ibc Interdisciplinary Bio Central

IBC 2014;6:1, 1-6 • DOI: 10.4051/ibc.2014.6.1.0001

Open Access

RAPID REPORT

Simple Statistical Tools to Detect Signals of Recent Polygenic Selection

Davide Piffer*

Ulster Institute for Social Research, London, UK

Subject areas; General

Author contribution; D.P. wrote this article

***Correspondence** and requests for materials should be addressed to D.P. (pifferdavide@gmail.com).

Editor; Keun Woo Lee, Gyeongsang National University, Korea

Received February 07, 2014;
Accepted February 24, 2014;
Published February 25, 2014

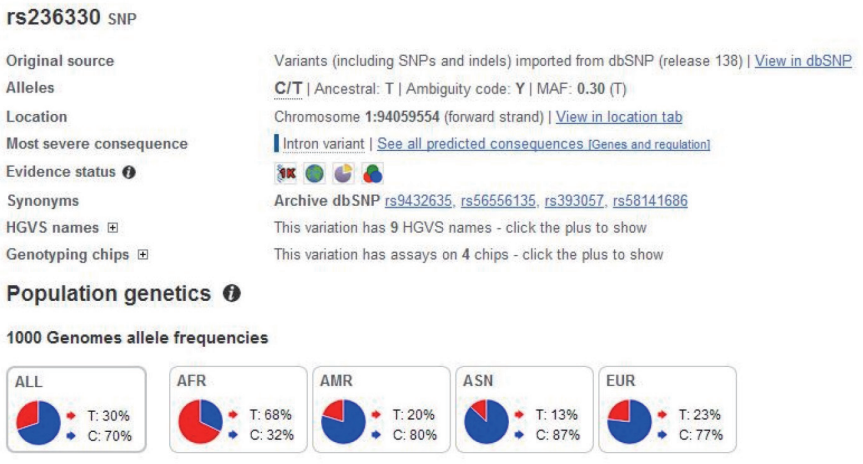
Citation; Piffer, D. Simple Statistical Tools to Detect Signals of Recent Polygenic Selection. IBC 2014, 6:01, 1-6. doi: 10.4051/ibc.2014.6.1.0001

Funding; N/A

Competing interest; All authors declare no financial or personal conflict that could inappropriately bias their experiments or writing.

SYNOPSIS

A growing body of evidence shows that most psychological traits are polygenic, that is they involve the action of many genes with small effects. However, the study of selection has disproportionately been on one or a few genes and their associated sweep signals (rapid and large changes in frequency). If our goal is to study the evolution of psychological variables, such as intelligence, we need a model that explains the evolution of phenotypes governed by many common genetic variants. This study illustrates simple statistical tools to detect signals of recent polygenic selection: a) ANOVA can be used to reveal significant deviation from random distribution of allele frequencies across racial groups. b) Principal component analysis can be used as a tool for finding a factor that represents the strength of recent selection on a phenotype and the underlying genetic variation. c) Method of correlated vectors: the correlation between genetic frequencies and the average phenotypes of different populations is computed; then, the resulting correlation coefficients are correlated with the corresponding alleles' genome-wide significance. This provides a measure of how selection acted on genes with higher signal to noise ratio. Another related test is that alleles with large frequency differences between populations should have a higher genome-wide significance value than alleles with small frequency differences. This paper fruitfully employs these tools and shows that common genetic variants exhibit subtle frequency shifts and that these shifts predict phenotypic differences across populations.



Key Words: polygenic adaptation; evolution; intelligence; height; selective sweep

ABSTRACT

A growing body of evidence shows that most psychological traits are polygenic, that is they involve the action of many genes with small effects. However, the study of selection has disproportionately been on one or a few genes and their associated sweep signals (rapid and large changes in frequency). If our goal is to study the evolution of psychological variables, such as intelligence, we need a model that explains the evolution of phenotypes governed by many common genetic variants. This study illustrates simple statistical tools to detect signals of recent polygenic selection: a) ANOVA can be used to reveal significant deviation from random distribution of allele frequencies across racial groups. b) Principal component analysis can be used as a tool for finding a factor that represents the strength of recent selection on a phenotype and the underlying genetic variation. c) Method of correlated vectors: the correlation between genetic frequencies and the average phenotypes of different populations is computed; then, the resulting correlation coefficients are correlated with the corresponding alleles' genome-wide significance. A significant difference between the allele frequencies for the three races was found. Post-hoc test revealed that East Asians had significantly higher frequencies of IQ increasing alleles than Africans. In contrast, the distribution of height increasing alleles did not differ among races. The second prediction, that alleles with large frequency differences between populations had a higher genome-wide significance value than alleles with small frequency differences, was confirmed by the analysis of the Pygmy vs non Pygmy data set.

INTRODUCTION

Polygenic adaptation (or weak widespread selection) is a model proposed to explain the evolution of highly polygenic traits that are partly determined by common, ancient genetic variation[1,2]. In contrast to the commonly held view within the academic community, which views adaptation as involving selective sweeps that drive beneficial alleles from low to high frequency in a population[3] polygenic adaptation involves modest changes in allele frequencies at many loci[1].

This model produces testable predictions[1]: a) populations with greater phenotypic values for a given trait (under the assumption of equal environments) should have higher average frequencies of trait increasing alleles. b) A positive correlation between frequency shifts (population differentiation) and effect size or genome-wide significance of allelic associations with a trait. This can be tested with a statistical tool borrowed from psychometric research, that is the method of correlated vectors[4].

Genetic variation mainly occurs between large continental groups (i.e. races), but some variation exists also within races, such as between ethnic groups. Turchin et al. (2012)[5] found significant differences in height increasing allelic frequencies between northern and southern Europeans. However, a drawback of that study is that it did not explicitly associate frequency differences with height differences between populations.

Piffer[6] showed that human populations have different frequencies of alleles increasing educational attainment and IQ and that these frequencies are positively correlated with national IQs and national scores on standardized educational attainment tests. A correlational matrix showed that the three most significant SNPs were highly correlated among each other and that, after principal component analysis, the 10 SNPs loaded highly on a single component that explained 45% of the variance. Two other SNPs, which appear to be related to IQ across several studies, were highly correlated with this factor.

However, the significance of the frequency differences between populations was not assessed. One of the aims of this paper is to overcome that shortcoming, and to test the significance of the differences between races in IQ-increasing allele frequencies.

Principal component analysis (PCA) is a new method for detecting signals of polygenic adaptation[6], when the aim is to identify the underlying component that accounts for the nonrandom distribution of allele frequencies, indicating deviation from random drift, which would produce inconsistent associations among alleles, thus assesses the strength of selection for a polygenic trait. This factor should be clearly interpretable, that is the majority of trait increasing alleles ought to load positively on it. Moreover, it ought to be significantly correlated with population trait values (e.g. average height, average IQ). A benefit of this method over ANOVA is that it exploits the higher resolution provided by the analysis of a much greater number of populations (e.g. 14 populations in 1,000 Genomes or >50 in the Allele Frequency Database). Thus, even when only a few genes are known to exert an effect on a trait, it is possible to identify a factor that likely accounts for their nonrandom distribution across populations.

The aim of this study is to extend the analysis to another highly polygenic trait (height) with phenotypic variation among populations, in order to test the following hypotheses:

a) Taller populations have higher average frequencies of height increasing alleles (and populations with higher IQs have higher frequencies of IQ increasing alleles).

b) There is a positive correlation between frequency shifts (population differentiation) and effect size or significance in GWAS of allelic associations with height.

RESULTS

IQ and Educational attainment

When considering only the 10 educational attainment alleles,

the frequencies did not significantly differ between the three races, F (2, 27) = 2.589, $P = 0.094$.

In the second ANOVA with 10 educational attainment alleles and the addition of 2 IQ increasing alleles, the frequencies differed significantly across the three races, F (2, 33) = 4.127, $P = 0.025$. Tukey post-hoc comparisons of the three groups indicated that the East Asian group (M = 47.41, 95% CI [25.63, 69.20] had significantly higher allele frequencies than the African group (M = 18.50, 95% CI [7.70, 29.29]), $P = 0.024$. Comparisons between the European group (M = 35.5, 95% CI [25.68, 53.98] and the other two groups were not statistically significant at $P < 0.05$.

Height

Frequencies of 89 SNPs from the Giant Consortium[7] with a P value = $< 1 \times 10$^-8 (corresponding to $P = 0.01$ after correction for multiple comparisons, see Johnson et al., 2010) were obtained from 1,000 Genomes. A one-way ANOVA was performed to test for different frequencies of height increasing alleles across the three human races (African, East Asian, European). Allele frequencies did not significantly differ between the three races, F (2, 264) = 0.598, $P = 0.551$.

Thus, no clear structure was visible at the level of continental groups. In order to get a more fine-grained analysis, data from all the 14 populations were used.

A polygenic score for the 14 populations was calculated as the average of all the 89 alleles. The correlation between the polygenic score and height was strongly positive and significant (r = 0.82; N = 7; $P = 0.02$).

In order to assess the underlying structure, 9 polygenic subscores were obtained, by dividing the sample in 9 variables, each variable being the average of 10 SNPs (the last one being the average of 9 SNPs, as there were only 89 SNPs), ordered according to their P value. A principal component analysis was carried out, treating the 9 polygenic scores as 9 different variables. Two components were extracted that explained respectively 64.59% and 23.62% of the variance. The two components were uncorrelated (r = -0.210), hence could not constitute an overarching factor. The first component was not clearly interpretable. The method of correlated vectors supported this, as there was no correlation between polygenic scores' PC1 loading and P value ($P = 0.033$; $P = 0.932$).The Kaiser-Meyer-Olkin (KMO) coefficient was satisfactory (0.622). Chi-squared (Bartlett's test of sphericity) = 137.62, df = 36, $P < 0.001$.

Table 1 reports the structure matrix with the loadings for the second PC. 6 of the 9 variables loaded positively and respectably high on the first factor.

Table 2 reports the Principal Component (PC2 scores) and average height for different populations. These were highly correlated with height (r = 0.98; N = 7; $P = 0.02$). Thus, this factor likely represents the strength of recent selection on a phenotype and the underlying genetic variation.

A prediction of the polygenic selection hypothesis is that SNPs with lower P values on average exhibit a stronger signal of selection. In order to test this prediction, the polygenic scores were ranked according to their P value, from lowest to highest. For each of the 9 polygenic scores, correlation with height was computed. The correlation coefficients were correlated with the polygenic score's P values.

Lower P values predicted higher polygenic scores' association with average height. Spearman rank correlation was $P = 0.683$ (N = 9; $P = 0.042$).

A moderate and positive but non-significant Spearman's rank correlation between polygenic scores' PC2 loading and P value was found ($P = 0.533$; N = 9; $P = 0.139$). That is, lower P values were associated with higher polygenic scores' loadings on PC2.

These results indicate that the SNPs with lower P values were better correlated, compared to less significant SNPs, to populations' average heights. Moreover, they predicted each polygenic

Table 1. Loadings on the second Principal Component

Polygenic Scores (ranked based on significance level)	PC2 Loading
1	0.875
2	0.836
3	0.017
4	-0.322
5	-0.715
6	0.814
7	0.008
8	0.348
9	-0.411

Table 2. Factor scores and average height for 1,000 Genomes populations

Population	PC2	Height (cm)
ASW	0.64	178
LWK	1.41	
YRI	0.65	
CLM	-0.08	
MXL	0.27	
PUR	-0.23	
CHB	-1.73	170.2
CHS	-1.68	
JPT	-1.61	170.7
CEU	0.89	179
FIN	0.45	179
GBR	0.72	178
IBS	0.29	
TSI	0.01	177

AFR; African, AMR; American, ASN; Asian, EUR; European, ASW; African ancestry in SW USA, LWK; Luhya, Kenya, YRI; Yoruba, Nigeria, CLM; Colombian, MXL; Mexican ancestry from LA, California, PUR; Puerto Ricans from Puerto Rica, CHB; Han Chinese in Bejing, China, CHS; Southern Han Chinese, JPT; Japanese in Tokyo, Japan, CEU; Utah Residents with Northern and Western European Ancestry, FIN; Finnish in Finland, GBR; British in England and Scotland, IBS; Iberian population in Spain, TSI; Toscani in Italy.

Table 3a. Frequency (%) of alleles associated with higher educational attainment (1,000 Genomes)

	IQ	rs9320913 (A)	rs3783006 (C)	rs8049439 (T)	rs13188378 (G)	rs11584700 (G)	rs4851266 (T)	rs2054125 (T)	rs3227 (C)	rs4073894 (A)	rs12640626 (A)	Average
AFR		19	35	58	0	8	4	0	12	11	17	16.4
AMR		40	43	58	3	11	27	4	58	12	62	31.8
ASN		39	29	72	1	31	56	0	84	6	73	39.1
EUR		50	42	65	6	23	37	6	48	21	57	35.5
ASW	86	23	32	50	0	7	9	0	16	7	25	16.9
LWK	74	17	40	60	0	9	1	0	10	17	17	17.1
YRI	71	18	32	61	0	6	5	0	11	6	11	15
CLM	83.5	42	53	55	3	9	23	3	62	22	57	32.9
MXL	88	30	34	56	1	9	30	5	68	5	73	31.1
PUR	83.5	51	43	64	7	15	25	6	42	11	53	31.7
CHB	105.5	42	23	76	1	30	57	0	87	5	75	39.6
CHS	106	40	33	64	1	25	59	0	87	5	75	38.9
JPT	105	35	30	78	0	38	51	0	78	7	70	38.7
CEU	100	49	46	61	8	21	40	6	45	18	56	35
FIN	97	52	33	61	6	27	35	10	59	25	57	36.5
GBR	100	49	44	67	6	26	41	8	40	18	61	36
IBS	97	43	54	71	0	21	29	4	71	21	43	35.7
TSI	100	52	43	69	5	18	34	3	45	23	59	35.1

AFR; African, AMR; American, ASN; Asian, EUR; European, ASW; African ancestry in SW USA, LWK; Luhya, Kenya, YRI; Yoruba, Nigeria, CLM; Colombian, MXL; Mexican ancestry from LA, California, PUR; Puerto Ricans from Puerto Rica, CHB; Han Chinese in Bejing, China, CHS; Southern Han Chinese, JPT; Japanese in Tokyo, Japan, CEU; Utah Residents with Northern and Western European Ancestry, FIN; Finnish in Finland, GBR; British in England and Scotland, IBS; Iberian population in Spain, TSI; Toscani in Italy.

Table 3b. Frequency (%) of alleles associated with higher IQ (1,000 Genomes)

Population	rs236330 C	Rs324650 T
AFR	32	26
AMR	80	61
ASN	87	91
EUR	77	46
ASW	40	32
LWK	31	25
YRI	28	24
CLM	84	58
MXL	85	65
PUR	68	57
CHB	92	92
CHS	89	92
JPT	79	88
CEU	81	45
FIN	75	52
GBR	71	46
IBS	61	21
TSI	82	46

score's loading on the PC2, albeit not significantly.

An analysis was also carried out on African populations, comparing Pygmy and non-Pygmy groups. The average frequencies did not differ between the two groups (39.2 and 40.3, respectively). However, the 10 height increasing alleles with the lowest *P* value were at higher frequencies among non-Pygmy (M = 0.44, SD = 0.41) than Pygmy populations (M = 0.40, SD = 0.39). A t-test showed that the frequencies did not significantly differ: t(18) = -0.226, *P* = 0.824, Cohen's d = 0.10.

Since this suggested the presence of a selection signal on SNPs with lower *P* values, a further analysis was carried out, to determine whether the difference in allele frequencies between the two groups was greater for the most significant SNPs.

A significant correlation was found between non Pygmy-Pygmy difference for each SNP and *P* value (Spearman's *P* = 0.41; N = 24; *P* = 0.047). Thus, alleles with lower *P* values exhibited higher population differentiation in the expected direction, with lower frequencies among Pygmies compared to non-Pygmy Africans.

DISCUSSION

This study found partial support for the first prediction, that trait increasing alleles are present at higher frequencies among populations with higher trait values. This was confirmed only with regards to IQ plus educational attainment increasing alleles, where a significant difference between the allele frequencies for the three races was found. Post-hoc test revealed that East Asians had significantly higher frequencies of IQ increasing alleles than Africans. In contrast, the distribution of height increasing alleles did not differ significantly across the three human groups, despite the much greater power provided by the larger number of SNPs (89 for height vs 12 for IQ).This suggests that recent directional selection on intelligence was much stronger than for stature. A similar finding was reported by Miller and Penke (2007), comparing evidence for directional selection on intelligence and an anatomical indicator other

than height: brain size.

The second prediction, that alleles with large frequency differences between populations had a higher genome-wide significance value than alleles with small frequency differences, was confirmed by the analysis of the Pygmy vs non Pygmy data set. Moreover, the significance also predicted the allele's association with the average population phenotype in the 1,000 Genomes data set, across seven populations, so that populations' average height was better predicted by alleles with lower *P* values.

Thus, the method of correlated vectors should be used as a confirmatory tool to aid in interpreting components that emerge from principal component analysis of gene frequencies.

The analysis of the entire data set, for all the 14 populations, produced mixed results. A principal component analysis extracted two factors that were uncorrelated. However, the factor explaining the largest variance was not clearly interpretable, as it had only a modest and non-significant correlation with height ($r = 0.276$). The second factor, explaining about 24% of the variance, was strongly related to height ($r = 0.984$; $P < 0.01$). The method of correlated vectors corroborated this interpretation, as lower *P* values were associated with higher polygenic scores' loadings on PC2.

The first component probably represents pleiotropic effects. Lettre et al. (2008)[8] suggest that loci associated with height may be pleiotropic, influencing the risk or severity of other diseases. The absence of a significant difference between pygmy and non-pygmy African groups suggests that the atypical height of the Pygmies is due only in part to selection on common variants and that a few mutations that evolved independently among various Pygmy groups, possibly influencing growth, thyroid function and sexual selection, account for their atypical growth pattern[9]. Drastically reduced gene expression has been proposed as another mechanism to account for the short stature of the Pygmies, in particular the expression of Growth Hormone (GH) and Growth Hormone receptor (GHR) genes, is 1.8-fold and 8-fold reduced, respectively[10].

CONCLUSION AND PROSPECTS

The results of this study confirm that common genetic variants exhibit subtle frequency shifts and that these shifts predict phenotypic differences across populations.The statistical tools proposed in this article proved overall successful for testing this.

MATERIALS AND METHODS

Height increasing alleles were obtained from a meta-analysis by the GIANT consortium[7]. Educational attainment alleles were taken from Rietveld's meta-analysis[11] and IQ alleles from Piffer[6].

Allele frequencies were taken from 1,000 Genomes (www.1000genomes.org) and the Allele Frequency Database (ALFRED; Alfred.med.yale.edu).

A one-way ANOVA with Rietveld's 10 educational attainment alleles was performed, to test for different frequencies of intelligence alleles across three human races (African, East Asian, European). The American group was excluded as it comprises populations with a substantial degree of European genetic admixture. Frequencies were taken from 1,000 Genomes and are reported in Table 3a and 3b.

Another one-way ANOVA was performed adding the 2 IQ SNPs (rs236330 and rs324650) reported in Piffer's study[6].

Frequencies of 89 SNPs from the Giant Consortium[7] with a *P* value = $< 1 \times 10^{-8}$ (corresponding to $P = 0.01$ after correction for multiple comparisons, see Johnson et al., 2010[12]) were obtained from 1,000 Genomes. A one-way ANOVA was performed to test for different frequencies of height increasing alleles across the three human races (African, East Asian, European).

For each height increasing allele, the frequencies for two Pygmy populations (Biaka, Mbuti) and 4 non-pygmy African populations were obtained from ALFRED. A total of 24 SNPs were found on ALFRED and are reported in supplementary Table 3. The average frequencies for the 2 groups (Pygmy vs non Pygmy) for all the alleles were computed.

REFERENCES

1. Jonathan, K. P., and Anna Di, R. (2010). Adaptation – not by sweeps alone. *Nature Reviews Genetics* 11, 665-667.
2. Pritchard, J. K., Pickrell, J. K., and Coop, G. (2010). The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current biology : CB* 20, R208-215.
3. Smith, J. M., and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical research* 23, 23-35.
4. Jensen, A. R., and Weng, L. J. (1994). What is a good g? *Intelligence* 18, 231-258.
5. Turchin, M. C., Chiang, C. W., Palmer, C. D., Sankararaman, S., Reich, D., Genetic Investigation of, A. T. C., and Hirschhorn, J. N. (2012). Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nat Genet* 44, 1015-1019.
6. 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.
7. Lango Allen, H., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N., Rivadeneira, F., Willer, C. J., Jackson, A. U., Vedantam, S., Raychaudhuri, S., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467, 832-838.
8. Lettre, G., Jackson, A. U., Gieger, C., Schumacher, F. R., Berndt, S. I., Sanna, S., Eyheramendy, S., Voight, B. F., Butler, J. L., Guiducci, C., et al. (2008). Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat Genet* 40, 584-591.
9. Migliano, A. B., Romero, I. G., Metspalu, M., Leavesley, M., Pagani, L.,

Antao, T., Huang, D. W., Sherman, B. T., Siddle, K., Scholes, C., et al. (2013). Evolution of the pygmy phenotype: evidence of positive selection fro genome-wide scans in African, Asian, and Melanesian pygmies. *Human biology* 85, 251-284.

10. Bozzola, M., Travaglino, P., Marziliano, N., Meazza, C., Pagani, S., Grasso, M., Tauber, M., Diegoli, M., Pilotto, A., Disabella, E., et al. (2009). The shortness of Pygmies is associated with severe under-expression of the growth hormone receptor. *Molecular Genetics and Metabolism* 98, 310-313.
11. Rietveld, C. A., Medland, S. E., Derringer, J., Yang, J., Esko, T., Martin, N. W., Westra, H. J., Shakhbazov, K., Abdellaoui, A., Agrawal, A., et al. (2013). GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* 340, 1467-1471.
12. Johnson, R., Nelson, G., Troyer, J., Lautenberger, J., Kessing, B., Winkler, C., and O'Brien, S. (2010). Accounting for multiple comparisons in a genome-wide association study (GWAS). *BMC Genomics* 11, 724.