ONLINE APPENDIX

for

Genetics and Economics

Jonathan P. Beauchamp, [1] [2] David Cesarini, [3] Magnus Johannesson [4]
Matthijs J.H.M. van der Loos, [5] Philipp D. Koellinger,
Patrick J.F. Groenen, James H. Fowler, [6] J. Niels Rosenquist,
A. Roy Thurik, Nicholas A Christakis

11/3/2011

[1] Corresponding author. Email: jpbeauch@fas.harvard.edu
[2] Harvard University
[3] New York University
[4] Stockholm School of Economics
[5] Erasmus University Rotterdam
[6] University of California, San Diego

In this Online Appendix, we provide more details on the data, methods, and results of the Genome-Wide Association Study (GWAS) of educational attainment described in the main paper. In a GWAS, tens or hundreds of thousands of genetic markers are individually tested for association with a trait of interest. In the first stage of our GWAS, we analyzed data on about 7,500 individuals from the Framingham Heart Study who have been genotyped at over half a million SNPs and we searched for SNPs that correlate with educational attainment. In the replication stage of our GWAS, we attempted to replicate the 20 most significant associations from the first stage in the independent Rotterdam Study with data on more than 9,500 genotyped individuals.

1 Data

1.1 Framingham Heart Study

Six decades ago, the U.S. Public Health Service selected the town of Framingham, Massachusetts, as the site for a major study on cardiovascular diseases. This study became known as the Framingham Heart Study. Participants can be divided into three groups of roughly equal size: the Original Cohort, the Offspring Cohort and the Third Generation Cohort. The study was initiated when the Original Cohort was formed in 1948. A total of 5,209 individuals, representing two thirds of all adults domiciled in Framingham at the time, were enrolled. In 1971, 5,124 biological descendants of members of the Original Cohort, as well as their spouses, were also enrolled. These 5,124 individuals are referred to as the Offspring Cohort. Finally, in 2002 the study was expanded to include biological descendants of the Offspring cohort. These 4,095 individuals are the Third Generation Cohort. A total of 14,531 individuals were thus enrolled in the Framingham Heart Study (not including certain other smaller, ancillary cohorts).

Study participants regularly come to a central facility for medical examinations and the collection of demographic and background data. During several of these examinations, data on educational attainment was obtained. Recently, biological specimens to be used for genotyping were also collected from a large number of subjects. Below, we describe the genotyping of the Framingham Heart Study participants as well as the construction of the educational attainment variable.

1.1.1 Genotyping

Out of the 14,428 members of the three main cohorts, a total of 9,237 individuals have been genotyped (4,986 women and 4,251 men). The fraction of members who provided DNA samples differed somewhat across the three cohorts, with 29% percent of Original Cohort members, 73% percent of Offspring Cohort members, and 95% percent of Third Generation members being genotyped. This is a high response rate considering that the provision of genetic information was entirely voluntary and given that most of the Original Cohort members and many members of the Offspring Cohort were deceased when the collection of genetic data began. Genotyping was conducted using the Affymetrix 500k chip - an array which contains 500,568 single nucleotide polymorphisms (SNPs), which are specific genetic markers that exhibit variation between individuals (Affymetrix, 2009).

1.1.2 Educational Attainment

The measures of educational attainment varied by cohort. Original Cohort members were asked to indicate their highest educational attainment on a scale with nine categories, ranging from "fourth grade or less" to "graduate education". We converted responses in each of the nine categories to years of educational attainment. All members of this cohort were aged 28 or above when they responded to the question. Thus, it can be assumed that respondents had completed their lifetime education when the question was posed.

Most members of the Offspring Cohort responded to the question "How many years of school did you complete? " in the third examination. We used responses to this question as the primary measure of educational attainment, excluding a small number of individuals who had not attained an age of 25 when the examination took place. Some individuals who failed to respond to the question in the third examination had answered a similar question ("Education years completed") in the second examination. When responses to this question were available, they were used to replace missing values for those individuals who were at least 25 years of age when the examination was administered.

Finally, for the third generation cohort, data on educational attainment is based on responses to the question "What is the highest degree or level of school you have completed? ". This question was administered in the first and only examination of the cohort, and there were eight response categories, ranging from "no schooling" to "graduate or professional degree". Again, we only included responses from individuals who had attained an age of 25 when the exam was administered.[1]

Out of the 9,237 individuals with genotypic data, educational and basic demographic data is available for 8,496. These individuals constitute our baseline sample. Some descriptive statistics for the baseline sample, disaggregated by cohort, are given in Table II.

1.2 Replication with the Rotterdam Study

The Rotterdam Study (Hofman et al., 2009) is a prospective cohort study that currently consists of three cohorts. The first cohort, called RS-I, was successfully recruited in the well-defined Ommoord district in Rotterdam from January 1990 to September 1993 and contains 7,983 participants. The participants were all 55 years of age or older when entering the study and the oldest participant at the start was 106 years old. The second cohort, RS-II, recruited an additional 3,011 participants from February 2000 to December 2001 and consisted of individuals who became 55 years old since the initial study and of individuals aged 55 years or more who moved into the Ommoord district. The last cohorts was recruited from February 2006 until December 2008 and comprises 3,932 individuals aged 45 years or more living in the district and who had not been previously interviewed. Together, the three cohorts contain data on 14,926 individuals aged 45 years or more.

1.2.1 Genotyping

From the 14,926 individuals in the three Rotterdam cohorts, 10,211 have been satisfactorily genotyped (4,324 males and 5,887 females). In RS-I, 5,974 participants (75%) have been genotyped; the corresponding numbers for RS-II and RS-III are 2,157 (72%) and 2,080 (53%), respectively. Genotyping was done with the Illumina 550K array for RS-I and RS-II and with the

[1] Approximately five percent of respondents had not attained an age of 25 when the question was administered.

Illumina 610K array for RS-III[2]. Because the Framingham and Rotterdam studies used different types of arrays, we used imputed data for the association analysis in the Rotterdam Study. However, as we describe below, calculation of the principal components was based on the original genotyped data of the Rotterdam Study.

1.2.2 Educational Attainment

As for the Framingham cohorts, the measures of educational attainment varied slightly over the Rotterdam cohorts. None of the surveys included questions asking directly for the number of years of attained education. Therefore, measures of educational attainment were converted into years of educational attainment.

Most of the participants of RS-I responded to the question "What is your highest attained education? ", with eight answer categories ranging from Primary Education to University. For RS-II, the question "What is the highest education level you have attended" was asked to the participants; the participants were also asked whether or not they completed that education level. Based on these two questions, we converted the highest completed education level to years of educational attainment. For RS-III, the question "What is the highest level of education you have completed? ", with six answer categories, was converted into years of educational attainment.

Educational attainment, basic demographics and genotypes are available for 9,535 out of 10,211 genotyped participants. For RS-I, RS-II, and RS-III, the sample sizes containing individuals with sufficient genotypic and phenotypic data are, respectively, 5,806, 1,665, and 2,064. Some descriptive statistics for this sample, disaggregated by study, are given in Table III.

2 Method

In this section, we detail the methods used to analyze the data. We begin by describing the methods used for the first stage of the GWAS, with the Framingham data; then, we describe the methods used for the second, replication stage of the GWAS, with the Rotterdam data.

2.1 Framingham Sample - First Stage

Data from the Framingham Heart Study was used for the first stage of the GWAS. In the first stage, all available genetic markers that passed a number of quality-control filters were tested for association with educational attainment. We first outline our implementation of standard quality control measures, designed to reduce problems that may arise due to genotyping errors. We then describe how we controlled for population stratification, a problem particular to genetic association studies. Finally, we explicate how we tested for association and how standard errors and p-values were adjusted to account for (i) the non-independence of the error terms within family and (ii) multiple hypothesis testing.

2.1.1 Preliminary Steps for the GWAS

Following usual practices (Pearson and Manolio, 2008; Sullivan and Purcell, 2008), we first applied a number of quality control measures to the sample comprising all 9,237 individuals with genetic data.

First, 499 individuals were dropped because they had a "missingness" larger than 0.05. An individual's missingness is the fraction of the SNPs in the employed array with missing data for

[2] A small part of RS-II was genotyped with the 610K array instead of the 550K array.

the individual. A high missingness can be suggestive that some problem occured in the genotyping procedure for this individual, and therefore that the nonmissing genotypic data might not be accurate enough. A requirement of less than 5% missingness is customary in the molecular genetics literature (Pearson and Manolio, 2008; Sullivan and Purcell, 2008).

Next, we excluded individual SNPs which failed one of three additional quality controls. First, SNPs with a missing data frequency greater than 2.5% were deleted. A high missingness can be suggestive that some problem occurred in the genotyping procedure for that SNP. Second, we eliminated SNPs for which the least common allele had an incidence smaller than 1% (this measure is also called the "minor allele frequency"). Coefficients on these SNPs will generally be imprecisely estimated and can thus be misleading. Finally, we excluded SNPs which failed a test of Hardy-Weinberg equilibrium at the 10^{-6} level. The null hypothesis of this test is that the observed genotype frequencies are equal to their theoretical expectations under random mating. A large departure from Hardy-Weinberg equilibrium may be an indication of genotyping errors. These three quality control measures are widely used by convention in the molecular genetics literature (Pearson and Manolio et al, 2008; Sullivan and Purcell, 2008).

From the 500,568 SNPs on our Affymetrix 500k array, 76,764 did not satisfy the missingness criteria, 61,293 did not satisfy the minor allele frequency criteria, and 16,991 did not pass the Hardy-Weinberg test. Applying all three filters leaves a total of 363,776 SNPs for analysis[3].

2.1.2 Population Stratification

Population stratification refers to differences in allele frequencies across subpopulations. Such differences can occur in the absence of random mating between subpopulations as a consequence of founder effects, genetic drift, and differences in natural selection pressures. When both the frequencies of alleles and environmental factors affecting a trait of interest vary across subpopulations, spurious associations between those alleles and the trait might result.

An interesting example of population stratification was provided by Hamer and Sirota (2000), who asked their readers to entertain the thought experiment of looking for genetic markers for chopstick use. Consider conducting such a study using a sample comprising, say, Caucasian and Asian individuals. Without population stratification controls, markers which differ significantly in frequency between the Caucasian and Asian subpopulations will be found to be associated with chopstick use, but those associations will of course be due to cultural differences, not to genetic differences. Although the individuals in the Framingham Heart Study are almost all of European ancestry, population stratification has been shown to be a concern even in samples of European Americans (Campbell et al., 2005).

Several approaches have been proposed in the literature to control for population stratification. We employed the EIGENSTRAT method developed by Price et al (2006), which has emerged as a standard approach. This method applies principal component analysis to the genotypic data to obtain the loadings of each individual on the 10 principal components associated with the 10 largest eigenvalues. These loadings are then added as control variables in the main regression specification. These 10 values contain information about population structure, so including them in an association test partly controls for population stratification.

Because principal component analysis assumes independent observations, we did not use our entire (family-based) sample to construct the principal components. Instead we used a subsample of 2,507 unrelated individuals to calculate the principal components of the genotypic data and

[3] Some SNPs failed to pass more than one filter.

then used a function of the EIGENSTRAT software to project the other individuals in the sample onto those principal components, thus obtaining the loadings of each individual on each of the top 10 principal components.

Consistent with standard procedures, we dropped outliers from the sample; outliers are defined as individuals whose ancestry was at least 6 standard deviations from the mean on one of the top ten inferred axes of variation (Price et al., 2006). 531 outliers were thus eliminated, leaving 8,207 individuals with satisfactory genotypic data. The final sample used for the GWAS comprised 7,574 individuals with satisfactory genotypic and phenotypic data[4].

2.1.3 Association Analysis

For each individual SNP that passed the filters, we ran the following regressions,

$$Edu=\beta_0+\beta_1\cdot SNP_S+PC\cdot\beta_2+X\cdot\beta_3+\varepsilon, \qquad (1)$$

where *Edu* is years of education, SNP_S is the number of copies of the minor allele (0, 1, or 2) an individual has at SNP *s*, *PC* is a vector of the 10 top principal components of the genome of the sample (to control for population stratification), and the vector *X* includes a cubic of birth year and a cubic of birth year interacted with gender. Notice that this regression specification assumes that years of education are linear in the number of minor alleles. The model is misspecified if, in expectation, the educational attainment of the heterozygotes is in fact not the midpoint of the two homozygotes[5].

Two complications arise when doing inference. The first is that the matrix $\Omega\equiv E[\varepsilon\varepsilon']$ is not diagonal, as the Framingham sample is family-based and related individuals share parts of their environments and large portions of their genomes. The second difficulty is that because a very large number of hypotheses are being tested, many SNPs will inevitably turn out to be statistically significant at conventional levels just because of sampling variation. We discuss these issues briefly in turn.

2.1.4 Modeling the Error Structure

We specify a parametric structure on the matrix Ω to account for the nonindependence of the error terms across individuals. In what follows, the subscripts *i* or *j* refer to individuals, $f\in\{1,...,F\}$ indexes families, and $g\in\{1,2,3\}$ refers to the three generations in the data.

First, assume that the error terms of individuals from different families are independent. We can write $\Omega=\text{diag}(\Omega_1,\Omega_2,...,\Omega_F)$, where $\Omega_f=E[\varepsilon_f\varepsilon_f']$ is the covariance matrix of the error terms for individuals in family *f*. To model the correlation structure of Ω_f, we follow the basic ACE model from the behavioral genetics literature (Falconer and Mackay, 1996; Neale and Cardon, 1992) and assume that phenotypic (outcome) variance is the sum of three independent latent variables: additive genetic factors, common environmental factors, and individual environment. More precisely, dropping individual subscripts for expositional convenience, we assume that the error can be written as,

---

[4] The sample size for each regression in the GWAS was generally a bit smaller than that, because for each SNP there were some individuals with missing genotypic information.

[5] In genetic parlance, the model assumes that all genetic variation is additive.

$$\varepsilon=\sigma_{\varepsilon}(aA_{-SNP_S}+cC+eE), \qquad (2)$$

where $\sigma_{\varepsilon}=\sqrt{\sigma_{\varepsilon}^2}$, $\sigma_{\varepsilon}^2=var(\varepsilon)$, and A_{-SNP_S}, C, and E are, respectively, the latent additive genetic (with SNP_S partialled out), common environmental, and individual environmental factors underlying educational attainment. To identify the model, we assume without loss of generality that the variables A_{-SNP_S}, C, and E are standardized to have mean 0 and unit variance. This implies that a^2, c^2 and e^2 sum to one.

The latent variable A_{-SNP_S} captures the variation in education that is attributable to additive genetic factors, which correspond to the sum of the individual effects of all individual alleles. Though genetic variation can also be attributable to the interaction of the two alleles at a given locus (dominance) and to the interaction of alleles at different loci (epistasis), the empirical evidence suggests that much of the genetic variation is additive for most traits (Hill et al, 2008); we therefore neglect these more complex sources of genetic variation. C captures the environmental factors that vary between the homes or families and that matter for educational attainment. Examples might be parental education, socioeconomic status, the quality of local schools, shared peer influences and certain elements of parenting style. Finally, E encompasses everything that is not captured by the other variables of the equation. Geneticists interpret E as a latent index of individual environment, but to the econometrician, E is simply an error term.

Our strategy is to obtain consistent estimates of the parameters a, c and e and then use these estimates to adjust the variance-covariance matrix to account for the within-family error structure. We make the simplifying assumptions that σ_{ε}^2 does not vary across generations: $\sigma_{\varepsilon|g=1}=\sigma_{\varepsilon|g=2}=\sigma_{\varepsilon|g=3}=\sigma_{\varepsilon}$. We also note that $E[\varepsilon|g]\approx E[\varepsilon]=0 \ \forall g$, since controls for age are included in (1).

Biometrical genetic theory implies that, if mating is random,

$$E[A_{-SNP_S,i},A_{-SNP_S,j}]=r_{ij}, \qquad (3)$$

where r_{ij} is Sewall Wright's coefficient of relationship. Wright's coefficient of relationship for two individuals is the probability that the alleles of the two individuals at a random locus are identical copies of the same ancestral allele (i.e. that they are identical by descent). For instance, for full siblings, $r=\frac{1}{2}$, and likewise for a parent and his/her offspring; for a grandparent and his/her grandchild, $r=\frac{1}{4}$; and for cousins, $r=\frac{1}{8}$. We follow the behavioral genetic literature and assume that full-siblings completely share their common environment. Modeling the transmission of common environment from parent to child is more complicated and no generally agreed upon model exists (See Feldman et al., 2000, for an accessible introduction). We assume that,

$$E[C_{i\,g},C_{j\,g+1}]=\gamma, \qquad (4)$$

where *i* is the father or the mother of *j*. From these assumptions, it is possible to work out the entire correlation structure of Ω_f; the results are shown in Table I.

2.1.5 Inference under Multiple Hypothesis

A challenging issue that arises in genome-wide association studies is how to properly do statistical inference given the large number of hypotheses being considered (one for each SNP). Several methods have been proposed to address this issue. The most stringent solution is to use the Bonferroni correction, in which the conventional significance threshold is divided by the number of tests performed to obtain a Bonferroni-corrected significance threshold or, equivalently, all p-values are multiplied by the number of tests performed to obtain Bonferroni-corrected p-values. In the first stage study with the Framingham data, 363,776 tests were performed (one for each SNP that passed the quality-control filters), thus yielding a Bonferroni-corrected significance threshold of $0.05/363{,}776=1.37\cdot10^{-7}$. However the Bonferroni approach is generally agreed to be overly conservative, because SNPs that are close to one another are generally correlated and thus not statistically independent[6]. The most utilized threshold in the literature for large GWAS's based on 500,000-SNP array data was set by the Wellcome Trust Case Control Consortium at $5\cdot10^{-7}$ (Wellcome Trust Case Control Consortium, 2007).

However, as we discuss below, previous experience with false positives in the field of medical genetics has led researchers to be cautious in interpreting results that have not been replicated in an independent sample. Hence, the above significance thresholds must be seen as suggestive only - the ultimate demonstration of a true association requires replication in an independent sample.

2.1.6 Estimation Procedure

We ran at total of 363,776 regressions, one for each individual SNP. Properly accounting for the correlation structure of the error term in each of these regressions would have been very computationally demanding. Therefore, as a first step, we used the software PLINK (Purcell et al., 2007; Purcell, 2008) to estimate regression (1), neglecting the non-independence of the error terms. This procedure gives correct, consistent estimates of $\hat{\beta}_1$ and $\hat{\sigma}_\varepsilon^2$, but the standard errors of these estimates are downward biased.

Next, we kept the 98 SNPs whose Bonferroni-corrected p-values for $\hat{\beta}_1$ were significant at the five percent level and obtained consistent estimates of the standard error of $\hat{\beta}_1$ for those SNPs, taking the correlation structure of the error term into account. To do so, we calculated the empirical correlation in the residuals from regression (1) for all full siblings pairs, all parent-child pairs, and for all aunt/uncle-nephew/niece pairs (there were approximately 4,950 full siblings pairs, 5,300 parent-child pairs, and 5,900 aunt/uncle-nephew/niece pairs, depending on the SNP). We then obtained consistent estimates of a^2, c^2, and γ by solving the following system of 3 equations with 3 unknowns:

[6] When two SNPs are correlated, geneticists say that they are in "linkage disequilibrium".

$$\hat{\rho}_{FS}(\varepsilon_i;\varepsilon_j|i;j \text{ are full siblings}) = \frac{1}{2}\hat{a}^2+\hat{c}^2$$

$$\hat{\rho}_{PC}(\varepsilon_i;\varepsilon_j|i;j \text{ are parent-child}) = \frac{1}{2}\hat{a}^2+\hat{\gamma}\,\hat{c}^2 \qquad (5)$$

$$\hat{\rho}_{AUC}(\varepsilon_i;\varepsilon_j|i;j \text{ are Aunt/uncle-nephew/niece}) = \frac{1}{4}\hat{a}^2+\hat{\gamma}\,\hat{c}^2$$

From this, we obtained $\hat{\Omega}_f \ \forall f=1..F$, as well as the following consistent estimator of the variance covariance matrix of the regression coefficients:

$var(\hat{\beta})=(\Sigma_{f=1}^{F} X_f^T X_f)^{-1}(\Sigma_{f=1}^{F} X_f^T \hat{\Omega}_f X_f)(\Sigma_{f=1}^{F} X_f^T X_f)^{-1}$.

As expected, the 98 p-values from the second step were all larger than those from the first step.

2.2 Rotterdam Study - Replication Stage

In the second stage, we attempted to replicate in the Rotterdam Study - an independent sample - the 20 most significant associations from the first stage of the GWAS. As we discuss below, such a replication step is now seen as necessary in the genetics community to validate the associations from the first stage.

2.2.1 Population Stratification

To control for population stratification, the top ten principal components of the genetic data were computed. The same quality-control measures as for the Framingham data were applied to all 10,211 genotyped individuals in the Rotterdam cohorts using the PLINK software (Purcell et al., 2007; Purcell, 2008). First, the individual missingness filter of 0.05 did not lead to the exclusion of any individuals in any of the cohorts. For RS-I 561,466 SNPs were available for analysis, of which 18,261 did not satisfy the missingness criteria, 24,977 did not satisfy the minor allele frequency criteria, and 4,082 did not pass the test of Hardy-Weinberg equilibrium. Of the 537,405 SNPs available for analysis in RS-II, 19,944 did not satisfy the missingness criteria, 23,986 did not satisfy the minor allele frequency criteria, and 1,002 did not pass the Hardy-Weinberg test. Finally, for RS-III 587,388 SNPs were available for analysis, of which 4,992 did not satisfy the missingness criteria, 33,625 did not satisfy the minor allele frequency criteria, and 1,366 did not pass did not pass the Hardy-Weinberg test. Applying all three filters left 517,397 SNPs for RS-I, 493,193 SNPs for RS-II, and 548,197 SNPs for RS-III for the analysis[7].

After quality control, the filtered data were used to compute the first 10 principal components for each of the cohorts independently using the EIGENSTRAT software. Outliers whose ancestry was at least 6 standard deviations from the mean on one of the top ten inferred axes of variation were removed. This procedure removed 229 individuals from RS-I, 86 individuals from RS-II, and 109 individuals from RS-III, thus leaving 5,745 individuals in RS-I, 2,071 individuals in RS-II, and 1,971 individuals in RS-III with sufficient genotypic data. Finally, keeping only individuals with complete genotypic and phenotypic data left 5,583 individuals in RS-I, 1,601 individuals in RS-II, and 1,958 individuals in RS-III.

[7] As in the Framingham sample, some SNPs failed to pass more than one filter.

2.2.2 Association Analysis

As mentioned above, the genotypic data for the Framingham Heart Study and for the Rotterdam Study come from different genotyping platforms. Consequently, many of the 20 most significant SNPs from the first stage were not directly available in the Rotterdam Study and had to be imputed. Imputation is performed by using the correlation structure of an independent, more densely genotyped sample to infer the genotypes at the SNPs that have not been genotyped in the sample of interest.

Only SNPs with a minor allele frequency greater than 0.01, a p-value greater than $1 \cdot 10^{-6}$ on the test of Hardy-Weinberg equilibrium, and a missingness less than 2% were used for the imputation. Imputation was performed with the software MACH (Li and Abecasis, 2006) using the HapMap samples (The International HapMap Consortium, 2003) as reference.

The association analysis was performed on the imputed data for the 20 SNPs using the Mach2qtl software (Li and Abecasis, 2006) through a web-based tool called GRIMP (Estrada et al., 2009).

For each SNP, the model in Equation (1) was estimated[8]. The regression analysis has been performed for each cohort independently and cumulative betas, standard errors, and p-values were obtained from a meta-analysis through the software Metal (Abecasis et al., 2007).

3 Results

3.1 First Stage Results from Framingham Data

In Table IV, we report results for the 20 SNPs which attained the highest statistical significance. The first column gives the rs number of each SNP with the chromosome on which it is located in parentheses. The second column shows the regression coefficients. The estimates are clustered around 0.25 for most SNPs, meaning that in our sample, the difference between the two homozygotes is about 0.5 years in educational attainment. It is important to emphasize that the reported estimates are likely to be subject to substantial upward bias because of a "winner's curse" type of selection bias (Zhong and Prentice, 2008). Put simply, the likelihood that a SNP passes the significance threshold obviously depends on the change in mean phenotype associated with having an additional minor allele in the particular sample studied. The SNPs that emerge as the most significant are therefore likely to be associated with greater differences in means than one might expect if a new, independent sample were drawn. The estimated effect sizes are usually smaller in follow-up studies than in the original study, even when replication attempts are successful (Ioannidis et al., 2001). Thus, the regression coefficients for each of the top SNPs do not give an unbiased estimate of the corresponding population parameters.

In the third column we report the raw p-value of each SNP. Four of the SNPs reached the conventional significance threshold of $5 \cdot 10^{-7}$ established by the Wellcome Trust Case Control Consortium. As shown in the fourth column, none of the SNPs survive a Bonferroni correction at

[8] Note that here SNP_S is the SNP dosage (a fractional number between 0 and 2 equal to the expected number of minor allele copies of the SNP from the imputation) instead of an integer indicating the exact number of minor allele copies.

the ten percent level, the two lowest Bonferroni-corrected p-values being 0.11[9]. The top two hits – rs11758688 and rs12527415 – are in the vicinity of several known genes, the closest being the IER2 gene, which is located a little over 40,000 base pairs away from the two SNPs. In addition, rs17350845 is located in the MAPKAP2 gene, and rs9646799 is located 79,000 base pairs away from the ITGA4 gene. The other two "significant" SNPs do not appear to be located near coding regions of the genome.

In Table V, we report the SNPs (from the above set of 20 SNPs) which are near any known genes along with the distances in base pairs. Ten of the 20 SNPs are in the vicinity of at least one gene. Three SNPs - rs17350845, rs10436961 and rs4845129 - are actually located within the MAPKAP2 gene. In addition, SNP rs11225388 is located inside the MMP27 gene.

3.2 Replication Stage Results from Rotterdam Data

Table VI reports the results of the replication attempt of the top 20 SNPs from the first stage with the Rotterdam data. The first column reports the rs number of the SNP and the chromosome number in parentheses. The second column contains the estimated beta coefficients. The third column presents the nominal p-values and the fourth column reports the Bonferroni-corrected p-values that have been adjusted for 20 tests (because replication was attempted for 20 SNPs).

As evidenced by the results in the fourth column, none of the top 20 SNPs has a statistically significant association with educational attainment in the Rotterdam data. In fact, the signs of the estimated beta coefficients from the first stage and the replication stage are only identical for 9 of the 20 SNPs.

[9] As a robustness check, we also computed standard errors by clustering at the level of the family. In general, the clustered standard errors were considerably smaller than the standard errors used to compute the p-values reported in Table III, and eight of the top twenty hits survived the Bonferroni correction at the ten percent level.

4 REFERENCES

Abecasis, Goncalo, Yun Li, and Cristen Willer. 2007. “Metal.” Available at http://www.sph.umich.edu/csg/abecasis/Metal.

Affymetrix. 2009. “AffymetrixGenome-Wide Human SNP Array 5.0.” Available at http://www.affymetrix.com/support/technical/datasheets/genomewide_snp5_datasheet.pdf.

Campbell, Catarina D., Elizabeth L. Ogburn, Kathryn L. Lunetta, Helen N. Lyon, Matthew L. Freedman, et al. 2005. “Demonstrating Stratification in a European American Population.” *Nature Genetics*, 37(8): 868–872.

Estrada, Karol, Anis Abuseiris, Frank G. Grosveld, André G. Uitterlinden, Tobias A. Knoch, and Fernando Rivadeneira. 2009. “GRIMP: a web- and grid-based tool for high-speed analysis of large-scale genome-wide association using imputed data.” *Bioinformatics*, 25(20): 2750–2752.

Falconer, Douglas S., and Trudy F.C. Mackay. 1996. *Introduction to quantitative genetics.* London; New York: Longman.

Feldman, Marcus W., Sarah P. Otto, and Freddy B. Christiansen. 2000.“Genes, Culture and Inequality.” In *Meritocracy and Economic Inequality*, eds. Kenneth Arrow, Samuel Bowles and Steven Durlauf, 61–85. Princeton: Princeton University Press.

Hamer, Dean H., and Leo Sirota. 2000. “Beware the Chopstick Gene.” *Molecular Psychiatry* 5(1): 11–13.

Hill, William G., Michael E. Goddard, and Peter M. Visscher. 2008. “Data and theory point to mainly additive genetic variance for complex traits.” *PLoS Genetics* 4(2): e1000008.

Hofman, Albert, Monique M.B. Breteler, Cornelia M. van Duijn, Harry L.A. Janssen, Gabriel P. Krestin, et al. 2009. “The Rotterdam Study: 2010 Objectives and Design Update.” *European Journal of Epidemiology* 24(9): 553–572.

Ioannidis, John P.A., Evangelia E. Ntzani, Thomas A. Trikalinos, and Despina G. Contopoulos–Ioannidis. 2001. “Replication Validity of Genetic Association Studies.” *Nature Genetics* 29 (2001), 306–309.

Li, Yun, and Gonçalo R. Abecasis. 2006. “Mach 1.0: Rapid Haplotype Reconstruction and Missing Genotype Inference.” *American Journal of Human Genetics* S79: 2290.

Neale, Michael C., and Lon R. Cardon. 1992. *Methodology for genetic studies of twins and families.* Dordrecht: Kluwer Academic Publishers.

Pearson, Thomas A., and Teri A. Manolio. 2008. "How to interpret a genome-wide association study." *JAMA*, 299(11): 335-1344.

Price, Alkes L., Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, et al. 2006. “Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies.” *Nature Genetics*, 38(8): 904–909.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Manuel A.R. Ferreira, David Bender, et al. 2007. “PLINK: A Toolset for Whole-Genome Association and Population-Based Linkage Analysis.” *American Journal of Human Genetics*, 81(3): 559–575.

Purcell, Shaun. 2008. “PLINK 1.05.” Available at http://pngu.mgh.harvard.edu/purcell/plink.

Sullivan, Patrick F., and Shaun Purcell. 2008. “Analyzing Genome-Wide Association Study Data: A Tutorial Using PLINK.” In *Statistical Genetics; Gene Mapping Through Linkage and Association*, eds. Benjamin M. Neale, Manuel A.R. Ferreira, Sarah Medland and Danielle Posthuma, 355–394. New York: Taylor & Francis Group.

The International HapMap Consortium. 2003. “The International HapMap Project.” 2003.

"The International HapMap Project," *Nature*, 426(6968): 789–796.

Wellcome Trust Case Control Consortium. 2007. "Genome-Wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls." *Nature*, 447(7145): 661-678.

Zhong, Hua, and Ross L. Prentice. 2008. "Bias-Reduced Estimators and Confidence Intervals for Odds Ratios in Genome-Wide Association Studies." *Biostatistics*, 9(4): 621-634.

5 Tables

TABLE I.

EXPECTED CORRELATIONS BETWEEN THE ERROR TERMS OF RELATED INDIVIDUALS

Relatedness	$E[A_iA_j]$	$E[C_iC_j]$	$E[\varepsilon_i\varepsilon_j]$
Full siblings	$\frac{1}{2}$	1	$\sigma_\varepsilon^2(\frac{1}{2}a^2+c^2)$
Half siblings	$\frac{1}{4}$	$\frac{1}{2}$	$\sigma_\varepsilon^2(\frac{1}{4}a^2+\frac{1}{2}c^2)$
Parent-child	$\frac{1}{2}$	γ	$\sigma_\varepsilon^2(\frac{1}{2}a^2+\gamma c^2)$
Grandparent-grandchild	$\frac{1}{4}$	γ^2	$\sigma_\varepsilon^2(\frac{1}{4}a^2+\gamma^2c^2)$
Full cousins	$\frac{1}{8}$	γ^2	$\sigma_\varepsilon^2(\frac{1}{8}a^2+\gamma^2c^2)$
Half cousins	$\frac{1}{16}$	$\frac{1}{2}\gamma^2$	$\sigma_\varepsilon^2(\frac{1}{16}a^2+\frac{1}{2}\gamma^2c^2)$
Aunt/uncle-nephew/niece	$\frac{1}{4}$	γ	$\sigma_\varepsilon^2(\frac{1}{4}a^2+\gamma c^2)$
Half aunt/uncle-nephew/niece	$\frac{1}{8}$	$\frac{1}{2}\gamma$	$\sigma_\varepsilon^2(\frac{1}{8}a^2+\frac{1}{2}\gamma c^2)$

NOTES: This table gives the assumed error structure for relatives in our sample. Full siblings have the same biological parents; “half siblings” share one biological parent; “full cousins” have the same two grand-parents on either the paternal or the maternal side; “half cousins” have only one grandparent in common; “half aunt/uncle-nephew/niece” refers to pairs of individuals where the father of one is the grandfather of the other, or the mother of one is the grandmother of the other.

TABLE II.

SUMMARY STATISTICS FOR THE FRAMINGHAM HEART STUDY

Cohort	Original Cohort	Offspring Cohort	Third Generation
Birthyear	1911	1937	1962
S.D.	6.83	9.64	7.88
# Obs	1461	3388	3647
1 if Female	0.599	0.553	0.530
S.D.	0.490	0.497	0.499
# Obs	1461	3388	3647
Educational Attainment	11.61	13.95	15.10
S.D.	3.21	2.52	1.97
# Obs	1461	3388	3647
1 if Caucasian	-	-	0.996
S.D.	-	-	0.064
# Obs	-	-	3647
1 if Married	0.88	0.82	0.68
S.D.	0.32	0.38	0.47
# Obs	1461	3092	3639

NOTES: This table gives some descriptive statistics, disaggregated by cohort, for the final sample of individuals for whom genotypic data and basic demographic information is available. Birth year is approximated by the distance in time between age at first examination and the average date on which the first examination was administered for each respective cohort. Marriage is a variable taking the value 1 if the individual was married when the first examination was administered.

TABLE III.

SUMMARY STATISTICS FOR THE ROTTERDAM STUDY

Cohort	Rotterdam Study I	Rotterdam Study II	Rotterdam Study III
Birthyear	1922	1935	1951
S.D.	9.12	7.97	5.76
# Obs	5806	1665	2064
1 if Female	0.588	0.524	0.561
S.D.	0.492	0.500	0.496
# Obs	5806	1665	2064
Educational Attainment	9.02	10.81	11.16
S.D.	2.80	2.55	2.86
# Obs	5806	1665	2064
1 if Married	-	0.71	0.80
S.D.	-	0.45	0.40
# Obs	-	1665	2056

NOTES: This table gives some descriptive statistics, disaggregated by cohort, for the final sample of individuals for whom genotypic data and basic demographic information is available. Marriage is a variable taking the value 1 if the individual was married (RS-II), or was married or living with a partner (RS-III) when the first examination was administered.

TABLE IV.

TOP 20 HITS FROM FIRST STAGE OF GWAS IN FRAMINGHAM DATA

SNP (Chromosome)	$\hat{\beta}$	p-value	Bonferroni	Sample	Minor Allele	Nearby Genes?
rs11758688 (6)	-0.253	$2.97 \cdot 10^{-7}$	0.107	7572	T	Yes
rs12527415 (6)	-0.253	$3.03 \cdot 10^{-7}$	0.109	7570	T	Yes
rs17365411 (2)	0.260	$3.73 \cdot 10^{-7}$	0.134	7559	C	No
rs7655595 (4)	-0.266	$3.99 \cdot 10^{-7}$	0.144	7486	G	No
rs17350845 (1)	-0.291	$6.22 \cdot 10^{-7}$	0.224	7415	C	Yes
rs12691894 (2)	-0.246	$6.67 \cdot 10^{-7}$	0.240	7572	G	No
rs9646799 (2)	0.271	$7.41 \cdot 10^{-7}$	0.267	7478	T	Yes
rs11722767 (4)	-0.257	$7.77 \cdot 10^{-7}$	0.280	7574	C	No
rs10947091 (6)	-0.245	$9.03 \cdot 10^{-7}$	0.325	7574	T	Yes
rs6536456 (4)	0.230	$1.32 \cdot 10^{-6}$	0.474	7513	C	No
rs1580882 (4)	0.229	$1.43 \cdot 10^{-6}$	0.516	7556	T	No
rs6536463 (4)	0.229	$1.48 \cdot 10^{-6}$	0.533	7571	G	No
rs1502720 (4)	0.228	$1.66 \cdot 10^{-6}$	0.560	7566	C	No
rs10436961 (1)	-0.268	$1.82 \cdot 10^{-6}$	0.657	7540	A	Yes
rs4845129 (1)	-0.265	$2.07 \cdot 10^{-6}$	0.745	7546	G	Yes
rs17365432 (2)	0.257	$2.32 \cdot 10^{-6}$	0.836	7573	G	No
rs11225388 (11)	0.261	$2.51 \cdot 10^{-6}$	0.904	7559	G	Yes
rs7743593 (6)	0.301	$2.68 \cdot 10^{-6}$	0.965	7545	C	Yes
rs10028331 (4)	-0.259	$2.93 \cdot 10^{-6}$	1.00	7565	G	No
rs11964691 (6)	0.307	$3.29 \cdot 10^{-6}$	1.00	7458	T	Yes

NOTES: This panel reports the top 20 hits from the first stage of the GWAS in the Framingham data.

TABLE V.

SUBSET OF TOP 20 HITS WITH NEARBY GENES

SNP	Nearby Genes (distance in kb)
rs11758688	NRM(-99), MDC1(-75), TUBB(-66), FLOT1(-48), IER3(-46), DDR1(98)
rs12527415	NRM(-95), MDC1(-71), TUBB(-62), FLOT1(-44), IER3(-42)
rs17350845	LGTN(-88), DYRK3(-51), DYRK3(-51), MAPKAPK2(0), IL10(67), IL19(98)
rs9646799	ITGA4(79)
rs10947091	NRM(-88), MDC1(-64), TUBB(-55), FLOT1(-37), IER3(-34)
rs10436961	LGTN(-76), DYRK3(-39), DYRK3(-39), MAPKAPK2(0), IL10(80)
rs4845129	LGTN(-85), DYRK3(-49), DYRK3(-49), MAPKAPK2(0), IL10(70)
rs11225388	MMP20(-79), MMP27(0), MMP8(8), MMP10(65), MMP1(85)
rs7743593	SLC16A10(-15), KIAA1919(21), REV3L(62)
rs11964691	SLC35B3(-28)

NOTES: This table reports the subset of SNPs from the twenty most significant hits from the first stage of the GWAS which are near known genes. Distance is listed in thousands of base pairs away from the gene of interest, with sign dictating whether the SNP is downstream (negative) or upstream (positive) from the encoding region of the gene. Using the PLINK retrieval interface, SNP annotations were created using the TAMAL database (Hemminger et al., 2006) based chiefly on UCSC genome browser files (Hinrichs et al., 2006), HapMap (Altshuler et al., 2005), and dbSNP (Wheeler et al., 2006).

TABLE VI.

REPLICATION OF TOP 20 HITS IN THE ROTTERDAM STUDY

SNP (Chromosome)	$\hat{\beta}$	p-value	Bonferroni	Sample	Minor Allele
rs11758688 (6)	-0.0674	0.1209	1	9142	T
rs12527415 (6)	-0.0689	0.1138	1	9142	T
rs17365411 (2)	-0.0314	0.4553	1	9142	C
rs7655595 (4)	0.0007	0.9877	1	9142	G
rs17350845 (1)	0.0175	0.7162	1	9142	C
rs12691894 (2)	0.0698	0.0998	1	9142	G
rs9646799 (2)	-0.0139	0.7579	1	9142	T
rs11722767 (4)	0.0007	0.9877	1	9142	C
rs10947091 (6)	-0.0674	0.1229	1	9142	T
rs6536456 (4)	0.0272	0.4917	1	9142	C
rs1580882 (4)	0.026	0.5105	1	9142	T
rs6536463 (4)	0.0209	0.5962	1	9142	G
rs1502720 (4)	0.026	0.5213	1	9142	C
rs10436961 (1)	0.0173	0.7196	1	9142	A
rs4845129 (1)	0.0175	0.7164	1	9142	G
rs17365432 (2)	-0.0326	0.4675	1	9142	G
rs11225388 (11)	-0.0648	0.1476	1	914	G
rs7743593 (6)	0.0172	0.7397	1	9142	C
rs10028331 (4)	-0.0226	0.6317	1	9142	G
rs11964691 (6)	-0.0441	0.4137	1	9142	T

NOTES: This panel reports the results of the replication attempt in the Rotterdam Study of the top 20 hits from the first stage of the GWAS in the Framingham data. The Bonferroni-corrected p-values have been adjusted for 20 tests. Imputation quality and imputation R^2 data was produced separately for each study (RS-I, RS-II, and RS-III) and is therefore not shown, but is available upon request.