Research Article for Psychological Science

Replicability and Robustness of GWAS for Behavioral Traits

Cornelius A. Rietveld,[1,2] Dalton Conley,[3] Nicholas Eriksson,[4] Tõnu Esko,[5] Sarah E. Medland,[6] Anna A.E. Vinkhuyzen,[7] Jian Yang,[7] Jason D. Boardman,[9,10] Christopher F. Chabris,[11] Christopher T. Dawes,[12] Benjamin W. Domingue,[9] David A. Hinds,[4] Magnus Johannesson,[13] Amy K. Kiefer,[4] David Laibson,[14] Patrik K. E. Magnusson,[15] Joanna L. Mountain,[4] Sven Oskarsson,[16] Olga Rostapshova,[14] Alexander Teumer,[17] Joyce Y. Tung,[4] Peter M. Visscher,*[,7,8] Daniel J. Benjamin,*[,18] David Cesarini,*[,19] Philipp D. Koellinger,*[,1,2,20] and the Social Science Genetic Association Consortium[21]

[1] Department of Applied Economics, Erasmus School of Economics, Erasmus University Rotterdam, 3000 DR Rotterdam, The Netherlands

[2] Department of Epidemiology, Erasmus Medical Center, Rotterdam 3000 CA, The Netherlands

[3] Department of Sociology, New York University, New York, New York 10012, United States of America

[4] 23andMe, Inc., Mountain View, California 94043, United States of America

[5] Estonian Genome Center, University of Tartu, Tartu 51010, Estonia

[6] QIMR Berghofer Medical Research Institute, 300 Herston Road, Brisbane, Queensland 4006, Australia

[7] Queensland Brain Institute, The University of Queensland, Brisbane, Queensland 4072, Australia

[8] University of Queensland Diamantina Institute, The University of Queensland, Princess Alexandra Hospital, Brisbane, Queensland 4102, Australia

[9] Institute of Behavioral Science, University of Colorado, Boulder, Colorado 80309, United States of America

[10] Department of Sociology, University of Colorado, Denver, Colorado 80217, United States of America

[11] Department of Psychology, Union College, Schenectady, New York 12308, United States of America

[12] Wilf Family Department of Politics, New York University.

[13] Department of Economics, Stockholm School of Economics, 113 83 Stockholm, Sweden

[14] Department of Economics, Harvard University, Cambridge, Massachusetts 02138, United States of America

[15] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, 171 77 Stockholm, Sweden

[16] Department of Government, Uppsala University, 75120 Uppsala, Sweden

[17] Interfaculty Institute for Genetics and Functional Genomics, Department of Functional Genomics, University Medicine Greifswald, Greifswald 17487, Germany

[18] Department of Economics, Cornell University, Ithaca, New York 14853, United States of America

[19] Center for Experimental Social Science, Department of Economics, New York University, New York, New York 10012, United States of America

[20] Faculty of Economics and Business, University of Amsterdam, 1018TV Amsterdam, The Netherlands

[21] See Supporting Online Material

* These authors contributed equally

To whom correspondence should be addressed: db468@cornell.edu (D.J.B.); dac12@nyu.edu (D.C.); p.d.koellinger@uva.nl (P.D.K.); peter.visscher@uq.edu.au (P.M.V.)

Keywords: behavior genetics, educational attainment, individual differences, genome-wide association study, population stratification

Abstract (max 150 words)

A recent genome-wide association study (GWAS) of educational attainment identified three single-nucleotide polymorphisms (SNPs) that, despite their small effect sizes (each $R^2 \approx 0.02\%$), reached genome-wide significance ($p < 5\times10^{-8}$) in a large discovery sample and replicated in an independent sample ($p < 0.05$). The study also reported associations between educational attainment and indices of SNPs called "polygenic scores." We evaluate the robustness of these findings. Study 1 finds that all three SNPs replicate in another large (N = 34,428) independent sample. We also find that the scores remain predictive ($R^2 \approx 2\%$) with stringent controls for stratification (Study 2) and in new within-family analyses (Study 3). Our results show that large and therefore well-powered GWASs can identify replicable genetic associations with behavioral traits. The small effect sizes of individual SNPs are likely to be a major contributing explanation for the striking contrast between our results and the disappointing replication record of most candidate gene studies.

Introduction

The discovery of genetic variants associated with behavioral traits could eventually be transformative for the social sciences, but the first step is identifying the specific genes associated with a trait. In psychology, the standard approach is the "candidate gene study." In a candidate gene study, a small set of genetic variants ("polymorphisms") is selected based on their hypothesized or known biological function, and these polymorphisms are tested for association with the trait. Most candidate gene studies are based on samples of several hundred participants and apply a significance threshold of 0.05 (for a review, see Ebstein, Israel, Chew, Zhong, & Knafo, 2010).

Despite the fact that such studies continue to be published in prominent journals, the successful replication of published genetic associations with behavioral traits is the exception, not the rule (Benjamin et al., 2012; Hewitt, 2012). In fact, the situation is so alarming that the editor of the leading field journal *Behavior Genetics* recently issued an editorial policy on candidate gene studies of behavioral traits that began "The literature on candidate gene associations is full of reports that have not stood up to rigorous replication" and went on to say "…it now seems likely that many of the published findings of the last decade are wrong or misleading and have not

contributed to real advances in knowledge" (Hewitt, 2012). *Psychological Science* has adopted the same strict standards for evaluating candidate-gene studies.

Why the findings from candidate gene studies of complex behaviors replicate inconsistently remains an open question, but it is commonly believed that low statistical power is a major contributing factor, and that the problem of low power is further compounded if the reported *p*-values correct for only a subset of the multiple hypotheses that were tested (Hewitt, 2012; Ioannidis, 2005). Candidate gene studies also cannot always adequately control for the well-known problem of "population stratification": genotypes may covary with unobserved environmental factors (Hamer & Sirota, 2000). For example, individuals with shared genetic ancestry (say, from the same ethnic group or from the same ancestral region) may also be more likely to share values, cultural practices, or exposure to other unobserved environmental confounds. Population stratification can give rise to associations driven by the shared environmental factors but spuriously attributed to the shared genotype (Cardon and Palmer, 2003). A finding may be confounded by population stratification even though it successfully replicates if the population structure that caused a spurious genetic discovery is also present in the replication samples.

As a result of the methodological limitations of candidate gene studies and the dramatic decline in the cost of genotyping, a paradigm shift took place around 2005 in medical research away from candidate gene studies to what are called "genome-wide association studies" (GWAS) (McCarthy et al., 2008; Pearson & Manolio, 2008; Visscher, Brown, McCarthy, & Yang, 2012). These are hypothesis-free studies in which researchers test the phenotype of interest for association with all of the (typically millions of) measured single-nucleotide polymorphisms (SNPs). Because of the large number of hypotheses tested, an association is only considered established if the SNP (i) reaches the "genome-wide significance" threshold of $p < 5\times10^{-8}$, and (ii) is subsequently successfully replicated in an independent sample at a nominal significance level of 0.05 (McCarthy et al., 2008).

Advocates of GWA studies argue that they overcome or mitigate many of the limitations of candidate gene studies. First, the large number of SNPs that are tested for association makes transparent the need to correct for multiple-hypothesis testing, which is achieved by imposing the genome-wide significance threshold of $p < 5\times10^{-8}$ (McCarthy et al., 2008). Moreover, GWA

studies, as a practical matter, tend to be based on larger samples (as indeed they must be to have any hope of identifying a SNP that reaches genome-wide significance).

Second, Bayes' Rule implies that conditional on observing an association at the genome-wide significance level, the association is likely to be true even if the study had only modest statistical power to detect the association in the first place; see Benjamin et al. (2012) for calculations.

Third, GWA data can be used to mitigate the potential confound of population stratification. In particular, it has become a common practice in GWASs to (a) estimate the first four principal components (PCs) of all the genotypes measured by the gene chip (the number four having emerged as a convention), (b) drop individuals who are genetic outliers as measured by these PCs, and then (c) include the PCs as control variables in the genetic association analysis. Intuitively, the PCs capture axes of correlation across the genome resulting from common ancestry. The PCs often have a geographic interpretation (Abdellaoui et al., 2013; Price et al., 2006, 2009). Controlling for PCs has become standard in GWA studies since Price et al. (2006) showed through simulation and empirical examples that doing so can eliminate spurious associations that are due to population structure. In Section *S-5* of the Supplemental Material, we illustrate the effectiveness of PCs using a simple placebo test. Specifically, we show that controlling for PCs eliminates a spurious association between educational attainment and a SNP for lactose intolerance that is known to vary in frequency across individuals with different ancestries (Bersaglieri et al., 2004; Campbell et al., 2005). (In contrast, the most common way of addressing population stratification in candidate-gene studies, namely including controls for self-identified race, does not eliminate the spurious association.)

There are thus many reasons to expect findings from GWA studies to replicate more consistently than findings from candidate gene studies. And experiences from the literature on complex anthropometric and medical traits suggest that GWA findings do in fact have a vastly superior replication record (Visscher et al., 2012). But do positive GWA findings from studies of complex *behavioral* traits similarly identify credible genetic associations that replicate consistently? And if the findings do replicate consistently, do they replicate consistently because what is being observed is a real genetic signal, or could it be that population stratification generates a spurious association in both the discovery sample and the replication sample? If GWA studies do identify credible and

replicable genetic associations, then they are a promising response to the non-replicability problem in gene-discovery research in the social sciences.

Until recently, virtually all GWA studies with positive findings have been studies of anthropometric or medical traits. For this reason, it may be inappropriate to infer from the superior replication record of GWA studies of medical traits that positive findings from GWA studies of behavioral traits are going to replicate consistently. If true genetic associations with behavioral traits have smaller effect sizes than true associations with anthropometric and medical traits, then GWA studies on behavioral traits will tend to generate less reliable results because they have lower power to detect true associations. Furthermore, while the convention of controlling for four PCs may be sufficient to minimize population-stratification concerns for anthropometric and medical traits, it might not be sufficient for behavioral traits, which may be characterized by more subtle population stratification.

While earlier GWA studies of behavioral traits (Benyamin et al., 2014; de Moor et al., 2012) have largely come up empty-handed (probably due to lack of power), a recent GWAS on educational attainment with a combined sample of over 100,000 individuals (Rietveld et al., 2013) identified three SNPs that meet the standard criteria for establishing a GWAS association, (i) and (ii) listed above. The effect sizes of the associations identified by Rietveld et al. are indeed small: the largest effect size corresponds to an R^2 of only approximately 0.02% (equivalent to about one month of schooling per allele). This is far smaller than the effect sizes for medical and anthropometric traits, for example, it is less than one tenth the R^2 of the largest associations discovered for height (R^2 = 0.4%; Lango Allen et al., 2010) and BMI (R^2 = 0.3%; Speliotes et al., 2010). The Rietveld et al. results therefore can serve as a test case for the robustness of the GWA approach to behavioral traits.

We sought to investigate (a) whether the Rietveld et al. results replicate in an independent sample with far more stringent controls for population stratification than are typically applied in GWA studies of medical and anthropometric traits, and (b) whether there is any evidence overall that the meta-analytic results are contaminated by unaccounted-for population stratification.

Study 1 – Replication of Specific SNPs in 23andMe with Extensive Controls for Stratification

Method

Study 1 sought to replicate the three genome-wide significant SNPs identified by Rietveld et al. in a new independent sample. The Rietveld et al. study tested approximately J = two million SNPs for association with educational attainment by running the following regression separately for each SNP $j \in \{1, 2, ..., J\}$:

$$y_i = \mu_j + \beta_j x_{ij} + \boldsymbol{\gamma}_j \boldsymbol{Z}_i + \epsilon_{ij}, \quad (1)$$

where y_i is the dependent variable (the phenotype); μ_j is a constant term; x_{ij} is the number of reference alleles (0, 1, or 2) individual i is endowed with at SNP j; β_j is the coefficient of interest; and $\boldsymbol{Z}_i$ is a vector of controls, which include age, sex, and the first four PCs of the variance-covariance matrix of the genotypic data. Rietveld et al. studied two dependent variables: *EduYears*, a measure of the number of years of schooling completed by the individual, and *College*, a binary variable equal to one if the individual had completed a college degree or its equivalent. (The point biserial correlation between the two measures is roughly 0.8; see Supplemental Materials *S-1*.) The tests of association with *EduYears* were conducted by running the linear regressions described above, and the tests of association with *College* were conducted analogously using logistic regressions.

We sought to replicate the original associations using data provided by 23andMe, a cohort based on a sample of volunteer participants (Eriksson et al., 2010) that was not included in the Rietveld et al. study. After applying quality-control filters and restricting to individuals of European descent who responded to a survey question about educational attainment, the sample size is N = 34,428. Because of the small effects, replication samples of this magnitude are required for adequate power. Given the sample size of 34,428, our power to replicate an association with R^2 = 0.02% at $p < 0.05$ is 75% (see Supplemental Materials *S-6*).

We used the same regression models (1) as in Rietveld et al., except that in our analysis, the vector of controls $\boldsymbol{Z}_i$ includes (in addition to age and sex) the first 25 PCs from the sample genotype covariance matrix—compared to only 4 PCs in Rietveld et al.—in order to reduce potential population-stratification confounding by partialing out more of the population structure.

Results

As shown in Table 1, all three SNP associations reported in Rietveld et al. replicate at a nominal significance level of 0.05, in the same direction and with similar effect sizes as in the original report. The replication of effect sizes suggests that the additional controls for population stratification from including more than 4 PCs made little difference (for related evidence, see also Supplemental Material *S-5*). As a caveat, we note that since all research participants are completely anonymous to us, we cannot rule out overlap between the 23andMe sample and the Rietveld et al. discovery or replication sample, in which case the new results would not be fully independent from the Rietveld et al. results. We believe, however, that such potential overlap is likely to be miniscule and is therefore unlikely to drive our replication findings.

Table 1. Replication in 23andMe Data

College	23andMe			Rietveld et al. Discovery			Rietveld et al. Replication		
	OR	*S.E.*	*p*-value	*OR*	*S.E.*	*p*-value	*OR*	*S.E.*	*p*-value
rs12206087	1.035	0.018	0.046	1.049	0.011	1.06×10^{-5}	1.042	0.021	0.022
rs11584700	**0.954**	**0.020**	**0.025**	**0.924**	**0.012**	$\mathbf{2.07\times10^{-9}}$	**0.923**	**0.022**	$\mathbf{4.86\times10^{-4}}$
rs4851266	**1.071**	**0.019**	**0.0001**	**1.069**	**0.012**	$\mathbf{2.20\times10^{-9}}$	**1.058**	**0.022**	**0.003**
N		34,428			95,407-95,419			23,663-23,668	

EduYears	23andMe			Rietveld et al. Discovery			Rietveld et al. Replication		
	Beta	*S.E.*	*p*-value	*Beta*	*S.E.*	*p*-value	*Beta*	*S.E.*	*p*-value
rs12206087	**0.058**	**0.020**	**0.004**	**0.106**	**0.018**	$\mathbf{4.19\times10^{-9}}$	**0.077**	**0.034**	**0.012**
rs11584700	-0.053	0.025	0.032	-0.086	0.021	2.98×10^{-5}	-0.126	0.041	9.61×10^{-4}
rs4851266	0.086	0.021	4.28×10^{-5}	0.076	0.018	3.64×10^{-5}	0.103	0.035	0.002
N		34,428			101,048-101,061			23,523-23,573	

Notes: *Beta* is the effect of an increase in one reference allele on years of schooling estimated by the linear regression model (1) for *EduYears*, and *OR* is the odds ratio (the relative likelihood of attending college for an individual with one more reference allele) estimated by the analogous logistic regression for *College*. The Rietveld et al. sample sizes are given by a range because different sample sizes are available for the three SNPs. In Rietveld et al., rs9320913 is significantly associated with *EduYears* and rs11584700 and rs4851266 with *College*. Because rs9320913 is unavailable in the 23andMe data, we use rs12206087 as a (very reliable) proxy for rs9320913 (R^2 = 0.99; see Supplemental Materials *S-2.3*); the "rs12206087" results for Rietveld et al. are actually the results for rs9320913. The **bold** rows in the *College* panel indicate the SNPs found by Rietveld et al. to be associated with *College* in their discovery sample; analogously for the *EduYears* panel.

Study 2 – Robustness of Polygenic Scores in STR and QIMR with Two Distinct Methods of Controlling for Stratification

Method

While Study 1 used a new dataset to replicate the three genome-wide significant SNPs reported by Rietveld et al., Study 2 used some of the same data as in the original report to probe the robustness of Rietveld et al.'s reported "polygenic score" results to potential confounding from population stratification. Following Purcell et al. (2009), polygenic scores are commonly constructed in the GWA literature in order to allow investigators to evaluate the joint predictive power of a large number of SNPs (possibly including SNPs whose effects are too small or estimated too imprecisely to reach genome-wide significance).

Following a common approach in the genetics literature (Purcell et al., 2009; Yang et al., 2012), Rietveld et al. constructed a polygenic score ($\hat{g}_i$) for each individual *i* as equal to a weighted sum of the number of reference alleles (0, 1, or 2) across a set of SNPs, where the weights are derived from the regression coefficients from a GWAS of either *EduYears* or *College*. They then evaluated the predictive power of an individual's score $\hat{g}_i$ for the individual's educational attainment using two hold-out samples (i.e., samples excluded from the GWAS used for estimating the weights): the Swedish Twin Registry (STR) sample and the Queensland Institute of Medical Research (QIMR) sample. Although the original datasets are family-based samples, one member from each family was selected at random to be included in the analyses. In each sample (and for scores constructed using GWASs of each of *EduYears* and *College*), Rietveld et al. tested four scores constructed from increasingly large sets of SNPs, the sets of SNPs whose GWAS associations with educational attainment fell below the respective *p*-value thresholds: 5×10^{-8} (i.e., only the genome-wide significant SNPs), 5×10^{-5}, 5×10^{-3}, and 1 (i.e., all SNPs). For each polygenic score $\hat{g}_i$, Rietveld et al. examined its predictive power by running the regression:

$$EduYears_i = \mu + \beta\hat{g}_i + \boldsymbol{\gamma}\boldsymbol{Z}_i + \varepsilon_i, \quad (2)$$

where the dependent variable is always *EduYears* (never *College*), μ is a constant term; β is the coefficient of interest; and $\boldsymbol{Z}_i$ is a vector of controls, which include age, sex, and age×sex, but no

PCs (though PCs were included as controls in the GWA analyses that generated the weights for constructing the $\hat{g}_i$'s). Rietveld et al. found that the incremental predictive power of the score (i.e., the increase in R^2 from estimating regression (2) with the score as an independent variable relative to the R^2 without the score) was larger when more SNPs were included in the score. The score containing all SNPs, which had the largest incremental predictive power, accounted for approximately 2% of the variance across individuals in educational attainment.

To explore the robustness of original prediction findings, we re-ran these prediction analyses using two distinct methods that control more stringently for population stratification. In the first, we estimated the same regression model (2), except that we additionally included in the vector of controls the first 20 PCs as control variables. In the second, we estimated mixed linear models (Kang et al., 2010) in place of the regression models. Conceptually, these models involve two steps: (i) the genome-wide data are used to estimate the degree of genetic similarity between each pair of individuals in the sample, and (ii) unlike in standard regression where the covariance of the error term (in an educational-attainment regression) between any two individuals is assumed to be zero, the covariance is fitted as an increasing linear function of the individuals' genetic similarity. In other words, to the extent that two individuals are more recently descended from a common ancestor (as very accurately measured by overall genetic similarity)—and thus are more likely be similar on unobserved environmental factors—these individuals are treated as correlated observations on the relationship between educational attainment and the score.

Results

The results are shown in Table 2. The upper panel shows the results from the association analyses with the scores constructed using different *p*-value thresholds. We separately report results from the STR and QIMR samples and separately for scores constructed from weights estimated using *College* and *EduYears*. The middle and lower panels show results, respectively, from regressions with 20 PCs included as controls and from mixed linear models. Each coefficient is the estimated effect of a one-standard-deviation increase in the score.

When all SNPs are used to construct the score, it has the predicted sign in all analyses and accounts for approximately 2% of the variance in educational attainment. In STR (the larger and therefore better-powered cohort), the polygenic score is statistically significant in all scenarios, even when

only genome-wide significant SNPs are included. The joint effect of the SNPs with $p < 5\times10^{-8}$ is approximately 0.1%-0.2% of variance in *EduYears* in STR. Since this polygenic score includes 3 SNPs (when constructed using *College*) or 5 SNPs (when constructed using *EduYears*), the results are roughly consistent with Rietveld et al.'s estimate that each of the most strongly associated SNPs explains approximately 0.02% of variance in *EduYears*. Overall, there is no systematic tendency for the predictive power of the scores to change when additional controls for stratification are included.

Table 2. Results from Additional Analyses of Polygenic Scores in QIMR and STR Data

			Queensland Institute of Medical Research (QIMR)				Swedish Twin Registry (STR)			
Threshold for inclusion of SNPs in PGS			$< 5\times10^{-8}$	$< 5\times10^{-5}$	$< 5\times10^{-3}$	All SNPs	$< 5\times10^{-8}$	$< 5\times10^{-5}$	$< 5\times10^{-3}$	All SNPs
No PC Adjustment	*College*	*Beta*	0.0495	0.1502	0.3581	0.5630	0.1706	0.1984	0.3469	0.5506
		S.E.	0.0554	0.0554	0.0551	0.0545	0.0496	0.0496	0.0496	0.0492
		p-value	0.3702	0.0067	9.1×10^{-11}	1.4×10^{-24}	6.1×10^{-4}	6.9×10^{-5}	3.1×10^{-12}	1.2×10^{-28}
		ΔR^2	0.0002	0.0021	0.0118	0.0291	0.0017	0.0023	0.0072	0.0180
	EduYears	*Beta*	-0.0221	0.2478	0.3326	0.5544	0.1353	0.2481	0.3194	0.5617
		S.E.	0.0554	0.0551	0.0551	0.0548	0.0496	0.0496	0.0496	0.0492
		p-value	0.6894	7.6×10^{-6}	1.7×10^{-9}	7.1×10^{-24}	6.5×10^{-3}	6.4×10^{-7}	1.3×10^{-10}	1.0×10^{-29}
		ΔR^2	0.0000	0.0056	0.0102	0.0282	0.0011	0.0037	0.0061	0.0188
Controlling for 20 PCs	*College*	*Beta*	0.0502	0.1617	0.3567	0.5508	0.1640	0.1968	0.3407	0.5445
		S.E.	0.0554	0.0554	0.0551	0.0548	0.0496	0.0496	0.0496	0.0492
		p-value	0.3671	3.5×10^{-3}	1.1×10^{-10}	1.5×10^{-23}	9.9×10^{-3}	7.8×10^{-5}	7.3×10^{-12}	4.9×10^{-28}
		ΔR^2	0.0002	0.0024	0.0117	0.0278	0.0016	0.0023	0.0069	0.0176
	EduYears	*Beta*	-0.0043	0.2604	0.3122	0.5372	0.1410	0.2423	0.3059	0.5588
		S.E.	0.0554	0.0554	0.0551	0.0548	0.0496	0.0496	0.0496	0.0492
		p-value	0.9400	2.6×10^{-6}	1.7×10^{-8}	1.8×10^{-22}	4.6×10^{-3}	1.1×10^{-6}	7.8×10^{-10}	2.0×10^{-29}
		ΔR^2	0.0000	0.0062	0.0090	0.0265	0.0012	0.0035	0.0056	0.0186
Mixed Linear Model Analysis	*College*	*Beta*	0.0538	0.1505	0.3383	0.4920	0.1697	0.1902	0.3387	0.5601
		S.E.	0.0571	0.0581	0.0574	0.0581	0.0504	0.0500	0.0500	0.0508
		p-value	0.3479	0.0096	3.8×10^{-9}	2.5×10^{-17}	7.5×10^{-4}	1.4×10^{-4}	1.1×10^{-11}	2.4×10^{-28}
		ΔR^2	0.0003	0.0021	0.0105	0.0222	0.0017	0.0022	0.0068	0.0187
	EduYears	*Beta*	-0.0083	0.2303	0.2666	0.4947	0.1410	0.2403	0.3104	0.5662
		S.E.	0.0581	0.0584	0.0561	0.0587	0.0513	0.0504	0.0500	0.0504
		p-value	0.8863	7.9×10^{-5}	1.9×10^{-6}	3.7×10^{-17}	6.0×10^{-3}	1.9×10^{-6}	5.1×10^{-10}	2.4×10^{-29}
		ΔR^2	0.0000	0.0049	0.0065	0.0225	0.0012	0.0034	0.0057	0.0191
		N	3,544				6,770			

Notes: *Beta* is the effect of a one-standard-deviation increase in the polygenic score on years of schooling estimated by the linear regression model (2) (the first two panel rows) or by the mixed linear model analysis (the last panel row); if *Beta* is positive, the score predicts *EduYears* in the same direction in the replication sample as in the discovery sample. *S.E.* is the estimated standard error of *Beta*, and the *p*-value is for a two-sided test. ΔR^2 is the increase in R^2 (in units of percentage points) from estimating a model that includes the polygenic score as an independent variable relative to estimating a model that excludes it.

Study 3 – Replication and Within-Family Robustness of Polygenic Scores in FHS

Method

The gold standard for ruling out concerns about population stratification is to show that the association holds within families. The original Rietveld et al. study reported within-family analysis using the pooled STR and QIMR sample. For this within-family analysis, the linear polygenic score constructed using all SNPs in the GWAS of *EduYears* is strongly associated with educational attainment, and the score constructed using a *p*-value threshold of 5×10^{-3} is marginally significant. Power was too low to draw conclusions about the scores constructed using *p*-value thresholds of 5×10^{-5} and 5×10^{-8} (which contain information from fewer SNPs). The STR and QIMR analyses were based on sample sizes of 2,774 DZ twin pairs and 572 full sibling pairs, respectively.

In Study 3, we use data from an independent sample, the Framingham Heart Study (FHS), to attempt to replicate the within-family analyses of the linear polygenic scores. FHS is an epidemiological study on three generations of individuals in the Massachusetts town of Framingham that was not included in any of Rietveld et al.'s analyses (see Supplemental Material *S-4*). In this sample, there are 395 families with two or more full biological siblings. Fewer SNPs are available in FHS than in STR and QIMR (see Supplemental Material *S-4.1*). Consequently, the polygenic scores in Study 3 are expected to have lower explanatory power than the analogous scores from Study 2. Our focus here is on examining, within the FHS dataset, how the estimated effect of the score is affected by restricting the analysis to within-family variation.

Our analyses proceeded in three steps. First, we applied quality controls to the data, pruned the SNPs for linkage disequilibrium, and then constructed the polygenic score using the meta-analytic results from Rietveld et al. Second, we identified all biological full siblings. Finally, we tested the score ($\hat{g}_i$) within-family by running regressions of the following form:

$$EduYears_i = \beta \hat{g}_i + \sum_{k=1}^{K} \gamma_k X_{ik} + \varepsilon_i, \quad (3)$$

where *i* indexes individuals, *k* indexes families, and X_{ik} is an indicator variable that takes the value 1 if individual *i* belongs to family *k* and 0 otherwise. Including the "family fixed-effect" X_{ik} is equivalent (except for the resulting R^2) to running a regression after both $EduYears_i$ and the score $\hat{g}_i$ are demeaned at the family level; hence the analysis uses only the within-family variation in *EduYears* and the within-family variation in the score. To account for the non-independence of the error term among siblings, we cluster the standard errors (Liang & Zeger, 1986) at the level of the family. Since we expected to have less power for this analysis than Rietveld et al. due to the smaller number of individuals and the smaller number of SNPs, we only ran these analyses for two scores, one constructed from all SNPs in the sample and one using a *p*-value threshold of 5×10^{-3}. (We did not conduct within-family tests of the individual SNPs because our statistical power would be less than 7%—too low to draw a meaningful conclusion regardless of how the analysis turns out; see Supplemental Material *S-6*.)

Results

Each coefficient in Table 3 is the estimated effect of a one-standard-deviation increase in the score. Columns 1 and 2 report the results from the new within-family analyses using the FHS data: whether or not we control for 20 PCs, both polygenic scores constructed from all SNPs and from SNPs reaching $p < 5\times10^{-3}$ are positively and significantly associated with educational attainment. In columns 3 and 4 we also report analyses analogous to those from Study 2 (i.e., excluding the family fixed effects, thus leveraging both between- and within-family variation in the score). In these analyses, both scores are positively associated with educational attainment, again with similar results with and without the PC controls. The score from SNPs reaching $p < 5\times10^{-3}$ is marginally significant, and the score from all SNPs is highly statistically significant.

Table 3. Results from Analyses of Polygenic Scores for *EduYears* in FHS

		Within-Family Variation Only		Between- and Within-Family Variation	
		(1)	(2)	(3)	(4)
Threshold for inclusion of SNPs in PGS		$< 5\times10^{-3}$	All SNPs	$< 5\times10^{-3}$	All SNPs
No PC Adjustment	*Beta*	0.2386	0.2677	0.1142	0.2448
	S.E.	0.0934	0.1034	0.0637	0.0614
	p-value	0.011	0.010	0.073	7.44×10^{-5}
	ΔR^2	0.0036	0.0037	0.0031	0.0141
Controlling for 20 PCs	*Beta*	0.2308	0.2642	0.1173	0.2534
	S.E.	0.0947	0.1044	0.0634	0.0621
	p-value	0.015	0.012	0.065	5.13×10^{-5}
	ΔR^2	0.0033	0.0036	0.0031	0.0140
N, #individuals		1,256		1,256	
N, #families with ≥ 2 children		395		395	

Notes: *Beta* is the effect of a one-standard-deviation increase in the polygenic score on years of schooling estimated by the linear regression model (3) (columns 1 and 2) or by linear regression model (2) (columns 3 and 4); if *Beta* is positive, the score predicts *EduYears* in the same direction in the replication sample as in the discovery sample. *S.E.* is the estimated standard error of *Beta*, and the *p*-value is for a two-sided test. ΔR^2 is the increase in R^2 (in units of percentage points) from estimating a model that includes the polygenic score as an independent variable relative to estimating a model that excludes it.

Summary

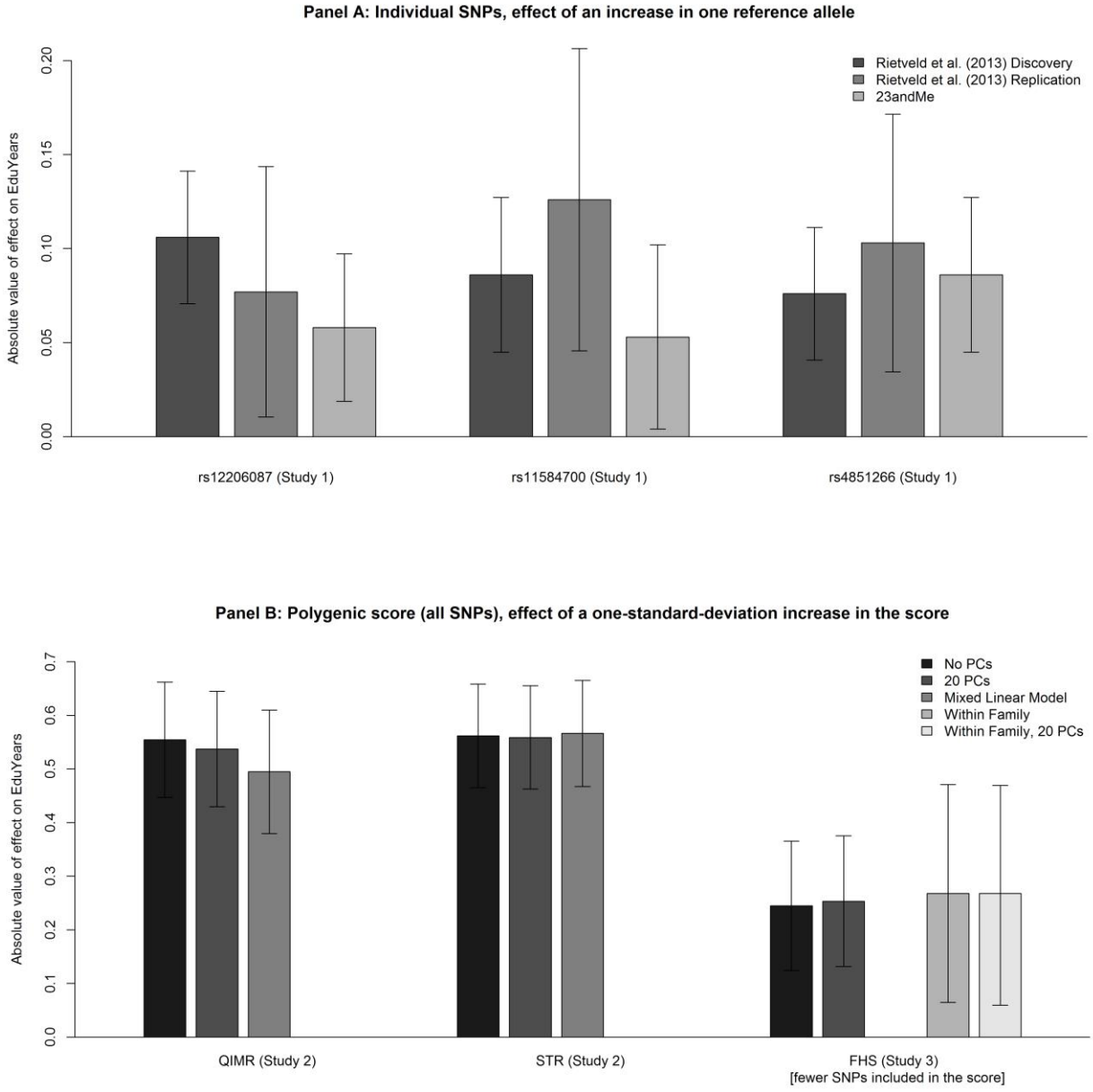
To summarize, in Study 1 we replicate in an independent sample the associations between educational attainment and Rietveld et al.'s (2013) three genome-wide significant SNPs, using more stringent controls for population stratification than is typical in the GWA literature; the next two studies show that polygenic scores robustly replicate in regressions with controls for population stratification and in within-family analyses.

To facilitate comparing the effect sizes across Studies 1-3 and Rietveld et al.'s analyses, Figure 1 shows 95% confidence intervals for the effect on *EduYears*. An effect size of 0.1, for example, is approximately 1 month of schooling. For the individual SNPs, the effects are per reference allele, and for the polygenic score containing all SNPs, the effects are per one-standard-deviation increase in the score. Panel A shows that the effect sizes of the genome-wide significant SNPs are comparable across datasets. The effect sizes of the polygenic scores in Panel B are also similar in QIMR and STR across the two different datasets and methods to control for population

stratification. The effect sizes of the polygenic score in FHS (Study 3) are not comparable to those from QIMR and STR; the effect sizes are attenuated in FHS because the scores are constructed from the smaller number of SNPs available in this sample (see Supplemental Material *S-4*). Within FHS, the effect sizes remain similar across different methods to control for population stratification, including the within-family analyses.

While these results are encouraging, we also note a potential limitation of this study. Our evidence, especially the finding that the score is significantly associated with educational attainment in within-family analyses, suggests that it is extremely unlikely that the findings of Rietveld et al. are largely an artifact of stratification. However, biases due to very subtle population stratification may still account for some of the observed relationships between educational attainment and some of the individual SNPs. This possibility cannot be *conclusively* ruled out until large enough family samples (e.g., N = 47,000 sibling pairs, see Supplemental Material *S-6*) are available to enable adequately powered within-family tests of association with individual SNPs. This potential limitation applies to all GWA studies. Our findings suggest, however, that the individual SNP associations with educational attainment are robust even when we include substantially more stringent controls than is standard in medical genetics.

Figure 1. Comparison of effect sizes across studies



Notes: Panel A shows the (absolute value of the) effect on years of schooling of a change in one reference allele for each of the three individual SNPs, with 95% confidence intervals. The results are a visual representation of the numbers in Table 1. Panel B shows the (absolute value of the) effect on years of schooling of a change in one standard deviation of the polygenic score that includes all SNPs, with 95% confidence intervals. The results for QIMR and STR are a visual representation of the numbers in the "all SNPs" columns of Tables 2. Similarly for the results for FHS and the "all SNPs" column of Table 3 (but note that the score in FHS is not comparable with the other two scores, since it is based on fewer SNPs).

Discussion

The contrast between the robustness of our findings and the disappointing replication record of most candidate gene studies of behavioral traits is striking. To draw the appropriate methodological conclusions, it is necessary to understand the causes of this difference.

A first major contributing factor is that the Rietveld et al. analyses were based on a sample size that was unprecedentedly large by the standards of social-science genetics. If, as now seems likely, the effects of individual genetic variants on most behavioral traits are small, much larger samples than are generally used are required to produce credible findings. This is a methodological lesson that applies to all studies whether they be GWA studies or not. However, as an empirical matter, candidate gene studies tend to be based on much smaller samples. Though it seems clear that much larger samples are needed, it is important to recognize that statistical power also depends on the reliability of the available phenotypic measure. Researchers will sometimes face a tradeoff between studying a studying a cruder variable available in a large sample (e.g., educational attainment) or more proximal variables available in a smaller sample (e.g., cognitive ability). Rietveld et al. (2013, SOM Section 7) provide a framework for quantifying this tradeoff.

A second contributing factor is that some of the discipline that comes from the hypothesis-based research of existing candidate gene studies is illusory: because a vast majority of genes are expressed in the brain (Ramsköld, Wang, Burge, & Sandberg, 2009), it is usually possible to create an ex post rationalization for an observed association between a candidate gene and a behavioral trait that sounds at least superficially biologically plausible. Thus, the main advantage of the candidate gene approach—namely the theoretical discipline that it imposes on the investigator—may be exaggerated.

We believe there are two key implications of our findings for research on genetics of behavioral traits. First, our results suggest that standard GWAS protocols from epidemiological research can indeed be successfully applied to the study of behavioral traits and may therefore offer a way to avoid the replication failures that are plaguing much research on the genetics of complex behavior. Second, even if (given the current state of biological knowledge) current candidate-gene approaches are not bearing fruit, this does not rule out an eventual "comeback" for hypothesis-based research in the genetics of behavioral traits. In fact, we envision that as the number of

credibly established associations from GWA studies rises, these discoveries will usher in a new era of "empirical candidate gene" studies in which the candidates are drawn from among the SNPs identified by GWA studies of related phenotypes. For example, the SNPs associated with educational attainment could be used as candidates to study cognitive and personality traits that may be part of the causal pathway. Such follow-up studies will of course need to be adequately powered to produce robust results, but since the GWAS results restrict the number of SNPs that are subsequently tested for association, the *p*-value threshold can be set much more liberally than the level of genome-wide significance.

What does the finding of small effect sizes—reported by Rietveld et al. and replicated here—imply about how genetic research in psychology should be conducted and what its payoffs will be for the field? An immediate implication is that current research using genotypic data in laboratory experiments is almost certainly underpowered, and therefore psychology should accelerate its move away from such methods, as they are unlikely to yield robust findings. A more subtle implication of the small effect sizes is that—as Turkheimer (2012) has persuasively argued—exuberant forecasts that the availability of genetic data will quickly transform the social sciences should be viewed skeptically. In principle, a genetic variant identified in an association study can explain a tiny part of the variation in the phenotype and yet point to an interesting biological system (and this has happened several times in medical genetics). In practice, it seems likely that SNPs with smaller effect sizes, on average, are more likely to operate on the phenotype through distal causal pathways involving a large number and many layers of mediating environmental factors. Therefore, it is conceivable that the identification of SNPs with very small effects will not lead to a useful psychological theory of the phenotype.

At present, it remains an open question to what extent the identification of individual SNPs will reveal new biological and psychological insights for highly polygenic behavioral traits. But we believe it is likely that genetic-association research will benefit psychology in the long run for at least two other reasons. First, even if genetic associations can only be discovered in samples of many tens of thousands of individuals, once the genetic variants to focus on have been identified, large-but-attainable samples of a few thousand individuals will provide sufficient statistical power to address interesting research questions, such as the nature and magnitude of gene-environment interactions.

Second, even though individual genetic variants have very small effects, polygenic scores can have large enough effects to be usable even in relatively small samples. The polygenic score explored here has modest explanatory power ($R^2 \approx 2\%$), but when the weights for constructing the score are estimated in larger samples, the explanatory power will be much greater. For example, Rietveld et al. (Table S26 in SOM) estimate that a polygenic score constructed using results from a discovery sample with N = 500,000 will have $R^2 \approx 12\%$. We anticipate that such sample sizes will be attainable in the next few years, making it possible to construct such a score. Once a polygenic score with $R^2 = 12\%$ can be calculated for each genotyped participant in a study, a sample of only 62 participants will be needed for 80% power to detect its effect.

In summary, our results suggest that in psychology, a shift away from candidate gene studies and toward GWA studies is likely to be fruitful. However, before the potential payoffs can be realized, the focus of much research on the genetics of behavioral traits will need to be reoriented, and new research infrastructures will need to be created—for example, to build much larger sample sizes than most GWA studies of behavioral traits have had access to. Nevertheless, we believe that this investment is worth making because it may lead to accumulation of reliable and replicable knowledge about the genetics of behavioral traits.

Author contributions

All authors contributed to the conception and design of the studies. Nicholas Eriksson, David Hinds, Amy Kiefer, Joanna Mountain, and Joyce Tung performed the analyses for study 1. Sarah Medland, Anna Vinkhuyzen, Jian Yang and Peter Visscher performed the analyses for study 2. Dalton Conley, Jason Boardman, Christopher Dawes, and Benjamin Domingue performed analyses for study 3. Cornelius Rietveld, Daniel Benjamin, David Cesarini, and Philipp Koellinger all wrote substantial portions of the article and the Supporting Online Material. Dalton Conley, Nicholas Eriksson, Tõnu Esko, Christopher Chabris, Magnus Johannesson, David Laibson, Patrik Magnusson, Sven Oskarsson, Olga Rostapshova, and Alexander Teumer critically reviewed and edited the manuscript.

Acknowledgments

This research was carried out under the auspices of the Social Science Genetic Association Consortium (SSGAC), a cooperative enterprise among medical researchers and social scientists

that coordinates genetic-association studies for social-science variables. This research was funded primarily by the Ragnar Söderberg Foundation (E9/11) and also by The Swedish Council for Working Life and Social Research (2006-1623) and the National Institute on Aging (NIA)/NIH through grants P01-AG005842, P01-AG005842-2052, P30-AG012810, and T32-AG000186-23. The formation of the SSGAC was made possible by an EAGER grant from the NSF and a supplemental grant from the NIH Office of Behavioral and Social Sciences Research (SES-1064089).

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Supplemental Material

Additional supporting information may be found on the journal's website.

The meta-analytic data from the GWAS on educational attainment (Rietveld et al. 2013) are publically available for follow-up research and replication purposes at http://ssgac.org/Data.php.

Further details about materials and methods used in Rietveld et al. (2013) can be found at http://ssgac.org/documents/Rietveld_et_al_2013_Science_SOM.pdf.

References

Abdellaoui, A., Hottenga, J.-J., de Knijff, P., Nivard, M. G., Xiao, X., Scheet, P., … Boomsma, D. I. (2013). Population structure, migration, and diversifying selection in the Netherlands. *European Journal of Human Genetics : EJHG*, *21*(11), 1277–85. doi:10.1038/ejhg.2013.48

Benjamin, D. J., Cesarini, D., Chabris, C. F., Glaeser, E. L., Laibson, D. I., Gudnason, V., … Lichtenstein, P. (2012). The promises and pitfalls of genoeconomics. *Annual Review of Economics*, *4*(1), 627–662. doi:10.1146/annurev-economics-080511-110939

Benyamin, B., Pourcain, B., Davis, O. S., Davies, G., Hansell, N. K., Brion, M.-J., … Haworth, C. M. A. (2014). Childhood intelligence is heritable, highly polygenic and associated with FNBP1L. *Molecular Psychiatry*, *19*, 253–258.

Bersaglieri, T., Sabeti, P. C., Patterson, N., Vanderploeg, T., Schaffner, S. F., Drake, J. A., … Hirschhorn, J. N. (2004). Genetic signatures of strong recent positive selection at the lactase gene. *American Journal of Human Genetics*, *74*(6), 1111–20. doi:10.1086/421051

Campbell, C. D., Ogburn, E. L., Lunetta, K. L., Lyon, H. N., Freedman, M. L., Groop, L. C., … Hirschhorn, J. N. (2005). Demonstrating stratification in a European American population. *Nature Genetics*, *37*(8), 868–72. doi:10.1038/ng1607

Cardon, L. R., & Palmer, L. J. (2003). Population stratification and spurious allelic association. *Lancet*, *361*(9357), 598–604. doi:10.1016/S0140-6736(03)12520-2

De Moor, M. H. M., Costa, P. T., Terracciano, A., Krueger, R. F., de Geus, E. J. C., Toshiko, T., … Boomsma, D. I. (2012). Meta-analysis of genome-wide association studies for personality. *Molecular Psychiatry*, *17*(3), 337–349. doi:10.1038/mp.2010.128

Ebstein, R. P., Israel, S., Chew, S. H., Zhong, S., & Knafo, A. (2010). Genetics of human social behavior. *Neuron*, *65*(6), 831–844.

Eriksson, N., Macpherson, J. M., Tung, J. Y., Hon, L. S., Naughton, B., Saxonov, S., … Mountain, J. (2010). Web-based, participant-driven studies yield novel genetic associations for common traits. *PLoS Genetics*, *6*(6), e1000993. doi:10.1371/journal.pgen.1000993

Hamer, D., & Sirota, L. (2000). Beware the chopsticks gene. *Molecular Psychiatry*, *5*(1), 11–13. doi:10.1038/sj.mp.4000662

Hewitt, J. K. (2012). Editorial policy on candidate gene association and candidate gene-by-environment interaction studies of complex traits. *Behavior Genetics*, *42*(1), 1–2. doi:10.1007/s10519-011-9504-z

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124.

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. a, Kong, S.-Y., Freimer, N. B., … Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, *42*(4), 348–354. doi:10.1038/ng.548

Lango Allen, H., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N., Rivadeneira, F., … O/'Connell, J. R. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, *467*(7317), 832–838. doi:10.1038/nature09410

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22. doi:10.1093/biomet/73.1.13

Mccarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., & Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, *9*(5), 356–369. doi:10.1038/nrg2344

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., & Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, *9*(5), 356–369.

Pearson, T. A., & Manolio, T. A. (2008). How to interpret a genome-wide association study. *JAMA : The Journal of the American Medical Association*, *299*(11), 1335–44. doi:10.1001/jama.299.11.1335

Price, A. L., Helgason, A., Palsson, S., Stefansson, H., St Clair, D., Andreassen, O. A., … Stefansson, K. (2009). The impact of divergence time on the nature of population structure: an example from Iceland. *PLoS Genetics*, *5*(6), e1000505. doi:10.1371/journal.pgen.1000505

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909. doi:10.1038/ng1847

Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F., … Moran, J. L. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, *460*(7256), 748–752. doi:10.1038/nature08185

Ramsköld, D., Wang, E. T., Burge, C. B., & Sandberg, R. (2009). An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Computational Biology*, *5*(12), e1000598. doi:10.1371/journal.pcbi.1000598

Rietveld, C. A., Medland, S. E., Derringer, J., Yang, J., Esko, T., Martin, N. W., … Koellinger, P. D. (2013). GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science*, *340*(6139), 1467–1471. doi:10.1126/science.1235488

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., … I McCarthy, M. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics*, *42*(11), 937–948. doi:10.1038/ng.686

Turkheimer, E. (2012). Genome wide association studies of behavior are social science. In K. S. Plaisance & T. A. C. Reydon (Eds.), *Philosophy of Behavioral Biology* (pp. 43–64). New York: Springer.

Visscher, P. M. M., Brown, M. A. A., McCarthy, M. I. I., & Yang, J. (2012). Five years of GWAS discovery. *The American Journal of Human Genetics*, *90*(1), 7–24. doi:10.1016/j.ajhg.2011.11.029

Yang, J., Ferreira, T., Morris, A. P., Medland, S. E., Madden, P. A. F., Heath, A. C., … Visscher, P. M. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics*, *44*(4), 369–375. Retrieved from http://www.nature.com/ng/journal/v44/n4/abs/ng.2213.html#supplementary-information