the Bering strait (Reich et al., 2012). The higher factor scores of East Asians possibly point to stronger selection pressure on educational attainment or cognitive ability following the ancestral split.

Indeed, geographic distance from East Africa was only a weak predictor of factor or polygenic scores (r= 0.21 and 0.29), suggesting that population history is not confounding the results. Further evidence comes from the weak correlations (r= 0.16-0.22) between all the pairwise geographic and polygenic distances (N=1176) .

Correction for different baseline frequencies of derived alleles did not significantly alter the polygenic scores. Interestingly, it produced a polygenic score identical to the factor score (r=0.999) obtained via factor analysis.

A limitation of this study is the reliance on GWAS hits for a complex phenotype such as educational attainment, which shares the majority of additive genetic variation with general intelligence, but also other personality and health-related traits (Krapohl et al., 2014 and 2015).

Another more obvious limitation is the small number of (independent) SNPs used for this analysis. More GWAS of intelligence or educational attainment are needed to shed light on worldwide patterns of polygenic selection on cognitive abilities.

.

**Appendix**

*SNPs frequencies and polygenic scores:*

<https://docs.google.com/spreadsheets/d/1IlcGnJRWcv5eaox-MQ8vp1pga36U1vNmw5AlkYVq7jQ/edit?usp=sharing>

Data files required to run the code:

osf.io/dk3f4

**References**

Davies, G., Marioni, R.E., Liewald, D.C., et al. (2016). Genome-wide association study of cognitive functions and educational attainment in UK Biobank (N =112 151). *Molecular Psychiatry*, 1-10. doi:10.1038/mp.2016.45

Ensembl, 2015: <http://www.1000genomes.org/faq/where-does-ancestral-allele-information-your-variants-come>

Henn, B.M., Botigué, L.R., Peischl, S., Dupanloup,I.,  Lipatov,M., Maples,B.K., Martin, A.R., Musharoff, S., Cann, H., Snyder,M.P., Excoffier, L., Kidd, J.M.,   Bustamante, C.D. (2015). Distance from sub-Saharan Africa predicts mutational load in diverse human genomes.PNAS ; published ahead of print December 28, 2015, doi:10.1073/pnas.1510805112

Krapohl, E., Rimfeld, K., Shakeshaft, N.G., Trzaskowski, M., McMillan, A., Pingault, J.-B., Asbury, K., Harlaar, N., Kovas, Y., Dale, P.S. & Plomin, R. (2014).The high heritability of educational achievement reflects many genetically influenced traits, not just intelligence. PNAS, 111, 15273–15278, doi: 10.1073/pnas.1408777111

Krapohl, E., Euesden, J., Zabaneh, D., Pingault, J.B.,, Rimfeld, K., von Stumm, S., Dale, P.S., Breen, G., O’Reilly, P.F., and Plomin, R. (2015). Phenome-wide analysis of genome-wide polygenic scores. *Molecular Psychiatry*, 1-6. doi:10.1038/mp.2015.126

Lawson Handley, L.J., Manica, A., Goudet, J., Balloux, F. (2007). Going the distance: human population genetics in a clinal world. *TRENDS in Genetics*, 23, 432-439.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research* 27, 209–220.

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 Quarterly,* 54, 168-200.

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

Reich, D., Patterson, N., Campbell, D., et al. (2012). Reconstructing Native American population history. *Nature*,16, 370-4. doi: 10.1038/nature11258.

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>