A **Genome'Wide Association Study** of **Educational Attainment**

Jonathan P. Beauchamp[!], David Cesarini[-]
Matthijs J.H.M. van der Loos[+] Philipp D. Koellinger[+]
Patrick J.F. Groenen[+] James H. Fowler[.] J. Niels Rosenquist[,]
A. Roy Thurik[+] Nicholas A. Christakis[,]

July 20, 2010

ABSTRACT

Twin and adoption studies have consistently found that genetic variation is an important source of heterogeneity in economic outcomes such as **educational** attain' ment and income. The advent of inexpensive, **genome'wide** scans is now making it increasingly feasible to directly examine specific genetic variants that predict individ' ual differences. In this paper, we conduct a **genome'wide association study** (GWAS) of **educational** achievement. In the first stage, we used data on over 360,000 genetic markers throughout the genome from the Framingham Heart **Study**, a family'based sample of nearly 8,500 individuals, and found a number of markers with suggestive as' sociations with **educational attainment**. The most promising variants were significant

at the 5 10-' level. In the second stage, we attempted to replicate the most significant first'stage associations using data from the Rotterdam **study**, an independent sample of over 9,500 individuals. None of the first'stage associations replicated, suggesting that the first'stage results were false positives. We discuss the challenges that arise

Corresponding author. Email: jpbeauch@fas.harvard.edu
#Harvard University
$New York University
"Erasmus University Rotterdam
%University of California, San Diego

when doing inference in genoeconomics research, emphasizing the importance of prop' erly correcting for multiple hypothesis testing and of replicating significant results in independent samples. We also discuss issues of power and sample size. We argue that if proper attention is given to these methodological challenges, the burgeoning field of ge' noeconomics will add a valuable new dimension to our understanding of heterogeneity in economic outcomes.

1 Introduction

Economists have long sought to better understand the causes of heterogeneity in out' comes between individuals. Economic outcomes such as household income and their pre' cursors, such as **educational attainment** and health, have been modeled to be a function of environmental inputs, genes, and interactions between the two (Becker, 1975; Becker and Thomes, 1979; Björklund, Lindahl and Plug, 2006). While a number of rich datasets and creative empirical strategies have enabled researchers to estimate parameters for environ' mental factors (Angrist and Krueger, 1999), the direct **study** of genetic markers has proven much more elusive as it was only recently that large datasets with comprehensively geno' typed subjects became available. Instead, studies on the effects of genes were historically limited to twin and adoption designs (Cesarini, 2010; Sacerdote, 2007, Taubman, 1976a). Such papers consistently find that the genetic relatedness between two individuals predicts their similarity on **educational attainment**, as well as on other economic outcomes.

In economics, the first paper on this topic is due to Taubman (1976a), who reported that

genetically identical (monozygotic, also known as MZ) twins exhibited greater similarity than fraternal (dizygotic, also known as DZ) twins on both **educational attainment** and income. Under some strong structural assumptions, the excess resemblance of MZ twins over DZ twins can be translated into a number ' heritability ' which measures the share of variance that can be explained statistically by genetic differences. Taubmanos finding of moderate heritability in **educational attainment** has subsequently been replicated in other samples of MZ and DZ twins (Miller, Mulvey and Martin, 2001) and confirmed in adoption studies (Sac' erdote, 2007) as well as papers using other pairings of relatives to identify environmental and genetic sources of variance (Behrman and Taubman, 1989). The heritability of **educational attainment** is typically estimated to be in the neighborhood of 40% (Björklund, Jäntti and Solon, 2005; Sacerdote, forthcoming).

While these studies are useful for underscoring the role of genes as a source of individual variation, they are limited both in an empirical sense, in that they require strong structural

---

assumptions, and in the practical sense, in that the policy implications of such exercises are often much less evident than is sometimes supposed (Goldberger, 1979; Jencks, 1980; Sacer' dote, forthcoming). Their overall usefulness is also undermined by the fact that heritability estimates cannot shed light on the specific genetic mechanisms that explain the **association** between genes and the outcome of interest (Jencks and Brown, 1977; Jencks, 1980). For **educational attainment**, a plausible hypothesis is that genome'level influences on mediating variables such as cognitive abilities, risk'taking and other aspects of personality explain a substantial share of the heritable variation.

The advent of inexpensive, **genome'wide** scans that contain information about hundreds of thousands of gene variants, and the inclusion of such information in large population datasets, means that it is becoming increasingly feasible to incorporate molecular genetic

data into economic analyses (Benjamin et al., 2007). This new focus on specific genes is ultimately likely to lead to a more sophisticated understanding of the pathways from specific genetic variants to complex social outcomes.

In this article, we contribute to the nascent pgenoeconomicsq literature by reporting the results of a **Genome'Wide Association Study** (GWAS) of **educational attainment**. In a GWAS, tens or hundreds of thousands of genetic markers are individually tested for asso' ciation with a trait of interest. In the first stage of our GWAS, we analyzed a dataset of nearly 8,500 individuals who have been genotyped for over half a million genetic markers, in order to search for specific gene variants associated with **educational attainment**. Our data is from the Framingham Heart **Study** (FHS), a longitudinal **study** initiated in 1948, which has been used extensively for medical studies. In the replication stage of our GWAS, we attempted to replicate the 20 most significant associations from the first stage using data from the Rotterdam **Study**, an independent dataset of more than 9,500 individuals who have also been genotyped.

We have three complementary aims in this article. First, we apply **genome'wide** associa' tion (GWAS) techniques to rich population datasets, looking at a specific outcome of great

interest to economists. Second, we seek to provide economists with an accessible discussion of how genetic studies should be performed and interpreted, given the numerous statistical challenges that arise. Finally, we also discuss how genetic data (and analyses) can improve existing empirical work in economics and, ultimately, inform policy. To preview our results, we found a number of promising markers ($p < \&\%^{-\&}$) in the first stage with the Framingham data, four of which were statistically significant at the * $\&\%^{-}$ level. However, the top 20 associations all failed to replicate in the independent Rotterdam **Study** sample, suggesting

that the first stage results were false positives. We discuss the implications of this for current efforts to find molecular genetic markers for economic behaviors and outcomes.

The paper is structured as follows. In Section II, we describe the Framingham and Rotterdam datasets and provide some descriptive statistics for our samples. In Section III, we introduce some basic concepts from molecular genetics. Section IV introduces the GWAS method, with careful emphasis on three complications that arise in such studies: (i) the possibility of genotyping errors, (ii) population stratification, (iii) inference under multiple hypothesis testing. We further discuss how to compute standard errors of the estimates to account for the family structure of the data. In Section V, we report our results, and in Section VI, we discuss the challenges that arise when doing inference in genoeconomics research, emphasizing power issues as well as the importance of properly correcting for multiple hypothesis testing and of replicating significant results in independent samples. In Section VII, we discuss how findings from molecular genetic studies such as this one can be brought to bear on fundamental economic questions. Section VIII concludes.

2 Data

2.1 Framingham Heart **Study**

Six decades ago, the U.S. Public Health Service selected the town of Framingham, Massa' chusetts, as the site for a major **study** on cardiovascular diseases. This **study** became known

as the Framingham Heart **Study** (FHS). Participants in FHS 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. FHS refers to these 5,124 individuals as the Off' spring 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 FHS (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 FSH participants as well as the construction of the **educational attainment** variable.

2.1.1 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 Genera' tion 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).

2.1.2 **Educational Attainment**

The measures of **educational attainment** varied by cohort. For details, see Appendix A. Original Cohort members were asked to indicate their highest **educational attainment** on a scale with nine categories, ranging from pfourth grade or lessqto pgraduate educationq. 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 pHow many years of school did you complete?qin 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 (pEducation years completedq) 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 pWhat is the highest degree or level of school you have completed?q. This question was administered in the first and only examination of the cohort, and there were eight response categories, ranging from pno schoolingqto pgraduate or professional degreeq. 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

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

statistics for the baseline sample, disaggregated by cohort, are given in Table II.

2.2 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.

2.2.1 Genotyping

From the 14,926 individuals in the three Rotterdam **Study** 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 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**.

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

2.2.2 **Educational Attainment**

As for the Framingham cohorts, the measures of **educational attainment** varied slightly over the Rotterdam **Study** 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 pWhat is your highest attained education?q, with eight answer categories ranging from Primary Education to University. For RS'II, the question pWhat is the highest education level you have attendedqwas 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 pWhat is the highest level of education you have completed?q, 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 suffi cient 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.

3 Elementary Genetic Concepts

Genome'wide association studies exploit a particular source of genetic variation among

individuals, namely variation in the single nucleotides that constitute the basic building blocks of DNA. Nucleotides come in four different forms: A (adenine), C (cytosine), T (thymine), and G (guanine)[3]. At each locus (position in the genome), a personos genotype consists of two pairs of nucleotides. Due to a property of the genome called complementarity,

[3]More precisely, these are the nitrogenous bases associated with the nucleotides.

the nucleotide A is always paired with the nucleotide T, and the nucleotide C is always paired with the nucleotide G. Because the second nucleotide of a pair can be directly identified from knowledge of the first one, we will often refer to a locus as consisting of two single nucleotides rather than of two pairs of nucleotides.

At each locus, one pair of nucleotides comes from the father and the other comes from the mother. For example, suppose that at a given locus there are two possible alleles (genetic variants) in the population, C and T (corresponding to the pair C and G and to the pair T and A). If both parents are pheterozygousqat that locus, meaning they each have a C and a T allele (the ordering is irrelevant), then a given offspring has a 25% chance of being phomozygousqfor C (CC), a 25% chance of being homozygous for T (TT) and a 50% chance of being heterozygous (CT).

Whenever a nucleotide varies across individuals, it is said to be a single nucleotide poly' morphism, or SNP. Humans share most of their genetic material: less than one percent of DNA sequences differ between two unrelated individuals. Therefore, **genome'wide** scans are conducted by genotyping DNA markers where it is known that there is significant variation across individuals. The markers are also selected to be fairly evenly distributed across the entire genome, and since SNPs that are close to one another tend to be highly correlated, unobserved SNPs can usually be imputed with high accuracy. It is for these reasons that

the term **genome'wide** scan is used to describe the genotyping of approximately 500,000 SNPs, despite the fact that the human genome is comprised of approximately three billion nucleotide base pairs.

Genes are sequences of nucleotide base pairs which code for proteins or RNA products. Following this process, these proteins and RNA products begin a cascade of interactions that regulate bodily structures and functions. While in some rare cases one allele can single' handedly lead to a disease (such as Huntingtonos disease), the vast majority of phenotypes are ppolygenicq, meaning they are influenced by multiple genes (Mackay 2001). Moreover, many traits of interest to economists are several links removed from the original genotype

in the chain of causation. Therefore, it would almost certainly be inaccurate to expect to find genes with a proximal effect on **educational attainment**. Instead, a more plausible hypothesis is that for complex socioeconomic outcomes, genetic effects are environmentally mediated. That said, simple **association** models between candidate SNPs and phenotype are still useful because the main effects discovered suggest areas for further exploration of mediation and moderation effects, and in some cases they point to the biological systems (e.g. the dopaminergic system) that are likely to influence the outcome being studied.

4 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 Heart **Study** data; then, we describe the methods used for the second, replication stage of the GWAS, with the Rotterdam **Study** data.

4.1 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 geno' typing 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.

4.1.1 Preliminary Steps for the GWAS

Following usual practices (Manolio et al, 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 pmissingnessqlarger than 0.05. An individualos missingness is the fraction of the SNPs in the employed array with missing data for 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 (Manolio et al, 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 missing' ness 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 pminor allele frequencyq). Coeffi cients 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 &%-& level. The null hypothesis of this test is that the observed genotype frequencies are equal to their theoretical expecta' tions 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 (Manolio et al, 2008; Sullivan and Purcell, 2008).

From the 500,568 SNPs on our Affymetrix 500k array, 76,764 did not satisfy the miss' ingness 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[4].

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

4.1.2 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 strat' ification. 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 compo' nents 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 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[5].

4.1.3 **Association** Analysis

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

$$Edu / \beta_1 \$ \beta_" \quad SNP' \$ PC \beta_\# \$ X \beta_\$ \$ E, \tag{1}$$

where Edu is years of education, SNP' 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[6].

Two complications arise when doing inference. The first is that the matrix ! # E0EE/1 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 diffi culty 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.

[5]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.

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

We discuss these issues briefly in turn.

4.1.4 Modeling the Error Structure

We specify a parametric structure on the matrix ! to account for the nonindependence of the error terms across individuals. In what follows, the subscripts i or j refer to individuals, / % (&, ..., F) indexes families, and g % (&,',() refers to the three generations in the data.

First, assume that the error terms of individuals from different families are independent. We can write ! /diag"!$,!#, ...,!$ #, where !+ / E0E+ E/+ 1 is the covariance matrix of the error terms for individuals in family /. To model the correlation structure of !+ , 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,

$$E / \sigma \text{"} aA_{-'\%\&_s} \$ cC \$ eE\#, \quad (2)$$

where σ / !σ#, σ#/ var"E#, and $A_{-'\%\&_s}$, C, and E are, respectively, the latent additive genetic (with SNP' partialled out), common environmental, and individual environmental factors underlying **educational attainment**. To identify the model, we assume without loss of generality that the variables $A_{-'\%\&_s}$, C, and E are standardized to have mean % and unit variance. This implies that a#, c# and e# sum to one.

The latent variable $A_{-'\%\&_s}$ 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# does not vary across genera' tions: σ \,)" / σ \,)# / σ \,)$/ σ . We also note that E0E*g1 $ E0E1/% &g, since controls for age are included in (1).

Biometrical genetic theory implies that, if mating is random,

$$E0A-'\%\&_{s}\,!-,A-'\%\&_{s}\,!.1 / r-., \tag{3}$$

where r-. is Sewall Wrightos coeffi cient of relationship. Wrightos coeffi cient 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 / " $_{\#}$, and likewise for a parent and his/her offspring; for a grandparent and his/her grandchild, r / " $_{\%}$; and for cousins, r / " $_{(}$. 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,

$$E0C-.,C.,"1 / 7, \tag{4}$$

where i is the father or the mother of j. From these assumptions, it is possible to work out

the entire correlation structure of !+ ; the results are shown in Table I.

4.1.5 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 %.%*/(+(,,,+ / &.(, &%-. 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 independent7. The most utilized threshold in the literature for large GWASos based on 500,000'SNP array data was set by the Wellcome Trust Case Control Consortium at * &%- (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.

4.1.6 Estimation Procedure

We ran at total of 363,776 regressions, one for each individual SNP. Properly account' ing for the correlation structure of the error term in each of these regressions would have

[7]When two SNPs are correlated, geneticists say that they are in Vlinkage disequilibriumV.

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 $\hat{\beta}_{"}$ and $2\sigma_{\#}$, but the standard errors of these estimates are downward biased.

Next, we kept the 98 SNPs whose Bonferroni'corrected p'values for $\hat{\beta}_{"}$ were significant at the five percent level and obtained consistent estimates of the standard error of $\hat{\beta}_{"}$ 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 $a_{\#}$, $c_{\#}$, and 7 by solving the following system of 3 equations with 3 unknowns:

$$\begin{aligned} 2\varrho_{\$'}\text{"E-,E.*i, j are full siblings\# / "} &\quad {}_{\#}\, 2a_{\#} \$\ 2c_{\#} \\ 2\varrho_{\&\#}\text{"E-,E.*i, j are parent'child\# / "} &\quad {}_{\#}\, 2a_{\#} \$\ 272c_{\#} \\ 2\varrho_{")\#}\text{"E-,E.*i, j are Aunt/uncle'nephew/niece\# / "} &\quad {}_{\%}2a_{\#} \$\ 272c_{\#} \end{aligned} \tag{5}$$

From this, we obtained $2!_{+}$, &/ / &..F, as well as the following consistent estimator of the variance covariance matrix of the regression coeffi cients:

$$\mathrm{var}(2\beta) \approx \mathrm{var}(\beta)^2 + 2\mathrm{cov}(\beta,\beta)$$

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

4.2 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.

4.2.1 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 **Study** 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[8].

After quality control, the filtered data were used to compute the first 10 principal com' ponents 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 suffi cient 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.

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

4.2.2 **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 & &%-& 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 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[9]. 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).

5 Results

5.1 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 coeffi cients.

[9]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 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 pwinneros curseqtype of selection bias (Zhong and Ross, 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 coeffi cients 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 * &%- established by the Wellcome Trust Case Control Consortium. As shown in the fourth column, none of the SNPs survive a Bonferroni correction at the ten percent level, the two lowest Bonferroni'corrected p'values being 0.11[10]. The top two hits rrs11758688 and rs12527415 rare 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 psignificantqSNPs 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.

[10]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.

5.2 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 **Study** data. The first column reports the rs number of the SNP and the chromosome number in parentheses. The second column contains the estimated beta

coeffi cients. 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 sta' tistically significant **association** with **educational attainment** in the Rotterdam **Study** data. In fact, the signs of the estimated beta coeffi cients from the first stage and the replication stage are only identical for 9 of the 20 SNPs.

6 Multiple Testing, False Positives, and Power Consid! erations

Genetic **association** studies are becoming increasingly common in the social sciences, in' cluding economics. Many of the studies carried out to date have examined the relationship between some economic phenotype ' typically an experimentally elicited preference parame' ter ' and a relatively small number of genetic markers. The markers in these studies are generally selected based on some a priori hypothesis derived from information about their biological function. These studies are therefore known as pcandidate'geneqstudies. For instance, because the dopamine receptor D% (DRD4) gene affects dopamine receptors in the brain and because dopamine plays an important role in learning and the processing of reward (Schultz, Dayan and Montague, 1997), many researchers initially suspected that variation in the DRD4 gene translates in variation in risk preferences. Indeed, two early genoeconomics studies, Dreber et al. (2009) and Kuhnen and Chiao (2009), both reported an **association** between a particular variant of DRD4 and experimentally elicited risk preferences. Carpen'

ter et al. (2010) subsequently failed to replicate this finding, however, and in fact reported

a borderline significant **association** in the opposite direction.

Several other hypothesis'based **association** studies of economic phenotypes have now appeared in the literature (Israel et al., 2008; Roe et al., 2009; Zhong et al., 2009a, Zhong et al., 2009b), only one of which replicated the main result in an independent sample prior to publication (Israel et al., 2008). Nevertheless, a replication attempt of the Israel et al. (2008) results using a different **study** population was not successful (Apicella et al., 2010). Failed replications are not unique to economics. After the decoding of the Human Genome Project in the early 2000s, there was a pgold rushqto find the genetic markers that correlate with important diseases. Unfortunately, later meta'analyses and review studies revealed that nearly all of these associations failed to replicate (Ioannidis, 2005; Ioannidis, 2007). The most plausible interpretation of this fact is that the initially reported findings were spurious, though it is possible that some of the published associations were in fact true only in the particular sample in which they were originally found because of treatment effect heterogeneity or that they were falsely non'replicated (Ioannidis, 2007).

The abundance of false positives can be attributed to two main factors. The first is that most **association** studies are seriously underpowered, and thus the probability that an **association study** will detect a true signal is vanishingly small. This point is now close to universally accepted in the molecular genetics community. Accumulating evidence from the genetics literature suggests that most true causal variants have very small effect sizes (Sklar et al. 2009, Manolio et al. 2009, Visscher 2008), and that for most complex traits few, if any markers have an R^2 greater than 1%. In the case of height, which is one of the most widely studied traits with a very high heritability, the top 40 SNPs that have been found explain only about 5% of the variance (Manolio et al., 2009). Several non'replicated published results notwithstanding, there is no reasons to expect traits of interest to economists to be any different in this respect. In the case of traits such as behavior in economic games, the situation may be even worse. To illustrate, Figure 1 shows power graphs for different

significance levels and population effect sizes. It is interesting to note that for a marker with a R^2 of 0.1%, a sample of about 4,000 subjects is required to have power of 50% at the a / %.%* level. An R^2 of 0.1%, which implies a correlation of about %.%(between the marker and the phenotype, is not implausibly low, in our opinion, because economic preferences are very distal from genes in the chain of causation and are likely measured with considerable noise. Even for a marker with a very large R^2 of 1% ' larger than the R^2os of all the markers that have so far been found to predict height ', a sample of about 780 subjects is needed to have power of 80% at the a / %.%* level ' a significance threshold that is much larger than what is generally relevant in **association** studies given the multiple testing involved.

The second main factor is the problem of multiple hypothesis testing. Researchers go to great pains to assemble datasets with genotypic information. Adding rich phenotypic information to such datasets is relatively inexpensive. With little theory to discipline the empirical work, there is a great risk that the p'values reported in the final manuscripts do not fully take into account the number of hypotheses that were tested before the final specification was arrived at. Besides the obvious need to adjust p'values for the number of phenotypes and markers tested, there is the additional problem of model selection uncertainty. For example, if additive and non'additive models are both estimated, but only findings from the model with lower p'values are reported, the resulting inference will of course be incorrect. If any of these problems is not fully accounted for when results are entered into the scientific record, the reported p'values will be diffi cult to interpret.

The results of this paper, in which initially promising results failed to replicate, are particularly humbling. They highlight the pitfalls of multiple testing ' even when proper correction is made for it ' and suggest that published but non'replicated associations should be approached with great caution until they have been corroborated in independent sam' ples. Economists are of course aware of the problem of multiple hypothesis testing (see e.g. Leamer, 1983), but the scale of the problem in the context of molecular genetic data use is

quite daunting and, in our opinion, not suffi ciently appreciated. The problem is only going

to be exacerbated as more and more sophisticated sequencing technologies are developed.

7 Discussion

In this paper we have provided an illustration of how **genome'wide association** techniques can be used to **study** economic phenotypes. We have also emphasized the numerous pitfalls that arise in a **study** of this kind. Our initial examination of the Framingham data revealed a number of markers with psuggestiveqassociations with **educational attainment**. These results were obtained by accounting for the family structure of the data in standard error estimation and using modern methods to deal with the problem of population stratification. This paper is, to our knowledge, the first to use molecular genetic variants in a large population sample to attempt to predict **educational attainment**.

Despite initially promising results, an attempt to subsequently replicate these results in a large, independent sample failed. Our power analyses show that for plausible effect sizes of individual alleles, very large samples will be needed to detect true causal variants predicting socioeconomic outcomes. This conclusion is entirely consistent with an emerging consensus in genetic epidemiology, according to which most diseases are caused by a large number of genes with small effects.

Studies of twins, adoptees and other sibling types show that a number of socioeconomic indicators, including **educational attainment**, are heritable and it seems highly plausible that such markers will eventually be identified. A legitimate question is therefore how the dis' covery of genetic markers that correlate with socioeconomic outcomes can advance economic research and influence policy. We believe that molecular genetic data will ultimately serve

as a great aid to economists wishing to obtain a better understanding of individual'level het' erogeneity in economic behavior. Knowledge of genetic risk factors, and an understanding of the relevant biological pathways, may ultimately prove helpful in designing interventions that target vulnerable individuals and improve their labor market outcomes (Benjamin et

al., 2007). Such basic science may also prove useful in providing consumers with better infor' mation about their genetic endowments. The idea of genetic testing is common in medicine, where genetic traits are important in the diagnosis (and prognosis) of a variety of conditions including Huntingtonos disease and breast cancer (Walker, 2007; Bouchard et al., 2002). More basic science is needed before molecular genetic data can provide an individual with a significant amount of information about her paptitudesqor some other heritable traits that is rewarded on labor markets. However, this may change with the next generation of full genome scans. Though the ramifications of such developments are likely to be broad and on the whole positive, they also raise a host of ethical problems about the use of the additional information, especially in labor and insurance markets (Tabarrok, 1994; Benjamin et al., 2007).

Finding the genes that affect economic preferences and outcomes can also be seen as a first step toward understanding the biological pathways though which those genes express themselves and this, in turn, could lead to a better understanding of the biological basis of economic decision making. A better understanding of the biological pathways may help in' spire more comprehensive theories about economic decision making and economic outcomes. For example, there is a heterogeneous collection of mechanisms that could potentially explain the heritable variation in income (Bowles et al., 2005). Molecular genetic data may help shed light on the complicated pathways from genes to complex socioeconomic outcomes.

A better understanding of the biological pathways may also have other positive conse' quences beyond their use in designing treatments or targeting interventions aimed to help vulnerable individuals. Consider, for example, the use of molecular genetic data to detect omitted variable bias. It has long been known that differences in genetic endowments across individuals might bias estimates of causal effects, as genes are typically unobserved (Taub' man, 1976b).[11] As genetic and biological markers are now being included in major social

[11] One proposed solution to this problem is within'family estimation. However, several authors have noted that such an approach is only an improvement over cross'sectional estimates under conditions which may not hold in practice (Griliches, 1979; Bound and Solon, 1999).

science surveys, such as the National Longitudinal **Study** of Adolescent Health (Add Health), the Wisconsin Longitudinal **Study** (WLS), and the Health and Retirement Survey (HRS), it may soon be feasible to directly control for previously unobserved differences in genetic endowments in standard regressions.

A second empirical use of genetic data relates to exploiting the random component of genetic markers inheritance by using them as instruments to better estimate causal effects. This use of genetic markers was anticipated by Davey'Smith (2002) and is already used in economic analyses (Ding et al., 2007; Norton and Han, 2008). As with any instrumen' tal variable analysis, the challenge is to show convincingly that the exclusion restriction holds. Given that many genes are pleiotropic, the plausibility of any analysis which uti' lizes genes as instruments hinges critically on how well the function of the genetic variant is understood. Economists would likely be less skeptical that a particular gene satisfies the exclusion restriction if the biological pathways were better understood. For example, if one wanted to instrument for smoking behavior in an education production function, the use of a genetic variant affecting nicotine metabolism in the liver such as the CYP2A6 gene (Audrain'

McGovern et al., 2007) would seem a priori to be a better instrument than a variant in the DRD2 gene, which affects reward systems in the brain (Lerman et al., 2006).

8 Conclusion

Studies of twins, adoptees and other sibling types show that a number of socioeconomic indicators, including **educational attainment**, are heritable. The advent of inexpensive, **genome'wide** scans of variation is now making it increasingly feasible to directly examine specific genetic variants that predict individual differences. In this article, we have provided an illustration of such an analysis. Applying the GWAS technique to over 360,000 SNPs throughout the genome in a family'based sample of over 8,500 individuals, we identified a number of promising markers. However, our attempt to replicate these findings in an inde'

pendent sample were not successful, prompting us to conclude that the original associations were probably spurious.

We believe that there are a number of valuable lessons for economists interested in **study**' ing the molecular genetic basis of economic behaviors and outcomes. It would benefit the genoeconomics enterprise if economists and other social scientists learn from the mistakes that have been made in genetic research in the past decade, rather than repeat them in our own discipline. In particular, our results add weight to the hypothesis that the genetic variance of most complex traits is highly diffuse throughout the genome and that no common markers covered by existing genotyping platforms have large effects on economic outcomes and behaviors. This is consistent with what has been found for most other traits. An impor' tant implication of this result is that future **association** work in economics and other social

sciences should use much larger samples than what has so far been the norm. It is likewise crucial that the p'values reported in papers be adjusted to take into account the full range of hypotheses that were tested, including pretesting of various model specifications. Most importantly, results should be replicated in independent samples. If these methodological challenges are taken seriously, we are optimistic that the burgeoning field of genoeconomics will add a new dimension to our understanding of heterogeneity in economic outcomes.

9 REFERENCES

Abecasis, Goncalo, Yun Li, and Cristen Willer, Metal, available at http://www.sph.umich.edu/csg/abecasis/Metal/ (2007).

Affymetrix, 2009, Affymetrix R " **Genome'Wide** Human SNP Array 5.0. Available at www.affymetrix.com/support/technical/datasheets/genomewideRsnp5Rdatasheet.pdf.

Altshuler David, Lisa D Brooks, Aravinda Chakravarti, Francis S Collins, Mark J Daly and Peter Donnelly, pA haplotype map of the human genome,qNature 2005 (437), 1299'1320.

Angrist, Joshua and Alan Krueger, pEmpirical Strategies in Labor Economics,qChapter

23 in O. Ashenfelter and D. Card, eds., The Handbook of Labor Economics, Volume III, North Holland, 1999.

Apicella, Coren L, David Cesarini, Magnus Johanneson, Christopher T Dawes, Paul Lichtenstein, Björn Wallace, Jonathan Beauchamp and Lars Westberg, pNo **association** be' tween oxytocin receptor (OXTR) gene polymorphisms and experimentally elicited social preferences,qPlos ONE, forthcoming.

Audrain'McGovern, Janet, Nael Al Koudsi, Daniel Rodriguez, E Paul Wileyto, Peter

G Shields and Rachel F Tyndale, pThe Role of CYP2A6 in the Emergence of Nicotine Dependence in Adolescents,qPediatrics, 119 (2007), 264'274.

Becker, Gary S, Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education. (New York, NY: Columbia University Press, 1975).

Becker, Gary S and Nigel Tomes, pAn Equilibrium Theory of the Distribution of Income and Intergenerational Mobility,qJournal of Political Economy, 87 (1979), 1153r1189.

Behrman, Jere R and Paul Taubman, 1989, pIs Schooling pMostly in the Genesq? Nature' Nurture Decomposition Using Data on Relatives,qJournal of Political Economy, 97 (1989), 1425'1446.

Benjamin, Daniel J, Christopher F Chabris, Edward L Glaeser, Vilmundur Gudnason, Tamara B Harris, David Laibson, Lenore Launer, and Shaun Purcell, pGenoeconomics,q in Biosocial Surveys, Maxine Weinstein, James W. Vaupel, and Kenneth W. Wachter, eds. (Washington, D.C.: The National Academies Press, 2007).

Björklund, Anders, Mikael Lindahl, and Erik Plug, pThe Origins of Intergenerational Associations: Lessons from Swedish Adoption Data,qQuarterly Journal of Economics, 121 (2006), 999r1028.

Bouchard, Louise, Ingeborg R. Blancquaert, François Eisinger, William D. Foulkes, Gareth Evans, Hagay Sobol, and Claire Julian'Reynier, pPrevention and genetic testing for breast cancer: variations in medical decisions,qSocial Science & Medicine, 58 (2004), 1085r1096.

Campbell, Catarina D, Elizabeth L Ogburn, Kathryn L Lunetta, Helen N Lyon, Matthew L Freedman, Leif C Groop, David Altshuler, Kristin G Ardlie., and Joel N Hirschhorn, pDemonstrating stratification in a European American population,qNature Genetics, 2005

(37), 868'872.

Carpenter, Jeffrey, Justin Garci and J Loji Lum (2009). pCan dopamine receptor genes explain economic preferences and outcomes?qManuscript.

Cesarini, David, Christopher T. Dawes, Magnus Johannesson, Paul Lichtenstein and Björn Wallace, pGenetic Variation in Preferences for Giving and Risk'Taking,qQuarterly Journal of Economics, 124 (2009), 809r841.

Cesarini David, pEssays on Genetic Variation and Economic BehaviorqPhD Dissertation, Massachusetts Institute of Technology, 2010.

Crisan, Liviu G, Simona Pana, Romana Vulturar, Renata M Heilman, Raluca Szekely, Bogdan Druga, Nicolae Dragos and Andrei C Miu (2009). pGenetic contributions of the sero' tonin transporter to social learning of fear and economic decision making,qSocial Cognitive and Affective Neuroscience, 4 (2009), 399'408.

Diabetes Genetics Initiative, pGenome'Wide **Association** Analysis Identifies Loci for Type 2 Diabetes and Triglyceride Levels,qScience 316 (2007), 1331'1336.

Ding Weili, Steven F. Lehrer, J. Niels Rosenquist and Janet Audrain'McGovern, pThe Impact of Poor Health on Academic Performance: New Evidence Using Genetic Markers,q Journal of Health Economics, 28 (2009), 578'597.

Dreber, Anna, Coren L. Apicella, Daniel T. A. Eisenberg, Justin R. Garcia, Richard Zamore, J. Kuji Lum and Benjamin C. Campbell, pThe 7R Polymorphism in the Dopamine Receptor D4 Gene (DRD4) Is Associated with Financial Risk'Taking in Men,qEvolution and Human Behavior, 30 (2009), 85r92.

Estrada, Karol, Anis Abuseiris, Frank G. Grosveld, André G. Uitterlinden, Tobias A. Knoch and Fernando Rivadeneira, pGRIMP: a web' and grid'based tool for high'speed analy' sis of large'scale **genome'wide association** using imputed data, qBioinformatics, 25 (2009),

2750s 2752.

Feldman, Marcus W, Otto, Sarah P. and Freddy B. Christiansen, 2000, Genes, Culture and Inequality, in Arrow, Kenneth, Samuel Bowles and Steven Durlauf, eds. Meritocracy and Economic Inequality. (Princeton, NJ: Princeton University Press, 2000).

Goldberger, Arthur S, pHeritability,qEconomica, 46 (1979), 327r47.

Griliches, Zvi, pSibling Models and Data in Economics: Beginnings of a Survey.qJournal of Political Economy, 87 (1979), 37'65.

Hamer, Dean H. and Leo Sirota, pBeware the chopstick gene,qMolecular Psychiatry 5 (2000), 11r13.

Harper Sarah J. and Philip LoGrasso, pSignaling for survival and death in neurones: the role of stress'activated kinases,qCellular Signaling 13 (2001), 299r310.

Hemminger Bradley M, Billy Saelim and Patrick F. Sullivan, pTAMAL: An integrated approach to choosing SNPs for genetic studies of human complex traits,qBioinformatics 22 (2006), 626'627.

Hill, William G., Goddard, Michael E and Peter M. Visscher, pData and theory point to mainly additive genetic variance for complex traits,qPLoS Genetics 4 (2008).

Hinrichs A, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, Hillman'Jackson J, Kuhn RM, Pedersen JS, Pohl A, Raney BJ, Rosenbloom KR, Siepel A, Smith KE, Sugnet CW, Sultan'Qurraie A, Thomas DJ, Trum' bower H, Weber RJ, Weirauch M, Zweig AS, Haussler D, Kent WJ. pThe UCSC Genome Browser Database: update 2006,qNucleic Acids Research 34 (2006), D590rD598.

Hofman, Albert, Monique M. B. Breteler, Cornelia M. van Duijn, Harry L. A. Janssen, Gabriel P. Krestin, Ernst J. Kuipers, Bruno H. Ch. Stricker, Henning Tiemeier, André G. Uitterlinden, Johannes R. Vingerling, and Jacqueline C. M. Witteman, pThe Rotterdam **Study**: 2010 objectives and design update,qEuropean Journal of Epidemiology 24 (2009), 553s 572.

Ioannidis, John PA, Evangelia E Ntzani, Thomas A Trikalinos and Despina G Contopoulos'

Ioannidis, pReplication validity of genetic **association** studies,qNature Genetics 29 (2001), 306r309 .

Ioannidis, John PA (2005) Why Most Published Research Findings are False, PLoS Medi' cine, 2: e124.

Ioannidis John PA, Patsopoulos NA, Evangelia E Ntzani, (2007) Heterogeneity in meta' analyses of **genome'wide association** investigations. PLoS ONE 2(9): e841.

Israel S, Lerer E, Shalev I, Uzefovsky F, Riebold M et al. (2009), pThe oxytocin receptor (OXTR) contributes to prosocial fund allocations in the dictator game and the social value orientations task,qPLoS One 4: e5535.

Jencks, Christopher (1980), pHeredity, environment, and public policy reconsidered,q American Sociological Review 45, 723r736.

Jencks, Christopher and Marsha Brown (1977). Genes and social stratification. In Ki' nometrics in Paul Taubman, ed. (Amsterdam: North'Holland.)

Knafo, Ariel, Salomon Israel, Ariel Darvasi, Rachel Bachner'Melman, Florina Uzefovsky, Lior Cohen, Esti Feldman, Elad Lerer, Efrat Laiba, Yael Raz, Lubov Nemanov, Inga Grit' senko, Christian Dina, Galila Agam, Brian Dean, Gary Bornstein and Richard P. Ebstein, pIndividual Differences in Allocation of Funds in the Dictator Game Associated with Length of the Arginine Vasopressin 1a Receptor (AVPR1a) RS3 Promoter Region and Correlation between RS3 Length and Hippocampal mRNA,qGenes, Brain and Behavior, 7 (2008), 266r 275.

Kuhnen, Camelia M and Joan Y. Chiao, pGenetic Determinants of Financial Risk Tak' ing,qPLoS ONE, 4 (2009).

Leamer, Edward E, pLetos Take the Con out of Econometrics,q American Economic Review, 73 (1983), 31r43.

Lerman Caryn, Christopher Jepson, E Paul Wileyto, Leonard H. Epstein, Margaret

Rukstalis, Freda Patterson, Vyga Kaufmann, Stephanie Restine, Larry Hawk, Ray Niaura and Wade Berrettini, pRole of functional genetic variation in the dopamine D2 receptor

(DRD2) in response to bupropion and nicotine replacement therapy for tobacco dependence: results of two randomized clinical trials,qNeuropsychopharmacology 31(2006), 231'42.

Li, Yun and Gonçalo R. Abecasis, pMach 1.0: rapid haplotype reconstruction and missing genotype inference,qAmerican Journal of Human Genetics S79(2006), 2290.

Mackay, Trudy FC, pThe genetic architecture of quantitative traits,qAnnual Review of Genetics, 35(2001), 303r339.

Manolio Teri A, Laura Lyman Rodriguez, Lisa Brooks, Goncalo Abecasis, pNew Models of collaboration in **genome'wide association** studies: The Genetic **Association** Information Network,qNature Genetics 39 (2007),1045r1051.

Miller, Paul, Charles Mulvey and Nick Martin, pGenetic and environmental contributions to **educational attainment** in Australia,qEconomics of Education Review, 20 (2001), 211r 224.

Norton, Edward C. and Euna Han, pGenetic Information, Obesity and Labor Market Outcomes,qHealth Economics, 17 (2008), 1089r1104.

Price, Alkes L., Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick and David Reich, pPrincipal components analysis corrects for stratification in **genome'wide association** studies,qNature Genetics, 38 (2006), 904r909.

Purcell Shaun, Benjamin Neale, Kathe Todd'Brown, Manuel A.R. Ferreira, David Ben' der, Julian Maller, Sklar P, Paul I.W. de Bakker, Mark J Daly and Pak C. Sham., pPLINK: A Toolset for Whole'Genome **Association** and Population'Based Linkage AnalysisqAmerican Journal of Human Genetics, 81 (2007), 559'75.

Purcell, Shaun, PLINK 1.05, available at http://pngu.mgh.harvard.edu/purcell/plink (2008).

Roe, Brian E, Michael R Tilley, Howard H. Gu, David Q Beversdorf, Wolfgang Sadee, Timothy C Haab, Audrey C Papp (2009), pFinancial and psychological risk attitudes asso' ciated with two single nucleotide polymorphisms in the nicotine receptor (CHRNA4) gene,q PLoS ONE, e6704.

Sacerdote, Bruce, pHow Large Are the Effects from Changes in Family Environment? A **Study** of Korean American Adoptees,qQuarterly Journal of Economics 122 (2007), 119'157.

Sacerdote, Bruce, forthcoming, Nature And Nurture Effects On Childrenos Outcomes: What Have We Learned From Studies Of Twins And Adoptees?, in Handbook of Social Economics (North Holland, Amsterdam, NL).

Schultz, W., Dayan, P. & Montague, P. R. (1997), pA neural substrate of prediction and reward,qScience, 275(5306), 1593'1599.

Sullivan, Patrick F and Shaun Purcell, pAnalyzing **genome'wide association study** data: a tutorial using PLINK,qin Statistical Genetics; Gene Mapping Through Linkage and As' sociation, Neale, Benjamin M Neale, Manuel AR Ferreira, Sarah Medland and Danielle Posthuma, Editors; Taylor & Francis Group: New York (2008), 355'394.

Tabarrok, Alexander, 1994, pGenetic testing: An economic and contractarian analysis,q Journal of Health Economics 13, 75'91.

Taubman, Paul, pThe Determinants of Earnings: Genetics, Family and Other Environ' ments: A **Study** of White Male Twins,qAmerican Economic Review, 66 (1976a), 858r870.

Taubman, Paul. 1976b. pEarnings, Education, Genetics, and Environment,qJournal of Human Resources, 11(4): 447r461.

Walker, Francis, pHuntingtonos Disease,qLancet, 359 (2007), 218r228.

Watson, James D. and Francis H.C. Crick, pMolecular Structure of Nucleic Acids,qNa' ture, 171 (1953), 737r738.

Wellcome Trust Case Control Consortium, pGenome'wide **association study** of 14,000 cases of seven common diseases and 3,000 shared controls,qNature, 447 (2007), 661'678.

Wheeler David L, Tanya Barrett, Dennis A Benson, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Ron Edgar, Scott Federhen, Lewis Y Geer, Wolfgang Helmberg, Yuri Kapustin, David L Kenton, Oleg Khovayko, David J Lipman, Thomas L Madden, Donna R Maglott, James Ostell, Kim D Pruitt, Gregory D Schuler, Lynn M Schriml, Edwin Sequeira, Stephen T Sherry, Karl Sirotkin, Alexandre

---

Souvorov, Grigory Starchenko, Tugba O Suzek, Roman Tatusov, Tatiana A Tatusova, Lukas Wagner and Eugene Yaschenko. pDatabase resources of the National Center for Biotechnol' ogy Information,qNucleic Acids Research 2006 (34), D173'180.

Zhong, Hua, Ross L L. Prentice, pBias'reduced estimators and confidence intervals for odds ratios in **genome'wide association** studies,qBiostatistics, 9 (2008), 621'634.

Zhong, Songfa, Salomon Israel, Hong Xue, Pak C Sham, Richard P Ebstein and Soo Hong Chew (2009a), pA neurochemical approach to valuation sensitivity over gains and losses,qProceedings of the Royal Society B: Biological Sciences, 276, 4181'4188.

Zhong, Songfa, Salomon Israel, Hong Xue, Pak C Sham, Richard P Ebstein and Soo Hong Chew(2009b), pMonoamine oxidase A gene (MAOA) associated with attitude towards longshot risks,qPlosONE, 4(12), e8516.

10 Tables and Figures

TABLE I.

EXPECTED CORRELATIONS BETWEEN THE ERROR TERMS OF RELATED INDIVIDUALS

	E0A-A.1	E0C-C.1	E0E-E.1
Relatedness			
Full siblings	$\frac{"}{\#}$	&	$\sigma_{\#"\#}a_{\#}\$\ c_{\#\#}$
Half siblings	$\frac{"}{\%}$	$\frac{"}{\#}$	$\sigma_{\#"\%}a_{\#}\$\ \frac{"}{\#}\ c_{\#\#}$
Parent'child	$\frac{"}{\#}$	7	$\sigma_{\#"\#}a_{\#}\$\ 7c_{\#\#}$

Grandparent'grandchild	$\frac{"}{\%}$	$7_{\#}$	$\sigma_{\#}^{2}\,"_{\%}\,a_{\#}\$\,7_{\#}c_{\#}\#$
Full cousins	$\frac{"}{(}$	$7_{\#}$	$\sigma_{\#}^{2}\,"_{(}\,a_{\#}\$\,7_{\#}c_{\#}\#$
Half cousins	$\frac{"}{"\&}$	$\frac{"}{\#}\,7_{\#}$	$\sigma_{\#}^{2}\,"_{"\&}\,a_{\#}\$\,"_{\#}\,7_{\#}c_{\#}\#$
Aunt/uncle'nephew/niece	$\frac{"}{\%}$	7	$\sigma_{\#}^{2}\,"_{\%}\,a_{\#}\$\,7c_{\#}\#$
Half aunt/uncle'nephew/niece	$\frac{"}{(}$	$\frac{"}{\#}\,7$	$\sigma_{\#}^{2}\,"_{(}\,a_{\#}\$\,"_{\#}\,7c_{\#}\#$

NOTES: This table gives the assumed error structure for relatives in our sample. Full siblings have the same biological parents; phalf siblingsqshare one biological parent; pfull cousinsqhave the same two grand'parents on either the paternal or the maternal side; phalf cousinsqhave only one grandparent in common; phalf aunt/uncle'nephew/nieceqrefers 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	2064

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)	β̂	p'value	Bonferroni	Sample	Minor Allele	Nearby Genes?
rs11758688 (6)	'0.253 '.., &%-'		0.107	7572	T	Yes
rs12527415 (6)	'0.253 (.%(&%-'		0.109	7570	T	Yes
rs17365411 (2)	0.260 (.,(&%-'		0.134	7559	C	No
rs7655595 (4)	'0.266 (... &%-'		0.144	7486	G	No
rs17350845 (1)	'0.291 +." &%-'		0.224	7415	C	Yes
rs12691894 (2)	'0.246 +.+, &%-'		0.240	7572	G	No
rs9646799 (2)	0.271 ,.)& &%-'		0.267	7478	T	Yes
rs11722767 (4)	'0.257 ,.,, &%-'		0.280	7574	C	No
rs10947091 (6)	'0.245 ..%(&%-'		0.325	7574	T	Yes
rs6536456 (4)	0.230 &.(' &%-&		0.474	7513	C	No
rs1580882 (4)	0.229 &.)(&%-&		0.516	7556	T	No
rs6536463 (4)	0.229 &.)- &%-&		0.533	7571	G	No
rs1502720 (4)	0.228 &.++ &%-&		0.560	7566	C	No
rs10436961 (1)	'0.268 &.-' &%-&		0.657	7540	A	Yes
rs4845129 (1)	'0.265 '.%, &%-&		0.745	7546	G	Yes
rs17365432 (2)	0.257 '.(' &%-&		0.836	7573	G	No
rs11225388 (11)	0.261 '.*& &%-&		0.904	7559	G	Yes
rs7743593 (6)	0.301 '.+- &%-&		0.965	7545	C	Yes
rs10028331 (4)	'0.259 '..(&%-&		1.00	7565	G	No
rs11964691 (6)	0.307 (.'. &%-&		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).

39

Page 42

TABLE VI.

REPLICATION OF TOP 20 HITS IN THE ROTTERDAM **STUDY**

SNP (Chromosome)	β	β^2	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# data was produced separately for each **study** (RS'I, RS'II, and RS'III) and is therefore not

---

shown, but is available upon request.

Appendix ' Description of how the **educational attainment** vari' able was coded for the Framingham data

Members of the three cohorts routinely come to a central facility for detailed examina' tions and collection of survey data. Here we describe the construction of the **educational attainment** variable for the three cohorts.

Members of the Original Cohort were asked to indicate their highest **educational** attain' ment on a scale with nine categories. We assigned zero years of education to individuals who responded pNoneq, 2.5 years of education to individuals who responded pFourth grades or lessq, 6 years of education to individuals who responded pFifth, sixth or seventh gradeq, 8 years of education to individuals who responded pGrade school graduateq, 10 years of

education to individuals who responded pHigh school, not graduateq, 12 years to individuals who responded pHigh school graduateq, 14 years to individuals who responded pCollege, not graduateq, 14 years to individuals who responded pBusiness college, nursing, music or art schoolq, 16 years to individuals who responded pCollege graduateqand 18 years to individ' uals who responded pPost'graduateq.

For the offspring cohort our primary variable was the response to the question pHow many years of schooling did you complete?q, which was administered in the third exam' ination. Possible responses ranged from 1 to 16, with a seventeenth option available for individuals with 17 years of education or more. We assigned 18 years of education to indi' viduals who selected the p17 years of moreqto maintain consistency with the coding in other cohorts. When responses to this question was not available, we replaced, when possible, the missing value using responses to a question on **educational attainment** posed in the second examination (the two measures correlate at 0.9).

Finally, for the third generation cohort, we used responses to the question pWhat is the highest degree or level of school you have completed?q We assigned 4.5 years of education to individuals who responded pGrades 1'8q, 10 years of education to individuals who responded pGrades 9'11q, 12 years of education to individuals who responded pCompleted High School

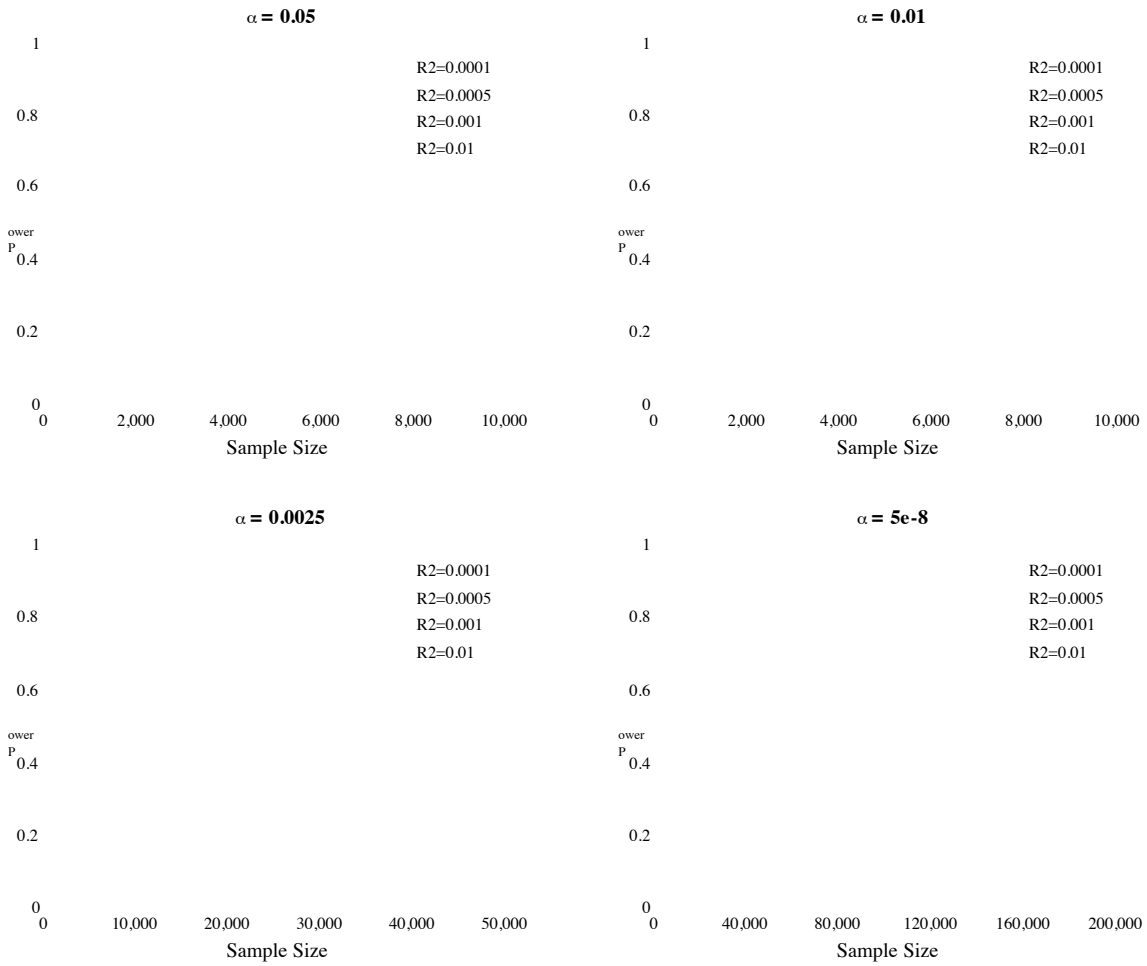
---

or GEDq, 14 years of education to individuals who responded pSome college but no degreeq, pTechnical School Certificateq or pAssociate Degree (Junior College, AA, AS)q. Finally, we assigned 16 years of education to individuals who responded pBacheloros Degree (BA, AB,BS)qand 18 years to individuals who responded pGraduate or Professional Degreeq.

Page 46

FIGURE I

POWER GRAPHS



NOTES: These graphs show the power to detect a true signal for a marker that explains a given fraction of the variance ($R_{\#}$) as a function of the sample size, for different significance thresholds.