Are Genetic Markers of Interest for Economic Research?

Steven F. Lehrer
Queen's University, NYU-Shanghai and NBER
sl164@nyu.edu

Weili Ding*
Queen's University and NYU Shanghai
dingw@queensu.ca

July 2016

Abstract

The idea that genetic differences may explain a multitude of individual-level behaviors and outcomes as studied by economists, is more than a bit controversial. Since an increasing number of datasets now contain measures of genetic variation, it is reasonable to postulate that incorporating genomic data into economic analyses will become increasingly common. However, there remains much debate among academics as to: First, whether and how ignoring genetic differences in empirical analyses would bias the resulting estimates; Second, since genetic characteristics are largely immutable, what types of policy guidance, if any, the incorporation of these variables into economic analyses may yield. In this paper, we revisit these concerns and survey the main avenues by which empirically oriented economic researchers have utilized measures of genetic markers to improve our understanding of economic phenomena. We discuss the strengths, limitations and potential of existing approaches and conclude by highlighting several prominent directions forward for future research.

Keywords: genetic markers; gene environment interactions; genome wide association studies; individual differences; intergenerational transmission

* We are extremely grateful to David Cesarini, Jason Fletcher, Susumu Imai and J. Niels Rosenquist for often heated conversations over the past twelve years that have helped us to think critically on whether and how economists should incorporate molecular genetic data within their analysis. We are also grateful to Arnaud Chevalier for his patience and support in preparing this draft. Similar to an Academy Awards acceptance speech, we simply do not have the space to list everyone with whom we have had informative debates including anonymous reviewers. Lehrer also would like to thank SSHRC for research support. We are responsible for all errors, glaring omissions and imprecise interpretations of evidence in the literature.

1. Introduction

It would not be an exaggeration to say that the mere mention of the word genetics to an economist a decade ago could raise alarm. This alarm may have been a response in part to one recalling the general response to then Harvard President Lawrence H. Summers' January 14, 2005 speech at an economics conference when discussing the under-representation of female scientists at elite universities.[1] This alarm perhaps was also triggered by memories of events approximately one decade earlier when Richard Herrnstein and Charles Murray attracted substantial controversy following the 1994 publication of the book titled "The Bell Curve", which was popularly (mis-)interpreted as ascribing the link between race and IQ to genetic factors.[2] Even recently, economists working on issues related to genetic factors continue to attract interdisciplinary criticism. For example, Ashraf and Galor's (2013) arguing for the importance of genetic diversity in explaining national income per capita drew a series of harsh responses from a long list of prominent scientists and anthropologists.[3] These three independent episodes occurred in a 20 year span have clearly indicated the controversy that one may encounter when interpreting or accounting for genetic factors within economic analyses. Thus, it would be unsurprising if individual researchers today would conclude that it is best to ignore genetic factors since the potential cost from subsequent criticism and potential damage to one's academic reputation could greatly outweigh any benefit one may receive from incorporating them. In short, genetic information becomes the hornet's nest that many economists stay away from.

In this paper, we argue that this would exactly be the wrong response. Not only has the role of empirical work in economics increased sharply over the last 20 years, but there is now a growing number of datasets that provide detailed information on genetic characteristics. Genetic markers are easily identifiable portions of our DNA code that are located at a specific known location on the chromosome. These measures are now being collected in multiple nationally representative social surveys and Conley (2009) suggests can be deployed to i) assess the direct impact of specific genetic influences on socioeconomic and behavioral outcomes, ii) explore genetic-environmental interactions, and iii) trace genealogies across time and space. Indeed, as

---

[1] A national media frenzy erupted focusing on an explanation that this under-representation may stem in part from "issues of intrinsic aptitude" between men and women, without considering the context in which the remarks were made. Summers' remarks made at the NBER Conference on Diversifying the Science & Engineering Workforce are posted online at http://www.harvard.edu/president/speeches/summers_2005/nber.php.

[2] Despite the hysteria at the time concerning this book and the link between IQ and genes, the authors (on page 311) made clear that while this explanation may hold water, they had no idea as to the importance. Specifically, (on page 311) in the concluding section they write, "If the reader is now convinced that either the genetic or environmental explanation has won out to the exclusion of the other, we have not done a sufficiently good job of presenting one side or the other. It seems highly likely to us that both genes and the environment have something to do with racial differences. What might the mix be? We are resolutely agnostic on that issue; as far as we can determine, the evidence does not yet justify an estimate."

[3] For example, d'Alpoim Guedes et al. (2012, 2013) present critiques against claims in Ashraf and Galor's (2013) and Ashraf and Galor (2012) replied to the first critique.

we discuss below, economists have done substantial applied research related to the first two themes. Further, simultaneous to the trend of growing importance of empirical economics, has been the development of a multitude of econometric strategies that exploit various research designs to identify causal impacts. These applied econometric methods have not solely transformed empirical practice within economics, but also in all other social science disciplines such as political science and sociology. We suggest that as economists increase their familiarity with genetic data, it is likely that they can develop methodological tools to generate new empirical strategies to shed light on the role of genetic factors that will be of interest to those within economics, as well as in many other scientific disciplines.

This paper can be viewed as an extension of the comprehensive reviews presented in Benjamin et al. (2007, 2012) and Lehrer (2015) that explore the use of genetic markers in studies within economics. While Benjamin et al (2007) coined the term "genoeconomics" for this field, the view that we advance is somewhat less ambitious. We argue that genetic markers are simply a new way to get inside the black box of individual permanent unobserved heterogeneity within numerous fields in economics. For example, in studies that explore labor supply, researchers often employ a fixed effect strategy to account for permanent unobserved differences in tastes or preferences across individuals. Similarly, when estimating wage, academic achievement or health outcome equations, researchers often employ fixed effects to capture permanent productivity characteristics of the individual. Genetic markers may be truly what is meant by permanent "unobserved" heterogeneity since they are assigned at conception, and (with the sole exception of monozygotic twins) differ markedly across individuals.[4]

While some economists have begun to incorporate data of genetic markers into their empirical analyses, their use remains scattered and limited to a handful of specific applications. This is somewhat surprising given the long history of economic research that explores how numerous traits and behaviors pass from one generation to the next. With data on genetic markers, perhaps one can gain understanding on how the transmission of genetic factors influences the transmission of outcomes. We should quickly state here that recent work by economists employing genetic data has attracted significant positive acclaims by researchers in other disciplines. Thus, our true aim of this survey is to reduce entry cost and hopefully attract other empirical economists to consider integrating genetic factors within their studies.

This paper is organized as follows. In Section 2, we provide a brief scientific primer on genetic terminologies. Section 3 reviews the four major strands of research in economics that has used genetic data to date. Section 4 proposes three directions for future research for economists. A concluding section summarizes our arguments and draws attention to how research using genetic data within economics is actually following trends in research within labor economics.

2. A Primer on Genetics

[4] Jencks (1980) may have been the first to point out that "genetic" does not imply "immutable". Thus, it may be the case that the effects of specific portions in an individual's DNA sequence on specific outcomes vary over the lifecycle, perhaps due to environmental stimuli. In a traditional fixed effects estimating equation, both the impact and stock of unobserved heterogeneity are assumed to be fixed over the period in which data is being analyzed.

In the Oxford Dictionary, the word "genome" is defined as a combination of the word "gene" and "chromosome". The genome is contained in all cells that have a nucleus and consists of more than 3.2 billion DNA base pairs located on 23 pairs of chromosomes. To help visualize the human genome, consider an instruction manual composed of 23 chapters (chromosomes) that is over 3.2 billion letters. The length of each chapter varies from 48 to 250 million letters (A, C, G, T) without any spaces. This genome that lies within each cell in our body is formed at conception when one member of each pair of chromosomes is inherited from the mother and the other from the father.

Using genetic data requires undertaking a molecular genetic approach to understand variation between individuals in the genetic code itself. This differs sharply from the approach in behavioral genetics that categorizes much earlier research in economics that aimed to understand the role of genetic factors using data collected from samples of twins or siblings. Briefly, and interested readers are referred to Behrman (2015) for a recent review of research using this approach, researchers begin by assuming that all variation in the outcome being investigated could be decomposed into additively separable genetic and environmental contributions. That is, the variance of a behavior being investigated is decomposed into three orthogonal components: additive genetic effects (A), common environment effects (C), and unique environment effects (E); hence the acronym ACE models.[5] If one further assumes that the same genetic or environmental factor has the same impact between monozygotic twins who share the same hereditary and environmental variables, within twin comparisons would thereby provide estimates of A+C. For dizygotic twins who share the same environmental variables and on average 50% of their genes, the same assumption on within twin comparisons would provide an estimate of ½*A*+*C*. Contrasting the monozygotic twins with dizygotic twins thus would isolate the hereditary effect *A*, by taking twice the difference between identical and fraternal twin correlations. Once *A* is obtained, *C* is then obtained by subtracting the estimated A from the identical twin correlation, whereas an estimate of E is given by subtracting the identical twin correlation from the number one. Within economics, Taubman (1976) is generally considered the first such study, which estimated that between 18% and 41% of variation in income across individuals was heritable.[6] Research using genetic data is now moving beyond variance decompositions between twins of different zygosity[7] into analyzing the impacts of specific portions of the genetic code.

[5] While not discussed in this paper due to space constraint, estimating heritability with data on twins has been taken under alternative assumptions including regression based methods (e.g., DeFries and Fulker, 1985), structural equation models (Boker et al., 2011), and generalized linear mixed models (Rabe-Hesketh et al., 2008).

[6] Jensen et al (1967) previously conducted a study published in a general interest scientific journal that tried to isolate the role of heredity, environment and luck in earnings.

[7] More recent research has shifted away from using data collected in traditional surveys to using data from either incentivized experiments or surveys, to explore the heritability in different measures of economic preferences (e.g. Wallace et al. (2007) and Cesarini et al. (2008, 2009, 2010, 2012). See Kohler et al. (2011) for a discussion of how to leverage twin-studies to model unobserved genetic endowments and causal pathways). In the section on gene by environment interactions (Section 3.4), we will discuss genome-wide complex trait analysis (GCTA), a method that uses restricted maximum likelihood estimation to estimate heritability from molecular genetic data.

Molecular genetics is the branch of genetics that studies the structure and function of DNA. The sequencing of the human genome in 2001 (Venter et al. 2001) has provided a means to measure genetic variation across individuals. One of the principal means through which genetic variation occurs is called a single nucleotide polymorphism (SNP); it refers to a mutation at a specific point in the genetic code where a single nucleotide is substituted (i.e. using the analogy before, for example, a single letter such as an A is substituted with a T at that point).[8] It has been estimated that there are only approximately two million sites on the genome where a SNP can be found. The genetic variants of SNPs are commonly referred to by the number of alleles. For example, at one of these specific locations, one's genotype can be denoted by the number of "risky" alleles (0, 1 or 2). Only a small minority of all of the known SNPs are considered to play important roles influencing the function and structure of the human body and these could be selectively advantageous or disadvantageous. In other words, while the human genome is over 3.2 billion chemical letters in length, less than 0.1% of these locations are believed to potentially account for observed differences in human behavior or outcome.

3. Categorizing Research by Economists Using Genetic Data

The majority of databases that contain genetic information were collected by medical scientists. However, a growing number of longitudinal databases that were designed for social scientists are adding genetic information. For example, the Add-Health Study has collected information on a few SNPs for primarily the sibling sample, the Health and Retirement Study has recently begun to make this information available from consenting participants, and the UK Biobank has linked genetic information to participants in the 1958 birth cohort study. To obtain measures of molecular genetic variation, a number of commercial entities have developed technologies that could measure several hundred thousand human SNPs simultaneously from blood or saliva samples.[9] Over the last decade, there have been a multitude of technological breakthroughs that not only make it easier to genotype more SNPs and other genetic variants, but to do so at lower costs. That said, in many datasets the genetic information provided is an *imputed* SNP, that is, it is calculated based on the high degrees of correlation between neighboring SNPs.[10] Prior to describing how such data could being utilized, it is important to point out that while some characteristics or health outcomes are known to be a unique result of a specific genetic difference, the majority of characteristics that economists are interested in are polygenic, meaning they are influenced by multiple genetic polymorphisms.

Before proceeding further, we would like to bring forth a controversial issue that some researchers in this area face by calling immutable characteristics such as SNPs, "treatments".

[8] Other sources of genetic variation could be mutations affecting repeated segments of DNA and include what are known as variable number of tandem repeat polymorphisms and copy number variation (CNV) polymorphisms. The interested reader is referred to a molecular genetics text for further reading on these mutations.

[9] There are methods that target the whole genome and others that are more selectively targeted. In general, the resulting data quality is heavily dependent upon the average number of times each base in the genome is actually 'read' during the sequencing process.

[10] The consequences of imputation have received scant attention thus far. It clearly generates measurement error, something that labor economists know a lot about and have a rich set of tools to offer to other researchers.

Many critics point to the impossibility of manipulating genetic traits that are fixed at conception in a manner that is analogous to administering a treatment in a randomized experiment. However, Greiner and Rubin (2011) argue that it is actually a matter of perception on those characteristics and perceptions are not as immutable.[11] Even without going all the way to the level of perception, if two individuals are the same in all important characteristics (age, gender, education, family background, residence, etc.) except for a specific SNP, then their difference in outcomes could still be attributed to this specific genetic difference, in which sense a specific SNP would be a legitimate "treatment" in the potential outcomes framework.

3.1 Candidate Gene Studies

Much of the earliest work by economists using genetic data is limited by the genetic information collected within the data being investigated. Generally, the initial genetic markers made available were those that at the time were hypothesized to be most important. These markers are called candidate genes. They were generally chosen to be genotyped (measured) since they were located in a particular chromosome region suspected of being involved in the outcome or its protein product may suggest that it could influence the outcome being investigated.

Many candidate gene studies in economics investigate whether specific SNPs correlate with measures of economic primitives such as risk aversion and delay discounting parameters, attempting to provide their biological micro foundations. Cesarini et al. (2009) suggest that approximately one-fifth of the variation in these measures is due to genetic factors. Initially, economists focused their candidate gene investigations on whether genes thought involved in the dopamine and serotonin system[12] in the brain's reward pathways explain primitives of behavior (e.g. Dreber et al., 2009, Kuhnen and Chiao, 2009).[13] While these early studies found some statistically significant associations, they were not replicated in samples of adolescents (Gee, 2014) and other samples analyzed by Carpenter et al. (2011) and Dreber et al (2011).

Associations between candidate genes and socio-economic outcomes have been undertaken in situations where the genetic basis for variation in outcomes was not established. For instance, DeNeve and Fowler (2015) and Kuhnen et al. (2013) respectively explore if there

[11] Specifically, Greiner and Rubin (2011) focus on circumstances under which race/gender can be appropriately called treatments. They argue that what causally explains gaps in outcomes between groups are not the groups themselves, but rather, are perceptions of the groups. In a genetic marker context, similar arguments could be made if employers or health insurers make decisions based on perceptions of the genetic characteristics of workers. While laws do exist in many countries, including the Genetic Information Nondiscrimination Act (GINA) in the United States, prohibit employment discrimination based on genetic information that forbid employers from asking about individuals' genetic information, including information about family members' health status, or family history. However, there are many reports that individuals voluntarily provide this information particularly to help develop workplace-based wellness programs and Baicker et al (2010) made a case this information provides benefits to employers.

[12] Dopamine and serotonin are two powerful neurotransmitters that affects one's mood and happiness. In general, neurotransmitters are chemical messengers which neurons use to tell other neurons that they have received an impulse.

[13] See Knafo et al (2008), Mertins et al (2011) and Zhong et al (2009) for studies that link specific genetic variants to outcomes measured in the laboratory.

existed statistically significant links between specific genetic markers and credit card debt and the number of credit lines opened. Since decisions on whether to issue a credit card as well as setting the credit limit are made by lenders and not by the borrowers themselves, whether there is a biological plausible mechanism underlying any such association should be justified. This is the practice within genetic epidemiology where researchers using candidate genes carefully explain how the putative candidate gene was chosen based on its relevance in the mechanism of the trait being investigated.

In short, studies that fall under the heading of candidate genes are likely undertaken based on convenience and have a poor track record when it comes to replication. Candidate genes studies also face concerns that they lack statistical power. Intuitively, if well-powered studies that search the entire genome for associations find only tiny effects, then the large effects found in many of these candidate gene studies with much smaller sample sizes are likely false positives.[14] We believe that despite the ease with which this research can be undertaken, candidate gene studies are unlikely to convince many in the research community.

An under-investigated aspect of candidate gene studies is whether the inclusion of genetic information changes the effects of other covariates. After all, if genetic factors are important, does their inclusion change estimates of other coefficients? In other words, is bias from omitted variables reduced? Or certain well-known covariates have approximated well for genetic factors? Answers to these question are important in understanding whether molecular genetic information is truly a valuable addition to many datasets.[15]

3.2 Moving Beyond Associations: Genetic Markers as Instruments

Perhaps the area that has attracted the most amount of debate among economists is whether or not genetic data can provide a source of exogenous variation to identify the impact of specific health conditions on socioeconomic outcomes. This source of identifying variation was first introduced in economics by Ding et al. (2009) who essentially used candidate genes as instruments to understand the impact of health outcomes on academic performance. Ding et al (2009)'s analysis underscores both the challenges researchers face when using genetic information as an instrument for specific health conditions and the need to investigate the sensitivity of one's conclusions to the identifying assumptions. We discuss these issues in further detail below.

The concept of comorbidity is well known in the medical sciences. It is defined as the simultaneous presence of two and more chronic poor health conditions in one individual. In

[14] Chabris et al., (2012) discuss the importance and challenge of replicating results from studies using genetic data.

[15] As we will discuss later, research focused on identifying genetic associations with outcomes of interest to economists has moved from using data collected on a few candidate genotypes to those measuring variation across the full genome. These studies have also proposed calculating genetic risk scores which are single variable measures that capture information contained in a multitude of SNPs. These covariates may help provide some preliminary information on the value of genetic information as a covariate but present difficulties in their interpretation.

empirical research, we are not provided with a single accurate measure of an individual's health, but rather proxies such as specific diagnoses. In their analyses, Ding et al (2009) show that using richer vectors of health conditions is important to identify the true effect of a specific condition since poor physical and mental health conditions are often positively correlated. By omitting the comorbid condition(s), different estimates may arise when using different estimators and specific instruments. The challenge of comorbidity has significant implications for researchers aiming to single out the role of a specific health condition and was first pointed out in economics due to this investigation using genetic information. Comorbidity also influences the general ways in which applied researchers select their instruments based on first stage relevance and whether they meet the exclusion restriction criteria.[16]

As with all studies that use instrumental variables to identify causal parameters, the plausibility of the (genetic) instrument comes into question. In a sense, one will never know whether a specific candidate gene is a valid IV since one cannot randomly assign genes to humans or create human equivalents to knockout mice but as discussed earlier the argument of Greiner and Rubin (2011) applies and a specific SNP could be a legitimate "treatment" in the causal framework. In addition, the role of individual genetic markers in many socioeconomic outcomes is likely quite small and likely explains less than 1% of the variation in that phenotype. This suggests that individual markers are likely weakly correlated.[17] Further, these genetic markers could be subject to the presence of dynastic effects since without more detailed data on parental outcomes and family environments (as well as parental genes), we cannot separate out the portion of the impact that is uniquely brought on by the child's gene.

Turning to the genetic marker itself, one may worry about population stratification[18] that there are subtle genetic differences between groups of individuals that are not accounted for and the gene being investigated is correlated with a missing genetic marker that is driving the results. Similarly, this may happen since genes located close together on the same chromosome are sometimes inherited as a group, so one may not be attributing the effect to the correct polymorphism. Given these potential confounders, researchers using genes as IVs could use the Conley et al. (2012) local to zero approximation sensitivity analysis.[19]

Conley (2009) argues that the phenomenon of pleiotropy presents a further challenge for the plausibility of a genetic instrument: since many genes code for proteins that may have multiple functions and effects, it is hard to know for certain that the instrument only affects outcomes through the endogenous regressor. Naturally, without random assignment one may

[16] Using genes as instruments has been subject to criticism as outlined in Cawley et al (2011) and Fang (2013), among others.

[17] For example, Wehby et al. (2011) uses two independent samples from Norway and the US to conduct IV analyses and finds weak correlations between maternal smoking and the genetic variant instrument sets.

[18] We discuss the term population stratification in further detail in the section on genome wide association studies, but the general idea is that there might be systematic differences in the frequency of risky alleles between groups, thereby leading to a form of omitted variable bias.

[19] This analysis involves making an adjustment to the asymptotic variance matrix, thereby directly affecting the standard errors. That is, a term that measures the extent to which the exogeneity assumption is erroneous constructed from prior information regarding plausible values of the impact of genetic factors on second stage outcomes is added to the variance matrix.

never be certain about the role of any specific genotype so this reinforces the need to investigate the robustness of results.

Among economists that use genetic markers as instruments, there is a major difference in how these variables are included in the first stage. Ding et al (2009) used a series of binary variables for each potential genetic polymorphism in the genes they investigated. A potential concern with datasets containing thousands of SNPs is that this may lead to the many instrument problem (Hausman et al 2012).[20] Other researchers treat the genetic information as a set of continuous variables and for each SNP include the count of the number of risk alleles. We argue that using a count variable not only makes it more challenging for researchers to interpret first stage relationships and assess if they are consistent with the scientific literature, but this also imposes a strong assumption that first stage outcomes are linear in the number of risk alleles (this assumes we know a lot more about the operation of genetic markers than the scientific literature has presently concluded). We would argue that there are benefits from allowing for non-linear relationships through using discrete indicator variable for each polymorphism of a SNP. First, one can easily test whether the linearity restriction from using a count is supported by the data. Second, this approach truly sheds more light on what features of the polymorphism are driving the estimated effect and one can then compare these results with those hypothesized in the scientific literature to gain more validity.

Studies that use genetic markers as instruments generally draw biological justification from results of published candidate gene studies, which as discussed are controversial.[21] The journal *Behavior Genetics* recently adopted stricter standards for publication of candidate gene studies (Hewitt 2012). To be considered for publication, any candidate gene study must be well powered, make corrections in statistical inference for multiple testing and any new finding must be accompanied by a replication.[22] Thus, when searching for a plausible genetic instrument by reviewing the literature, researchers should also justify their choice by considering the statistical power of the study.

The idea of using genetic information as a source of identifying variation also appears in the epidemiological literature where it is termed Mendelian randomization. Mendelian randomization was first proposed in Katan (1986) and applied with data in Davey-Smith (2003). Studies using Mendelian randomization all implicitly assume that there are no dynastic effects to invoke the term randomization. However, genes are inherited by design from one's parents who may also transmit environmental inputs and numerous behaviors across generations. In effect,

[20] To date, no research has investigated using the LASSO for variable selection in the first stage to determine which binary genetic variants in a large set of SNPs should be used as instruments.

[21] The scientific literature is populated with conflicting findings from candidate gene studies and many early studies failed to be replicated since, initially, researchers did not adjust for population stratification. Further, studies in this literature often suffer from low statistical power, coupled with potential publication bias as well as undisclosed pretesting, they could have led to too many false positives appearing in press.

[22] Chabris et al. (2013) illustrates a limitation of candidate gene studies. They replicate previously identified candidate genes using data from three independent longitudinal studies, and the results are disappointing since they found fewer significant associations than a traditional power analyses would have ex-ante predicted.

empirical economists can draw a parallel between the Mendelian randomization research design and the econometric analysis of a randomized experiment with non-compliance. Thus, under the assumption of no dynastic effects, Mendelian randomization can be viewed as an encouragement design. Since randomization (experiments) is often regarded as the gold standard in medical research, we would suggest that these studies be more accurately recast as being a Mendelian encouragement design.

A final variant on the instrumental variable strategy was introduced by Fletcher and Lehrer (2009ab, 2011) who exploit genetic inheritance within full biological siblings. Fletcher and Lehrer name a family fixed effects estimator with genetic instruments as the "genetic lottery". This genetic lottery might truly characterize Mendelian randomization, since by controlling for the family fixed effect one removes the dynastic effects (assuming they are constant) between full biological siblings. This strategy exploits variation in genetic inheritance and socio-economic outcomes between full-biological siblings and provides a means to test a key identifying assumption in a workhorse research design used in family and population economics that has been applied in almost every branch of empirical economics as well as behavioral genetics. That is, does the family fixed effects estimator sufficiently solve the underlying endogeneity problem? By using a bootstrapped Hausman test to compare a family fixed estimator to estimates with a family fixed effects IV (aka genetic lottery) one can find evidence that either refutes or is unable to reject the maintained assumptions. In each of their applications, Fletcher and Lehrer are able to reject that the family fixed effects estimators sufficiently solve the endogeneity problem in health, when estimating its effects on academic and early labor market outcomes. While labor economists have made substantial advances at estimating causal relationships, we believe that genetic information may hold more hope at identifying causal mechanisms, a topic we elaborate upon in our discussion of gene-environment interactions.

3.3 Economists Replicate Scientific Studies: Genome Wide Association Studies

Whereas research by economists using genetic markers as instruments displays a new use of these data, economists have also ventured into research methods common in medical science and by geneticists. This work led by economists who established the Social Science Genetic Association Consortium (SSGAC) involves the development of large networks of researchers and pooling of multiple datasets containing genetic information.[23] The primary goal is to conduct large scale genome wide association studies (GWAS) on a number of training datasets and to see if the results from the training datasets are replicated in other studies. Such design could yield more robust evidence (or refute) of the molecular genetic basis of outcomes of interest to economists. This approach strives to overcome many criticisms of candidate gene studies.[24]

[23] The economists who established the SSGAC are Dan Benjamin, David Cesarini and Phil Kollinger. Ambitious, the SSGAC may have been motivated by The Wellcome Trust Case Control Consortium's attempts to improve the understanding of the aetiological basis of several major causes of global disease by pooling databases collected by individual research teams. The consortium's pooled data approach has yielded important findings in the medical sciences, particularly in understanding the genetics of autism (Glesner et al., 2014) and schizophrenia (Ripke et al., 2014).

[24] The declining cost of genotyping and technological advances include the availability of canned software packages to do the analyses also likely played a large role in their growth. See McCarthy et al., (2008) among others for early examples of work in this area. Other work involves using what is termed genomic-relatedness-matrix restricted maximum likelihood (GREML) that for a sample of unrelated individual pairs

To date, the best example of research in this strand appears in Okbay et al (2016).[25] The authors first conduct a GWAS of about 300,000 people (named the discovery sample)[26] and find a total of 74 SNPs associated with educational attainment, where educational attainment is the amount of formal education completed. In aggregate, these 74 SNPs explain only 0.43% of the variation in educational attainment across individuals in the discovery sample. The economic significance of each of these 74 SNPs is also found to be quite small, since even in the strongest association reported, that is individuals with 0 instead of 2 copies of the risky allele of the genetic variant, it is shown to predict (on average) roughly 9 extra weeks of schooling. What is striking is that when the authors conduct a replication study with 110,000 individuals from the U.K. Biobank, they find that 72 of the initially identified 74 SNPs remain significantly associated with educational attainment. Thus, they are confident they have identified the molecular genetic basis for educational attainment.

Despite this confidence, Okbay et al (2016) are quite cautious in how one should interpret their findings since years of educational attainment is a complex phenomena. The results only point towards an association and one cannot separate if these identified genes are truly related to educational attainment or whether they explain the selection process that led one to complete more schooling.[27] Since there are more hypotheses of significant association than data points, one must make corrections for multiple testing, and they are careful to use an independent sample for the replication study. The authors take great care to convince the reader that the observed associations are unlikely to be spurious by both utilizing the latest quality control protocols in the medical genetics literature (Winkler et al., 2014) and carefully account for population stratification in their analysis. Specifically, the authors repeat the analysis where i) common support is imposed across samples by excluding dissimilar individuals, ii) account for high levels of principal components as additional controls to capture potentially confounding

estimates what portion of the total fraction of variance in a trait is attributable to the average effects of SNPs. That is, does genetic similarity predicts phenotypic similarity? We return to how genetic similarity is measured in the section on gene by environment interactions.

[25] Two examples of earlier papers by the SSGAC include Rietveld et. al (2013a) who combined data on 42 cohorts providing over 100,000 individuals to study which of approximately two million single nucleotide polymorphisms influences measures of educational attainment such as college completion and years of education. This research suggested three specific genetic variants. Subsequently, Rietveld et al (2014) verified the robustness of these findings using data from three new sources, as well as used exploited only genetic variation within families.

[26] The discovery sample pools numerous datasets and contains information from participants in 15 different countries.

[27] Instrumental variable estimators of the effects of years of schooling generally identify the causal effect of years of schooling only for the subsample whose behavior was influenced by the instrument. A popular example is compulsory schooling and often the resulting estimate compares individuals with 11 to 12 years of schooling. At present, with GWAS we do not know where in the decision process, the individual markers operate on individual behavior.

genetic differences across samples,[28] and iii) include family fixed effects in the analysis.[29] This paper provides a comprehensive guide on how to undertake and report results from a GWAS.

Many economists' first reaction to a GWAS is that it simply is data mining. After all, these studies are not motivated by any theory of why specific SNPs are being investigated and simply examine for an outcome of interest, whether it is associated with one or more of the (typically millions of) measured SNPs. Further, genetic researchers are generally solely interested in characterizing the variance of estimates of how much SNPs influence outcomes and point estimates are not usually the focus. While it could be of interest to discover what percentage of the variation in outcome an individual SNP can account for, this is definitely not how economists determine the relative importance of explanatory variables in outcome regressions such as wage, health, education and economic growth. In response, Rietveld et al (2013a, 2014) suggest examining polygenic scores in future research, and Papageorge and Thom (2016) present one of the first attempts to incorporate these scores in a labor economics application.

A polygenic score is constructed by adding up the individual alleles that are reliably related to this trait, where each allele is weighted by effect sizes estimated from a GWAS (Dudbridge, 2013). The underlying idea is that from the GWAS results we can give weights of relative importance to each SNP. Then, with a polygenic score, a researcher could exploit the joint predictive power of many SNPs when used as an input in an estimating equation. As an explanatory variable, these polygenic scores are constructed to explain more variation than individual SNPs and may provide clearer role on some combined genetic influence. The scores provide a means to identify individuals at high risk for certain outcomes. From an econometric perspective, this may reduce the chance of including irrelevant variables in a regression model and increase the resulting efficiency of estimates but come at a cost of placing strong functional form assumptions on the components of the score.[30] After all, the score is just a weighted linear combination which implicitly makes assumptions about relative substitutability of effects of different SNPs.

GWAS research with replication samples could be valuable to establish robust evidence of a main genetic effect. Referenced with GWAS results, studies using specific genes either in a candidate gene approach or as an instrumental variable may face significantly less opposition. We should point out, however, that evidence of main effects from these large scale GWAS requires the genetic variant to have a similar effect across all samples with respect to the same dependent variable, which likely differs on the basis of the environment and sample characteristics. It is reasonable to assume that specific genetic variants may only have significant

[28] Since data is pooled from different studies, the principal components off the gene chip (i.e. the correlation matrix of all the assayed SNPs) are measured. To control for population stratification, generally, the first four of these components are used to identify geographic ancestry within the sample.
[29] By exploiting variation within siblings, one controls for dynastic factors and any differences in genetic factors do not come from differences in sample composition. Since there is less variation and a smaller sample size, the effects are noisier relative to the discovery sample but the effect sizes are remarkably similar on average, enhancing confidence in the initial GWAS results.
[30] Debates about the relevance of polygenic scores exist outside of economics. Purcell et al., (2009) list concerns on their likely usefulness, whereas Belsky et al, (2012, 2013) are empirical examples illustrating potential benefits.

effects in particular environments or with specific types of samples, while being insignificant in all other cases. If either of the above scenarios hold, then a standard GWAS would never identify the main effect of this variant, despite the fact that there may be strong evidence of a significant heterogeneous impact of this variant with the environment.[31] Thus, there is interest in understanding the interplay between genes and the environment, a challenging area since the bar to find robust evidence of gene by environment interactions might be quite high since replication across similar contexts clearly places challenges on the date used to explore the interactive effect of genetic variants.

3.4 Gene by Environment Interactions

Recall that research in behavioral genetics (comparing twins) began by assuming the absence of gene by environment interactions, henceforth G*E. This assumption is now clearly rejected and researchers across a multitude of disciplines champion the importance of G*E effects. Among labor economists, James Heckman is perhaps best known for arguing of the importance of G*E effects in his testimony designed to convince policymakers to invest early in child development.[32]

To explore G*E effects requires rich longitudinal data with clean variation in environmental exposure to interact with genetic factors. Modelling G*E effects requires either exogenous variation in environmental factors or a clean econometric strategy that can identify unknown breakpoints in relationships between genetic factors and outcomes.[33] Rosenquist et al (2015) undertakes the latter approach using the threshold regression estimator by Hansen (1999) to estimate an augmented version of a linear age–period–cohort model, to understand the source of G*E with longitudinal data collected between 1971 and 2008 in the offspring cohort of the Framingham Heart study. Specifically, they test whether the well-documented association between the rs993609 variant of the FTO gene and body mass index (BMI) varies across birth cohorts, time period, and the lifecycle.[34] A key feature of the analysis is statistically testing for a

[31] We are grateful to Pietro Biroli for discussion that clarified why gene by environment interactions should not solely be motivated by results from GWAS.

[32] Heckman (2007) writes: "Third, the nature versus nurture distinction, although traditional, is obsolete. The modern literature on epigenetic expression and gene environment interactions teaches us that the sharp distinction between acquired skills and ability featured in the early human capital literature is not tenable (Rutter, (2006), Gluckman and Hanson (2005), Rutter et al., (2006)). Additive 'nature' and 'nurture' models, although traditional and still used in many studies of heritability and family influence, mischaracterize gene-environment interactions. Recent analyses in economics that break the 'causes' of birthweight into environmental and genetic components ignore the lessons of the recent literature. Genes and environment cannot be meaningfully parsed by traditional linear models that assign unique variances to each component. Abilities are produced, and gene expression is governed by environmental conditions (Rutter, (2006), Rutter et al., (2006). Behaviors and abilities have both a genetic and an acquired character. Measured abilities are the outcome of environmental influences, including *in utero* experiences, and also have genetic components."

[33] Fletcher and Conley (2013) argue that G*E interactions are most meaningful when they are based on exogenous environmental measures that are not themselves a function of genes. Pushing further, van IJzendoorn and Bakermans-Kranenburg (2012) advocate using randomized controlled trials to study how environmental changing interventions have differential effects as a function of genetic endowments.

[34] In other words, aggregate macro-environmental conditions are explored and one cannot identify mechanisms that operate within these three categories. It is worth pointing out that since all the data used

structural break of unknown timing across cohorts and checking the robustness of their finding by additionally controlling for family fixed effects. The selected breakpoint is based on the model that best fits the data using a grid search algorithm.[35]

Rosenquist et al (2015) find that there is a robust relationship between birth cohort and the FTO risk allele with BMI, with an observed inflection point for those born after 1942.[36] Specification tests of the unrestricted model that additionally control for gene*cohort effects and gene*age effects provide evidence that the inclusion of gene*contemporaneous period effects is statistically insignificant. Only if one were to ignore gene*cohort effects, would they find evidence that G*E effects are due to contemporaneous events for FTO and BMI. Upon reflection, this result is unsurprising since environments are highly correlated over the lifecycle for most individuals and there is limited variation in environmental conditions experienced to affect the penetrance of genetic influences.

The results also have important implications for how one interprets evidence from GWAS that pools data across samples. With GWAS researchers carefully account for population stratification when pooling data from different sources. However, the findings in Rosenquist et al (2015) raise the possibility that genetic associations may differ across birth cohorts due to variation in prevailing environmental contexts.[37] In other words, there may be a need to control for the need for environmental stratification. It remains an open question if the low replication rates of many GWAS due to ignoring environmental stratification that arises from differences in the period of time study subjects were born in and the historical moment researchers conduct their investigations.

To date, the majority of work by social scientists evaluating G*E effects does not explicitly consider the endogeneity of the environmental variables that were selected by the individual. Perhaps the best example of research in this stream is Biroli (2015) who situates his analysis within an economic framework.[38] Biroli (2015) integrates genetic factors inside the canonical model of health production due to Grossman (1972), allowing genetic variants to both potentially differentially affect the health production function and preferences related to the incentives related to health investment faced by individuals. Using data from both the

---

in this study was collected in a single small geographic area, any potential biases due to sorting across regions based on environmental conditions due to unobservables are reduced.

[35] Best fitting refers to explained variation. That is, the breakpoint identifies the point in time where the difference between birth cohorts in how genes influence obesity that would explains the most variation in the data. The econometric strategy does not identify the point where the relationship is most different.

[36] Consistent with Rosenquist et al. (2015), Biroli (2015) finds that the estimated interaction between the FTO genotype and caloric intake is stronger for individuals born in later cohorts.

[37] This criticism is not viewed favorably among geneticists and is more natural to economists who understand that GWAS just report associations, and not causal or structural parameters.

[38] As with candidate gene studies, concerns of low statistical power due to a combination of potential pre-testing and publication bias are likely valid in what we shall term “candidate gene*environment interactions”. Two examples of such studies: 1. Caspi et al (2002) finds that the effects of self-reported childhood maltreatment on adolescent antisocial behavior varied based on one’s MAOA gene. 2. Shanahan et al. (2008) discovers a significant interaction between a variant of the DRD2 dopamine receptor gene with factors such as having a parent that belongs to the PTA and how often parents discuss school related issues with the student.

Framingham Heart study and Avon Longitudinal Study of Parents and Children, Biroli finds evidence that genetic factors do change both the production function of BMI and the level of health investment. While this work extends our analysis of a workhorse model in health economics, the empirical analysis requires one to assume that caloric intake is exogenous and not a behavioral choice, otherwise biased coefficients may result.

Studies that have tried to exploit genetic variation within families have the potential to provide more compelling evidence of candidate G*E effects. Similar to Fletcher and Lehrer (2011), the idea is to exploit within family differences in genetic code to remove biases from dynastic effects. For example, Thompson (2014) exploits within-family variation in genetic inheritance, to see if there are differential responses of household income on child education outcomes by variants of the MAOA genes. The results indicate that the gradient is steeper for those with rarer variants.[39] Conley and Rauscher (2013) advise caution in this line of research. When they investigate how genetic traits may moderate the relationship between birthweight and several outcomes including high school GPA that exploits within twin- pair birthweight differences, the sole statistically significant G*E effect discovered has a sign that is the opposite of what had been suggested by prior scientific research.

Economists are well-aware of the benefits of comparative advantage. Thus, one can interpret the set of guidance provided in Conley and Rauscher (2013) as indicating potential benefits from interdisciplinary collaborations. Most economists are not trained to assess whether the estimated effects of specific SNPs are plausible in sign and magnitude.[40]

Dealing with potential 'environmental stratification' is explicitly yet indirectly considered in what is known as genome-wide complex trait analysis (GCTA), a variant on the behavioral genetic approach to measure heritability between genetically dissimilar individuals.[41] Yang et al., (2011) suggests that genetic similarity between two individuals is essentially estimated as a weighted correlation of their genotypes on the included SNPs and the goal is to restrict the analysis to unrelated individuals. This restriction is motivated by the assumption that individuals who are more genetically related share a more similar environment than unrelated individuals. Using a restricted maximum likelihood estimator (commonly referred to as REML in the literature), one can obtain estimates of heritability without resorting to twins data.[42]

---

[39] Thompson (2014) also point out that parents may make "compensating" investments in which more resources are allocated to the less able sibling to promote equality. Thus, one cannot rule out with the data that MAOA variants is correlated with the environmental conditions children receive from their parents after conception. Future research is need to see if a child's MAOA status induces differential treatment from their parents who invest in their children's human capital and if so, to what signals of MAOA status do parents respond, given that they are unlikely to have genotyped their children.

[40] There are examples of successful interdisciplinary collaborations reviewed in this chapter, including the multiple papers produced by the SSGAC that was lauded in an editorial *Nature* (Hayden, 2013), Rosenquist et al (2015), among others.

[41] Benjamin et al. (2012), use this approach in their analysis to explain heritability of economic preferences.

[42] As an example, Rietveld et al (2013b) point out that while twins studies suggest that genetic factors may account for as much as 30–40% of the variance in subjective well-being measures, additive effects of genetic polymorphisms that are common in the population can only explain 5-10% of the variation in

However, this identifying assumption appears strong to individuals trained as labor or urban economists since it would suggest that genetic markers must explain much selection of one's environment. To test this assumption, one may develop tests similar to understanding whether selection on observables leads to balance. We discuss further econometric and computational directions for economists in the next section.[43]

4: Potential Future Areas for Economists to Analyze Genetic Data

We suggest that there are, at a minimum, three main directions that economists can contribute to the literature using genetic data. First, economists can provide an understanding of genetic influences in a logically consistent framework. Economists should not fall victim to thinking that just because one's genetic code is fixed at conception, genetic-expression is also fixed. Labor economists treat fixed characteristics such as gender and race as having time-varying effects in empirical analysis. Research is needed to understand whether individual SNPs or polygenic scores are time varying and to what extent they are truly capturing a portion of what economists in longitudinal analysis refer to as permanent individual specific unobserved heterogeneity. Clarifying what is meant by a genetic disposition is an area where economists can contribute strongly.

Moreover, similar to Manski (2013) arguing in his view of the incredible certitude taken by many in the public policy community, economists should help other research communities to become much more comfortable with embracing uncertainty in how genetic effects operate. To an extent, the work of the SSGAC evaluating genome associations is striving to reduce the degrees of certitude represented in any given candidate gene study. Future research if well-powered can also be used to understand why across the databanks collected by the SSGAC, divergent main effects are observed. That is, do differences in specific life events of the participants across studies explain these divergence in the main effects? After all, different experiences may alter the magnitude of genetic expression on a given trait, or it may in fact instigate different genetic processes.

Building on the above point that suggests potential gene environment interplay is something that should be emphasized more strongly in the genetics literature, a second area where economists can contribute is by developing tools and research designs to shed new light on the pathways through which genetic factors influence socioeconomic outcomes. Lehrer (2015) suggests that researchers should consider working with more aggregated environmental factors and perhaps exploit regional environmental changes. Indeed, there is a large history in empirical microeconomics of exploring differences in environmental conditions or policies across regions as natural experiments. Thus, exploring genetic heterogeneity in the estimated effects with this research design seems to hold promise. An early example of such a study is Okbay et al (2016a) who compare cohorts prior to and post a suite of schooling reforms that,

---

these measures. While subjective well-being measures are not accurately measured, accounting for measurement error only increases the amount of explained variation from additive genetic effects to 12-18%.

[43] More recent developments for sequencing and linkage analysis that have been introduced in the genetics literature include Ott et al (2015) and Pabinger et al (2014); but to the best of our knowledge have yet to be used by economists.

most importantly, extended mandatory schooling from seven to nine years. The authors find that the association between educational attainment and the polygenic score constructed from their GWAS is roughly half as large among Swedish individuals in later cohort, suggestive that the Swedish reforms reduced the effects of genetic variants in generating differences in educational attainment.

With genetic data, it may be possible to use biological mechanisms to shed light on why treatment effect heterogeneity is observed. Researchers in empirical microeconomics already have sets of tools to explore whether interventions have different effects for subgroups defined on the basis of more aggregated predetermined characteristics such as gender and race.[44] It is also necessary to move beyond linear models in order to study G*E effects. Conti and Heckman (2010) provide a more general framework to operationalize and interpret gene-environment interactions.

As a whole, there is tremendous scope in this stream for both empiricists and econometricians to collaborate and develop methodological tools for G*E analyses. Researchers currently use G*E to both describe situations where the effect of exposure to an environmental factor on a behavior is conditional upon a person's genotype, as well as situations when the genotype's effect is moderated by some environmental effect. While statistically separating these pathways is needed, policy audiences do need to understand what is being identified. Lehrer (2015) suggests that researchers use *G*E responses* to refer to situations where the effect of exposure to an environmental factor on a behavior is conditional upon a person's genotype and *G*E modifications* to refer to differential genetic reactions to environment Personalized medicine and policies that would target by genotype may be interested in *G*E modifications,* whereas *G*E responses* may be more interesting for researchers trying to understand the heterogeneity in environmental effects on outcomes across population. In short, improving methodological tools can lead to more credible evidence from rigorous G*E studies, which could subsequently reshape theories on various health and socioeconomic outcomes.

Third, economists have the advantage to investigate the behavioral restrictions implicitly imposed by empirical methods used to both elucidate genetic associations and construct polygenic scores. Consider how Todd and Wolpin (2003) influenced researchers in the economics of education by highlighting the behavioral restrictions on an underlying model of human capital development that were implicitly made by researchers when estimating various equations that proxied for education production functions. Analogously, the socioeconomic outcomes being investigated in both candidate gene and GWAS studies are likely determined by complex processes that involve behavioral decisions. As another example that was discussed in the section considering genes as instruments, should risk alleles enter these estimating equations as a count assuming a linear effect or as a series of indicator variables? The consequences of using imputed versus actual SNPs also requires further evaluation. Further, it is worth stressing that much of the existing analysis in the scientific literature uses canned software that itself imposes additional assumptions on the underlying process generating the outcomes. The maintained assumptions in the methodology undertaken need to be made clear to the general research community.

[44] For example, Lee and Shaikh (2014) and Lehrer et al. (2016) provide a set of methodological tools to analyze heterogeneity in causal effects that can additionally incorporate corrections for multiple testing.

In GWAS, a number of potential methodological questions are worth considering. For example, should these equations be estimated for different health outcomes independently, or as a system of equations framework allowing for correlations in the residuals? Similar to evidence of the importance of comorbidity in Ding et al (2009), Boardman et al. (2015)'s GWAS investigation of the molecular basis of education and depression / self-rated health, points out that one may wish to disentangle whether a given genetic marker has an independent influence on outcomes or mediates the effect of these correlated outcomes on one another. Could we use LASSO estimator for GWAS studies as a means to shrink the variable set in place of REML estimators? Other directions include developing an optimal way to make corrections for multiple testing in settings where there are potentially more covariates (SNPs) than observations.

Labor economists have done much work developing methods to estimate both cross-sectional and panel data models with repeated cross sections. To an extent, work on GWAS is pooling many samples that did not choose individuals in the study via random sampling. In addition, between the different datasets, members of the population have an unequal probability of being observed. These sampling issues are normally not considered other than creating a form of balance the samples by adding controls for population stratification. Issues related to nonparameteric identification of population parameters in this setting when many choice based samples are combined (and as noted earlier, Rosenquist et al (2015) point out that differences across environments in these studies is not considered), seem important to correctly interpret the resulting estimates. Further, given the combination of these non-random samples, there may be methods to conduct efficient estimation from a combination of biased samples. While much work has been done across disciplines on topics related to combining non-random datasets, the mixing of types of datasets used in GWAS which range from case-control studies to random samples likely requires researchers to find better ways to combine samples in order to improve both the estimation and interpretation of results from GWAS.

Polygenic scores are beginning to generate much interest for assessing the explanatory power of an ensemble of genetic markers. Turning to the construction of polygenic scores,[45] should these be anchored in a metric that has economic significance such as earnings? How should researchers account for estimation error in these scores when including these measures as explanatory variables? Is there a partial identification approach to calculate polygenic scores? Labor economists have worked as applied econometricians for decades and as one becomes more familiar with not just genetic data but with the methods used to measure and analyze this data, we believe there is great scope to develop refinements that will have impacts not solely on the economics literature but in other disciplines that analyze these data.

Along these lines, to understand one of the main challenges with using polygenic scores, consider the ingredients to their construction faces a common problem empirical researchers face in practice when trying to decide what conditioning variables to include when estimating either a structural parameter or causal effects. After all, with GWAS, there is a related problem of selecting which SNPs from a potentially large set to include since ignoring potentially relevant variables that do not satisfy a strong sparsity condition would result in omitted variables bias.

[45] At present, most researchers who construct these scores rely on canned software routines such as PRSice for convenience and do not discuss the statistical and behavioral restrictions embedded.

This bias has been shown in Leeb and Pötscher (2008), Belloni et al (2014, 2016), among others, to exist with common econometric estimator and machine learning strategies who additionally use simulations to demonstrate that the sampling distribution of whichever estimator is not centered around the true parameter. Since GWAS is focused strictly on prediction and not causal effects, it is unclear how to adapt double-selection estimators to this setting. Since the ingredients to the construction of the polygenic score are biased and unlikely to be robust to even minor deviations from conditions implying perfect model selection in the GWAS itself, it is unclear whether this variable can accurately capture individual genetic predisposition. Future work that uses these scores in empirical applications may be better off creating binary indicators for having a very high or very low polygenic score, rather than assuming perfect model selection and treat the polygenic score as a continuous variable.

Finally, as the scientific literature is also now moving beyond only considering main genetic effects, it is worth pointing out that gene-gene interactions almost certainly do exist.[46] Indeed, both Ding et al. (2006, 2009) and Fletcher and Lehrer (2009b, 2011) consider such two-way interactions in their instrument set, but there is not much information even in the behavioral genetics literature on how and why these interactions operate. In other words, understanding the genetic architecture of a particular trait is one of the main goals and this challenge mirrors the steps required when labor economists create empirical models to understand the underlying data generation process. With newer and richer data, future research will be able to additionally explore the interactive effects between genes themselves as well as with environmental interactions and genetic networks. With the likely continued increasing focus of labor economists at understanding the origins of economic inequality, we believe that molecular genetic data may help shed new light on understanding the sources of unobserved heterogeneity.

5. Conclusions

Many labor economists including Claudia Goldin (1999) have described their research strategy as first finding a topic that one is passionate about, and then being the best detective one could be. Indeed, labor economists have for painstakingly long analyzed data, not solely to help reveal trends and patterns, but also to shed new light on drivers of human behavior. There is no question that genetic factors do play a role in nearly every socioeconomic outcome of interest to empirical economists and only recently have we begun to develop affordable and reliable technologies to measure this individual level of variation.

Over the past decade, a growing number of economists have begun to incorporate genetic markers in their empirical analyses. This area is quickly maturing and it is likely that many of the low hanging applications of genetic data have already been undertaken. At this stage, and similar to trends within labor economics, we are witnessing a shift towards researchers using much larger datasets to asses genetic associations as well as researchers developing new econometric tools to understand heterogeneity in genetic effects. These trends parallel those within labor

[46] See Lazopoulou et al. (2014) Huang et al. (2011), among others for evidence of significant gene-gene interactions in obesity. In behavioral genetics, additive genetic effects are associated with a narrow sense of heritability and broad-sense heritability refers to the proportion of trait variation that can be attributed to all types of genetic effects, including dominance, epistatic interaction, and additive effects.

economics where a growing number of studies are relying on the use of rich administrative databases to draw credible evidence as well as the development of econometric tools to shed light on treatment effect heterogeneity.

There is likely much more that can be done particularly along entering the black box of individual specific unobserved heterogeneity and exploring gene environment interactions. Evidence from these studies can in turn be utilized to help refine theories and potentially shape policies. While the idea of developing a separate field within economics called genoeconomics is clearly appealing to those in the area, we believe that there is more potential from incorporating genetic data within existing fields such as labor economics. For example, many labor economists when studying educational attainment must carefully deal with modelling individual choices of whether to continuing to study that is based on both the expected return, risk and costs. Yet, when researchers suggest genetic factors influence educational attainment, which aspect of the selection process are the markers related to?

We should caution that there are high start-up costs for researchers trained as economists in understanding genetics research, in part since research in other disciplines use different and complex jargons and are generally much less explicit about the behavioral assumptions.[47] That said, there are likely high returns for economists to develop richer and empirically tractable models to investigate the role of genetic factors that can challenge the maintained behavioral assumptions. In summary, while we are ultimately bullish on the future of genetic markers in economics, we believe such future may become more fruitful if us economists are more critical and disciplined in our embrace of genetic data.

[47] Similar to how the training of econometricians and microeconomic theorists rely on developing stronger backgrounds in specific branches of mathematics and statistics, economists will need to become much more familiar with the genetics literature. This is not unique to the field. Currently many economists are now learning machine learning tools to analyze large datasets (Athey, 2015), whereas many behavioral economists need to keep track of developments in the psychology and neuroscience literatures.

Bibliography

Ashraf, Q., and O. Galor. 2013. The "Out of Africa" hypothesis, human genetic diversity, and comparative economic development. *American Economic Review* 103(1): 1-46.

Ashraf, Q., and O. Galor. 2012. Response to comments made in a letter by d'Alpoim Guedes et al. on "The Out of Africa hypothesis, human genetic diversity and comparative development."http://www.econ.brown.edu/fac/Oded_Galor/Ashraf-Galor%20Response.pdf.

Athey, S. (2015), "Machine Learning and Causal Inference for Policy Evaluation", forthcoming in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

Baicker K, Cutler D, Song Z (2010), Workplace wellness programs can generate savings. *Health Affairs* 29(1), 1–8.

Behrman, J. R. (2015), Twin Studies in Economics in the Oxford Handbook of Economics and Human Biology (J. Komlos and I. Rashad eds.) Oxford University Press.

Belloni A, Chernozhukov V, Hansen C. (2014). Inference on treatment effects after selection amongst high-dimensional controls. *Review of. Economic Studies* 81(3), 608–650.

Belloni A, Chernozhukov V, Fernandez-Val, I. and Hansen C. (2014). Program Evaluation and Causal Inference with High-Dimensional Data forthcoming in *Econometrica.*

Belsky, D. W., Moffitt, T. E., Baker, T. B., Biddle, A. K., Evans, J. P., Harrington, H., … Caspi, A. (2013). Polygenic risk and the developmental progression to heavy, persistent smoking and nicotine dependence: Evidence from a 4-decade longitudinal study. *JAMA Psychiatry*, *70*, 534–542.

Belsky, D. W., Moffitt, T. E., Houts, R., Bennett, G. G., Biddle, A. K., Blumenthal, J. A., … Caspi, A. (2012). Polygenic risk, rapid childhood growth, and the development of obesity: Evidence from a 4-decade longitudinal study. *Archives of Pediatrics & Adolescent Medicine*, *166*, 515–521

Benjamin, D. J., D. Cesarini, C. F. Chabris, E. L. Glaeser, D. I. Laibson, V. Guðnason, T. B. Harris, L. J. Launer, S. Purcell, A. V. Smith, M. Johannesson, Patrik K.E. Magnusson, Jonathan P. Beauchamp, N. A. Christakis, C. S. Atwood, B. Hebert, J. Freese, R. M. Hauser, T. S. Hauser, A. Grankvist, C. M. Hultman, and P. Lichtenstein (2012), "The Promises and Pitfalls of Genoeconomics." *Annual Review of Economics* 4:627–662.

Benjamin, D. J., D. Cesarini, M. J.H.M. van der Loos, C. T. Dawes, P. D. Koellinger, P. K.E. Magnusson, C. F. Chabris, D. Conley, D. I. Laibson, M. Johannesson, and P. M. Visscher (2012), The Genetic Architecture of Economic and Political Preferences, *Proceedings of the National Academy of Sciences*, 109(21):8026–8031

Benjamin, D. J., C. F. Chabris, E. L. Glaeser, V. Gudnason, T. B. Harris, D. I. Laibson, L. Launer, and S. (2007), "Genoeconomics," In Weinstein, Maxine, James W. Vaupel, and Kenneth W. Wachter (eds.), *Biosocial Surveys.* Committee on Population, Division of Behavioral and Social Sciences and Education. Washington, D.C.: The National Academies Press.

Biroli, P. (2015), Genetic and Economic Interaction in the Formation of Human Capital: The Case of Obesity, *mimeo,* University of Chicago.

Boardman, J. D., Domingue, B. W., & Daw, J. (2015). What can genes tell us about the relationship between education and health? *Social Science & Medicine*, *127*, 171–180.

Boardman, J. D., Daw, J. , & Freese, J. (2013). Defining the environment in gene–environment research: Lessons from social epidemiology. *American Journal of Public Health*, 103(S1), S64–S72.

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., … Fox, J. (2011). OpenMx: An open source extended structural equation modeling framework. *Psychometrika*, *76*, 306–317.

Carpenter J, Garcia J, Lum J (2011), Dopamine receptor genes predict risk preferences, time preferences, and related economic choices, *Journal of Risk and Uncertainty* 42:233-261.

Caspi, A., McClay, J., Moffitt, T. E., Mill, J., Martin, J., Craig, I. W., Taylor, A., and Poulton, R. (2002), Role of genotype in the cycle of violence in maltreated children, *Science* 297(5582):851–854.

Cawley, J., E. Han and E. Norton (2011), The validity of genes related to neurotransmitters as instrumental variables, *Health Economics*, 20(3):884–888.

Cesarini D., Johannesson M, Magnusson P, Wallace B. (2012) The behavioral genetics of behavioral anomalies, *Management Science* 58(1):21-34.

Cesarini D., Johannesson M., Lichtenstein P., Sandewall O., Wallace B. (2010), Genetic variation in financial decision-making. *Journal of Finance* 65:1725-1754.

Cesarini, D., C.T. Dawes, M. Johannesson, P. Lichtenstein, and B. Wallace (2009), Genetic variation in preferences for giving and risk-taking, *Quarterly Journal of Economics* 124:809–842.

Cesarini D., Dawes C.T., Fowler J., Johannesson M., Lichtenstein P., Wallace B. (2008), Heritability of cooperative behavior in the trust game, *Proceedings of the National Academy of Sciences* 105:3271-3276.

Chabris, C.F., J.J. Lee, D.J. Benjamin, J.P. Beauchamp, E.L. Glaeser, G. Borst, S. Pinker, and D.I. Laibson (2013), Why Is It Hard to Find Genes that are Associated with Social Science Traits? Theoretical and Empirical Considerations, *American Journal of Public Health* 103(S1):S152–S166.

Chabris, C. F. , Hebert, B. M. , Benjamin, D. J. , Beauchamp, J. , Cesarini, D. , van der Loos, M. , … Laibson, D. (2012). Most reported genetic associations with general intelligence are probably false positives. *Psychological Science*, 23(11): 1314–1323.

Conley D. 2009. "The Promise and Challenges of Incorporating Genetic Data into Longitudinal Social Science Surveys and Research," *Biodemography and Social Biology*, 55(2): 238- 251.

Conley, Dalton and Emily Rauscher. 2013. "Genetic Interactions with Prenatal Social Environment: Effects on Academic and Behavioral Outcomes." *Journal of Health and Social Behavior* 54(1):109-27.

Conley, T. G., C. B. Hansen, and P.E. Rossi (2012). Plausibly exogenous. *Review of Economics and Statistics*, 94:260–272

Conti, G., and Heckman, J. J. (2010), Understanding the early origins of the education– health gradient: A framework that can also be applied to analyze gene– environment interactions, *Perspectives on Psychological Science* 5:585–605.

Davey Smith, G. (2003). "Mendelian randomization": can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, 32(1):1–22.

d'Alpoim Guedes, J., Bestor TC, Carrasco D, Flad R, Fosse E, et al. (2013) Is poverty in our genes? *Current Anthropology* 54: 71–79.

d'Alpoim Guedes, J., D. Reich, M. Herzfeld, N. Patterson, T. Bestor, D. Lieberman, J. Comaroff, et al. 2012. Response to Ashraf and Galor "The Out of Africa hypothesis, human genetic diversity and comparative economic development."http://ssrn.com/abstract=2155060.

DeFries, J. C. and Fulker, D. W. (1985). Multiple regression analysis of twin data. *Behavior Genetics*, *15*(5), 467–473.

DeNeve, J-E and J. Fowler (2015) Credit Card Borrowing and the Monoamine Oxidase A (MAOA) Gene, forthcoming in the *Journal of Economic Behavior and Organization*.

Ding W., Lehrer S.F., Rosenquist J.N., Audrain-McGovern J. (2009), The impact of poor health on academic performance: New evidence using genetic markers, *Journal of Health Economics* 28(3):578–97.

Ding, W., Lehrer, S.F., Rosenquist, N.J., and Audrain-McGovern, J. (2006), The Impact of Poor Health on Education: New Evidence Using Genetic Markers, *NBER Working paper 12304.*

Dreber, A., D.G. Rand, N. Wernerfelt, J.R. Garcia, M.G. Vilar, J. K. Lum and R. J. Zeckhauser (2011), Dopamine and Risk Choices in Different Domains: Findings Among Serious Tournament Bridge Players, *Journal of Risk and Uncertainty* 43:19-38.

Dreber, A. C. L. Apicella, D.T.A. Eisenberg, J. R. Garcia, R. S. Zamore, J. K. Lum, and B. Campbell (2009), The 7R polymorphism in the dopamine receptor D4 gene (DRD4) is associated with financial risk taking in men, *Evolution and Human Behavior* 30(2):85–92.

Dudbridge, F. (2013), Power and predictive accuracy of polygenic risk scores. *PLoS Genetics* 9, e1003348.

Fletcher, J. M., & Conley, D. (2013). The challenge of causal inference in gene-environment interaction research: Leveraging research designs from the social sciences. *American Journal of Public Health*, *103*(Suppl. 1), S42–S45.

Fletcher, J. M and Lehrer, S.F. (2009a), Using Genetic Lotteries within Families to Examine the Causal Impact of Poor Health on Academic Achievement. *National Bureau of Economic Research Working Paper Series No. 15148.*

Fletcher, J.M. and Lehrer, S. F. (2009b), The Effects of Adolescent Health on Educational Outcomes: Causal Evidence Using Genetic Lotteries between Siblings, *Forum for Health Economics & Policy* 12(2):Article 8.

Fletcher, J. M. and Lehrer, S.F. (2011), Genetic lotteries within families, *Journal of Health Economics* 30:647– 659.

Gee (2014) "All the Time in the World: An Examination of Time Preferences using Monetary Delay Discount Rates" *MA Research paper,* Queen's University.

Glessner, Joseph T, John J. Connolly and Hakon Hakonarson (2014), "Genome-Wide Association Studies of Autism", Current Behavioral Neuroscience Reports 1(4), 234-241.

Gluckman P.D., Hanson M. (2005), *The Fetal Matrix: Evolution, Development, and Disease* (Cambridge University Press, Cambridge, UK).

Goldin C. (1999) "The Economist as Detective." In M. Szenberg, ed., Passion and Craft: Economists at Work.

Greiner, James, and Donald Rubin. 2011. "Causal Effects of Perceived Immutable Characteristics." *The Review of Economics and Statistics* 93:775-785.

Grossman, M. (1972), On the concept of health capital and the demand for health, *Journal of Political Economy*, 80(2):223–255.

Hansen, B.E. (1999),Threshold effects in non-dynamic panels: estimation, testing, and inference, *Journal of Econometrics*, 93, 345–368.

Hausman , J. A. , Newey , W. K. , Woutersen , T. , Chao , J. C. , Swanson , N. R. (2012), Instrumental variables estimation with heteroskedasticity and many instruments, *Quantitative Economics* 3(2):211 – 255.

Hayden, E. C. (2013), Dangerous Work, *Nature* 502, 5-6 (03 October 2013).

Heckman, J. J. (2007), The economics, technology and neuroscience of human capability formation, *Proceedings of the National Academy of Sciences*, 104, 13250–13255.

Herrnstein, R. J. and Murray, C., (1994), *The Bell Curve.* New York: The Free Press.

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

Huang W, Y. Sun, and J. Sun (2011), Combined effects of FTO rs9939609 and MC4R rs17782313 on obesity and BMI in Chinese Han populations, *Endocrine 3*9(1), 69–74.

Jencks Christopher (1980), Heredity, Environment, and Public Policy Reconsidered. *American Sociological Review*. 45, 723–736

Jensen, Arthur R. (1967), "Estimating the Limits of Hereditability of Traits by Comparison of Monozygotic and Dizygotic Twins", *Proceedings of the National Academy of Science*, 58:149-156

Katan, M. B. (1986), Apolipoprotein E isoforms, serum cholesterol and cancer, *Lancet*, 327:507-508.

Knafo, A., Israel, S., Darvasi, A., Bachner-Melman, R., Uzefovsky, F., Cohen, L., Feldman, E., Lerer, E., Laiba, E., Raz, Y. (2008), Individual differences in allocation of funds in the dictator game associated with length of the arginine vasopressin 1 a receptor RS 3 promoter region and correlation between RS 3 length and hippocampal mRNA, *Genes, Brain and Behavior* 7(3):266–275.

Kohler, H.-P., Behrman, J. R., and Schnittker, J. (2011), Social Science Methods for Twins Data: Integrating Causality, Endowments, and Heritability, *Biodemography and Social Biology*, 57(1):88–141.

Lazopoulou N, Gkioka E, Ntalla I, Pervanidou P, Magiakou AM, Roma-Giannikou E, Chrousos GP, Papassotiriou I, Dedoussis G, Kanaka-Gantenbein C. (2014) The combined effect of MC4R and FTO risk alleles on childhood obesity in Greece *Hormones* (Athens). 2014 Nov 5.

Lee, S. and A. Shaikh (2014), Multiple Testing and Heterogeneous Treatment Effects: Re-Evaluating the Effect of PROGRESA on School Enrollment, *Journal of Applied Econometrics*, 29:612–626.

Leeb H, Pötscher B. M. (2008). Sparse estimators and the oracle property, or the return of Hodges' estimator. *Journal of Econometrics*, 142: 201–211.
Lehrer, S. F. (2015), Biomarkers as Inputs in the Oxford Handbook of Economics and Human Biology (J. Komlos and I. Rashad eds.) Oxford University Press.

Lehrer, S. F., V. R. Pohl and K. Song (2016), Targeting Policies: Multiple Testing and Distributional Treatment Effects, *mimeo*, Queen's University.

Manski, C. (2013), *Public Policy in an Uncertain World*, Harvard. University Press.

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, , Ioannidis JP, Hirschhorn JN.. (2008), Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Review Genetics* 9(5), 356–369.

Mertins, V., Schote, A.B., Hoffeld, W., Griessmair, M., Meyer, J., (2011), Genetic susceptibility for individual cooperation preferences: The role of monoamine oxidase a gene (MAOA) in the voluntary provision of public goods, *PLoS One*

Ott J, Wang J, and Leal SM (2015) Genetic linkage analysis in the age of whole-genome sequencing. *Nature Review Genetics* 16: 275-284

Okbay, A. J.P. Beauchamp, M.A. Fontana, J.J. Lee, T.H. Pers, C.A. Rietveld, P. Turley,..., P.M. Visscher, T. Esko, P.D. Koellinger, D. Cesarini, D.J. Benjamin, (2016a), Genome-wide association study identifies 74 loci associated with educational attainment, *Nature*, 533, 539–542, (26 May 2016)

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M.R., Zschocke, J., and Trajanoski, Z. (2014) A survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinformatics* 15, 256-278.

Papageorge, N. W. and K. Thom (2016), Genes, Education, and Labor Market Outcomes: Evidence from the Health and Retirement Study, *mimeo*, John Hopkins university.

Purcell, S. (2002). Variance Components Models for Gene-Environment Interaction in Twin Analysis. *Twin Research and Human Genetics*, 5(06):554–571.

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

Rabe-Hesketh, S., Skrondal, A., & Gjessing, H. K. (2008). Biometrical modeling of twin and family data using standard mixed model software. *Biometrics*, *64*, 280–288.

Rietveld, C. A., Tõnu Esko, Gail Davies, Tune H. Pers, Patrick A. Turley, Beben Benyamin, C. F. Chabris, Valur Emilsson, Andrew D. Johnson, James J. Lee, Christiaan de Leeuw, R. E. Marioni, Sarah E. Medland, Michael B. Miller, Olga Rostapshova, Sven J. Van der Lee, A. A.E. Vinkhuyzen, Najaf Amin, Dalton Conley, J. Derringer, C. M. van Duijn, R. Fehrmann, Lude Franke, Edward L. Glaeser, Narelle K. Hansell, C. Hayward, William G. Iacono, Carla A. Ibrahim-Verbaas, Vincent Jaddoe, Juha Karjalainen, David Laibson, Paul Lichtenstein, David C. Liewald, Patrik K.E. Magnusson, Nicholas G. Martin, Matt McGue, George McMahon, Nancy L. Pedersen, Steven Pinker, David J. Porteous, Danielle Posthuma, Fernando Rivadeneira, Blair

H. Smith, John M. Starr, Henning Tiemeier, Nicholas J. Timpson, Maciej Trzaskowski, André G. Uitterlinden, Frank C. Verhulst, Mary E. Ward, Margaret J. Wright, George Davey Smith, Ian J. Deary, Magnus Johannesson, Robert Plomin, Peter M. Visscher, Daniel J. Benjamin, David Cesarini and Philipp D. Koellinger (2014). "Common Genetic Variants Associated with Cognitive Performance Identified Using Proxy-Phenotype Method." *Proceedings of the National Academy of Sciences*, 111(38), 13790-4.

Rietveld, C. A., S. E. Medland, J. Derringer, J. Yang, T. Esko, N. W. Martin, H.-J. Westra, K. Shakhbazov, …, D. Conley, G. Davey-Smith, L. Franke, P. J. F. Groenen, A. Hofman, M. Johannesson, S. L.R. Kardia, R. F. Krueger, D. Laibson, N. G. Martin, M. N. Meyer, D. Posthuma, A. R. Thurik, N. J. Timpson, A. G. Uitterlinden, C. M. van Duijn, P. M. Visscher, D. J. Benjamin, D. Cesarini, and P. D. Koellinger (2013). "GWAS of 126,559 individuals identifies genetic variants associated with educational attainment." *Science*, 340(6139), 1467–71.

Rietveld, C. A., D. Cesarini, D. J. Benjamin, P. D. Koellinger, J-E. De Neve, H. Tiemeier, M. Johannesson, Patrik K.E. Magnusson, N. L. Pedersen, R. F. Krueger, and M. Bartels (2013b). "Molecular Genetics and Subjective Well-Being." *Proceedings of the National Academy of Sciences*, 110(24), 9692–9697.

Ripke, S and members of the Schizophrenia Working Group of the Psychiatric Genomics Consortium et al. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 511(7510): 421-427.

Rosenquist, J. N., Lehrer, S. F., Malley, A. J. O., Zaslavsky, A. M., Smoller, J. W., and Christakis, N. A. (2015). Cohort of birth modifies the association between FTO genotype and BMI. *Proceedings of the National Academy of Sciences* 112(2), 354-359.

Rutter M (2006*) Genes and Behavior: Nature–Nurture Interplay Explained* (Blackwell, Oxford).

Rutter M, T. E. Moffitt and A. Caspi (2006), Gene-environment interplay and psychopathology: Multiple varieties but real effects, *Journal of Child Psychology and Psychiatry Journal of Child Psychology and Psychiatry* 47(3-4), 226–261.

Rutter, M., and J. Silberg *(*2002*).* Gene–environment interplay in relation to emotional and behavioral disturbance. *Annual Review of Psychology*, 53, 463–490.

Shanahan, M. J., S. Vaisey, L. D. Erickson, and A. Smolen (2008), Environmental contingencies and genetic propensities: social capital, educational continuation, and dopamine receptor gene DRD2, *The American Journal of Semiotics* 114 (86):S260-86.

Taubman, Paul (1976), "The Determinants of Earnings: Genetics, Family, and Other Environments: A Study of White Male Twins," *American Economic Review* 66: 858- 870.

Thompson, O. (2014) Economic Background and Educational Attainment The Rhole of Gene-Environment Interactions, *Journal of Human Resources* 49(2), 263-294

Todd, P. E. and K. I. Wolpin (2003), "On the Specification and Estimation of the Production Function for Cognitive Achievement," *Economic Journal* 113(1), 3-33.

Van IJzendoorn, M.H., & Bakermans-Kranenburg, M.J., (2012). A sniff of trust: meta-analysis of the effects of intranasal oxytocin administration on face recognition, trust to in-group, and trust to out-group. *Psychoneuroendocrinology*, 37, 438- 443.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Gabor Miklos, G.L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Yuan Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S.C., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Lai Cheng, M., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F.,May, D., McCawley, S., McIntosh, T.,McMullen, I., Moy, M.,Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N., Tse, S., Vech, C.,Wang, G.,Wetter, J., Williams, S., Williams, M.,Windsor, S., Winn-Deen, E.,Wolfe, K., Zaveri, J., Zaveri, K., Abril, J.F., Guigó, R., Campbell, M.J., Sjolander, K.V., Karlak, Kejariwal, B., Mi, A.H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R.,Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y., Coyne, M., Dahlke, C., Deslattes Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros,W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E.,Wang, M.,Wen, M.,Wu, D.,Wu, M., Xia, A., Zandieh, A., Zhu, X., 2001. The sequence of the human genome. *Science* 291, 1304–1351.

Wallace, B., D. Cesarini, P. Lichtenstein, and M. Johannesson, 2007, Heritability of ultimatum game responder behavior, *Proceedings of the National Academy of Sciences* 104, 15631-15634.

Wehby, G., J. M. Fletcher, S. F. Lehrer, L. M. Moreno J. C. Murray, A. Wilcox and R. T. Lie (2011), A Genetic Instrumental Variables Analysis of the Effects of Prenatal Smoking on Birth Weight: Evidence from Two Samples, *Biodemography and Social Biology* 57(1): 3–32

Winkler, T. W., F. R Day, D. C. Croteau-Chonka, A. R Wood, A. E. Locke, R. Mägi, T. Ferreira, T. Fall, M. Graff, A. E. Justice, J. Luan, S. Gustafsson, J. C. Randall, S. Vedantam, T. Workalemahu, T. O. Kilpeläinen, A. Scherag, T. Esko, Z. Kutalik, the GIANT consortium, I. M. Heid and R. J.F. Loos (2014) Quality control and conduct of genome-wide association meta-analyses. *Nature Protocols* 9(5), 1192–1212.

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, and Visscher PM (2010),. Common SNPs explain a large proportion of the heritability for human heigh,. *Nature Genetics*, 42(7), 565–569.

Zhong, S., Israel, S., Xue, H., Ebstein, R.P., Chew, S.H. (2009), Monoamine oxidase a gene (maoa) associated with attitude towards longshot risks, *PLoS One* 4 (12).

Zuk, O., Hechter, E., Sunyaev, S. R. and Lander, E. S. (2012), The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109, 1193-1198.