Can social scientists use molecular genetic data to explain individual differences and inform public policy?*

Steven F. Lehrer, Queen's University, NYU-Shanghai and NBER

Weili Ding, Queen's University and NYU-Shanghai

NON-PRINT ITEMS

Abstract

This chapter surveys the relevance of genome sequencing to social science research. The cost of genome sequencing has rapidly fallen, delivering rich data on previously unmeasurable genetic differences across people that may directly and indirectly influence socioeconomic outcomes. We discuss how genomic research can inform policy challenges, suggesting that social scientists can play a valuable role in helping design policy based on research that links specific genetic factors to socioeconomic outcomes. Finally, we urge social scientists to use formal models and empirical tools to analyze this data and progress our understanding of the production processes underlying human developmental outcomes.

* We are grateful to both Gigi Foster and Brendan Wilson for their patience and detailed comments on earlier drafts, that both markedly improved the exposition and clarified many concepts in this chapter. We also wish to thank Pietro Biroli for initial encouragement to undertake this chapter. Lehrer also wishes to thank SSHRC for research support. The usual caveat applies.

Introduction

Heritability is generally defined as the proportion of variation in a population's observable characteristics or outcomes that is accounted for by genetic factors. The role of heredity in most socioeconomic outcomes ranging from income to educational attainment is not in itself a new revelation. However, until the human genome was decoded in 2001, it was considered unlikely that much could be done with this knowledge.

With the availability and sheer volume of datasets containing individual molecular genetic information growing at a rapid pace in recent years, the tantalizing possibility now exists to identify specific genes and the pathways through which they operate to drive important socioeconomic outcomes. More generally, Conley (2009) argues that this new information can be deployed to (1) assess the direct impact of specific genetic phenomena on socioeconomic and behavioral outcomes, (2) explore genetic–environmental interactions, and (3) trace genealogies across time and space. This knowledge may have substantial policy implications and may also be of use in refining social science theories to improve their realism and predictive accuracy.

This chapter focuses primarily on the findings of studies that fall under the umbrella of molecular genetics. These studies examine whether and how variation at specific locations in the individual genetic code is associated with individual socioeconomic or health outcomes. This approach differs from the main approach, drawn from behavioral genetics, that social scientists have historically employed to understand the role of genetic factors in explaining outcomes. Studies taking this more traditional behavioral genetics approach typically use data collected from family-based samples, such as twins or siblings. In this literature, researchers often assume that the driver of all variation in the outcome being investigated could be decomposed into additively separable genetic and environmental sources: the nature (genetic) effect and the nurture (environment) effect. Research using a behavioral genetics approach was recently surveyed in Behrman (2016) and first entered the economics literature in Taubman (1976). This approach has also been used with a sample of adopted children to understand the role of 'nurture' in producing outcomes (see, e.g., Sacerdote (2007)).

Findings from studies that use molecular genetic data have already produced profound implications for diagnostics, preventive medicine, and therapeutics. As our knowledge about the links between genes and complex socioeconomic outcomes such as educational attainment or behavioral traits continues to grow, societies will face critical questions, such as: should molecular genetic information be considered in the design of social and economic policies? Should genes come to play a central role in society's thinking about socioeconomic issues? In parallel, researchers are faced with the question of whether they wish to use these exciting new sources of data that allow them to enter the black box of what were previously known as individual fixed effects, or in other words, individual-specific permanent unobserved heterogeneity. Genetic markers may be truly what past researchers meant by permanent unobserved heterogeneity, since such markers are assigned at conception and, with the sole exception of monozygotic twins, differ markedly (potentially, according to 1000 Genomes Project Consortium (2015) on average at over 4.1 million locations on our DNA) across individuals. Social science researchers have historically employed fixed effects to capture permanent productivity characteristics of each individual, and data on the genetic markers themselves allows us to examine the nature and dimensions of these individual effects. Entering this black box, while tempting, may expose researchers to the accusation that their research endeavors or results implicitly promote social eugenics.

This chapter first updates the comprehensive reviews presented in Benjamin et al. (2007, 2012), Lehrer (2016), and Lehrer and Ding (2017) that explore the use of genetic markers in studies within economics by discussing the most recent findings. Second, this chapter contains a discussion of how genetic markers are influencing drug development, including consideration of the unintended consequences of policies that promote personalized medicine. With this chapter we aim to help social scientists interested in integrating genetic factors within their studies while being cognizant of the broader social and philosophical implications of such an effort. The chapter's sectioning is organized around the following questions: how is genetic data collected?; How is genetic data used by social scientists?; What policy implications arise from genetic evidence and the data collection methods that support it?; and finally,) Where might or should we go next?

In the next section, we provide a brief scientific primer on what genetic data is and how it is collected. We then summarize how social scientists, and particularly economists, have used this information in their empirical analyses. We draw distinctions between descriptive work, research that aims to establish evidence of causation in one primary direction, and studies that seek to identify gene-environment interactions in producing outcomes. Understanding how genetic markers associate with health and socioeconomic outcomes may have implications for public policy, a subject we then discuss with a heavy focus on innovation policy as it impacts the pharmaceutical industry, because genetic markers can be targeted in the delivery of specific treatment regimens (popularly referred to as 'personalized medicine'). We conclude the chapter by discussing promising directions for future research that continues economists' disciplinary tradition of simultaneously developing new tools and new models that incorporate the data drawn from those tools, so we can better understand how outcomes develop and – based on this knowledge – enrich both policy and academic discussions.

Scientific primer

As background, research on heredity dates back over 1000 years, but the mechanism of heredity that ignited the modern field of genetics did not receive widespread scientific attention until long after Gregor Mendel first published the fundamental laws of inheritance in 1866. Mendel's research was conducted with pea plants and led him to the insights that genes come in pairs and are inherited as distinct units, one from each parent. These insights were drawn from tracking the patterns of inheritance of seven different features between parental and offspring pea plants. Since Mendel was not an academic but rather a little known Central European monk, his work largely went unrecognized until 1900. Only recently has knowledge about the genetic factors that contribute to health and socioeconomic outcomes begun to emerge, and much of this knowledge has been generated in the last 15 years. To assist the reader who may be unfamiliar with the terms and jargon in the molecular genetics literature that we draw on in this section, we have created a glossary of several scientific terms at the end of this paper.

A brief review of the development of molecular genetics over the last century

Readers interested in an accessible review, designed for non-academic audiences, of the historical study of human heredity and its main findings related to the modern field of genetics are referred to Mukherjee (2017). The first half of the twentieth century saw the blossoming of what is now known as classical genetics. Many recent breakthroughs in our knowledge are due to a combination of important findings made in the second half of the twentieth century and recent technological advances. Perhaps one of the best known and most important findings in genetics that led to the development of molecular genetics as a field was published in 1953, when James Watson and Francis Crick described the double helix structure of deoxyribonucleic acid (DNA). DNA is composed of two strands of “nucleotides” coiled around each other and can be viewed as an immensely long ladder twisted into a helix, or coil where the nucleotides are linked (i.e. rungs on a ladder) together by hydrogen bonds. Each strand is composed of multiple instances of four complementary nucleotides: adenine (A), cytosine (C), guanine (G) and thymine (T). A nucleotide consists of a base (one of four chemicals: adenine, thymine, guanine, and cytosine) plus a molecule of sugar and one of phosphoric acid. These nucleotides are often referred to as the building blocks of DNA. A complementarity between the two strands of DNA arises since adenine on one strand always bonds with thymine on the other, and similarly, cytosine is always paired with guanine. The DNA “base-pairs” – i.e., the pairs of nucleotides that can be found at any cross-sectional slice of the two complementary DNA strands – are thus guanine-cytosine (“GC”) and adenine-thymine (“AT”).

DNA is hereditary material that contains detailed instructions, in the form of a set of biological messages, for how an organism needs to develop, live, and reproduce. DNA is located on 23 pairs of chromosomes in every cell of an organism that has a nucleus. To provide some additional intuitive understanding of the genome, we employ here the analogy provided in Lehrer and Ding (2017). Our DNA can be understood as an instruction manual composed of 23 chapters (chromosomes) that in total contain over 3.2 billion letters (DNA base pairs). The length of each chapter varies from 48 to 250 million letters (A, C, G, T) without any spaces. Although there are no spaces, a gene can be viewed as a paragraph in the chapter. Each gene is a segment of DNA that can vary in size from a few hundred letters (aka DNA bases) to more than 2 million letters. Thus, a single chromosome (chapter) can have hundreds or even thousands of genes (paragraphs) containing millions of letters.

The structure of DNA is formed at conception, when one member of each pair of chromosomes is inherited from the mother and the other from the father. Homologous chromosomes have the same genes arranged in the same order, but they have slightly different DNA sequences across individuals. Our DNA is able to produce variation in our individual outcomes insofar as part of the human genome – less than two per cent of it, according to modern measurement and classification techniques – encodes information to make proteins through the order, or sequence, of the nucleotides along each strand.[i] As we will shortly discus, knowledge that the DNA sequence of a gene determines the amino acid sequence of the resulting protein is crucial for how technologies have been developed to sequence our individual genetic code. In other words, one of the reasons that individuals differ from one another is that their DNA consists of different nucleotide sequences and, consequently, carry different biological messages regarding protein production.

Proteins are complex molecules involved in many critical functions of the body ranging from the production of antibodies to the transportation of substances, structure and sending messages. Hormones and enzymes that cause chemical changes and control all body processes are made of proteins. Proteins are required for the structure, function, and regulation of the body’s tissues and organs. For example, antibodies also known as immunoglobulins consist of proteins produced by white blood cells and play a critical role in the body’s immune response by specifically recognizing and binding to particular antigens, such as bacteria or viruses, and aiding in their destruction. Thus, if different amounts of proteins are produced in different people due to differences in the nucleotide sequences in their DNA, some individuals may produce lower levels of immunoglobulins than others and will be more likely to become ill when they encounter a virus. As another example, growth hormone is a protein produced by somatotropic cells that acts as a messenger to co-ordinate processes between different cells and organs to stimulate growth, which occurs at different times and rates across individuals.

Beginning in the mid-1970s, methods were developed to determine the order (sequence) of the nucleotides in a given sample of DNA to help uncover the genetic components of individual difference. To complete this goal, the United States Department of Energy and National Institutes of Health joined with numerous international partners in October 1990 to provide funding for and start the Human Genome Project. The Human Genome Project's goal was to sequence all 3.2 billion nucleotides, or base pairs, which is the complete set of DNA in the human body.[ii] DNA sequencing is the process of determining the precise order of nucleotides within a DNA molecule. Human beings are all 99.9 percent the same, DNA-wise, and human DNA is about 99 per cent the same as that of chimpanzees, our closest relatives. It has been suggested that if two individuals were selected at random they would only differ at about one in every 1200 to 1500 DNA base pairs. Most genome variations between a given pair of individuals are relatively small and simple, such as an A substituted for a T at a specific location on one strand. These single–base-pair differences across individuals are known as single nucleotide polymorphisms (SNPs).

The complete human genome sequence announced in June 2000 is a 'representative' genome sequence based on the DNA of just a few individuals. To accelerate the pace of medical discovery worldwide, all data generated by the Human Genome Project was made freely and rapidly available on the internet. In April 2003, researchers successfully completed the Human Genome Project.[iii]

Once the human genome was decoded, researchers in multiple disciplines strove to conduct studies that would elucidate how each of the many parts of our chromosomes works with the others in generating individual outcomes. Motivating much of this research is the goal of understanding individual outcomes that are hypothesized to be polygenic, i.e., due to multiple genes where each gene may play a small role. Unlike outcomes such as sickle cell disease and cystic fibrosis that can generally be explained by alterations in a single gene, many outcomes of research interest are the product of numerous genes, each with a small effect and often interplaying with the environment. The susceptibility of individual genes to contribute to certain outcomes may vary with environmental factors. Central to informing research objectives in this area, the National Institutes of Health began in 2005 to produce a catalog of common genetic patterns that is referred to as the HapMap (http://hapmap.ncbi.nlm.nih.gov/).

The HapMap can help to identify the locations on the human genome of outcome-relevant genetic variation, which aids researchers in developing hypotheses and new types of tests that measure genetic variation. In the first HapMap published in 2005, approximately one million SNPs were genotyped (i.e., this is the process of determining if there are differences in the genetic sequence of an individual relative to the DNA sequence of a reference individual. In 2007, the second HapMap was published, containing descriptions of over three million SNPs. Updated versions of the HapMap continue to be issued, rapidly expanding our knowledge of SNPs, in turn improving the accuracy of the unique individual-specific genetic fingerprints used in identity testing and other applications. It has been estimated that there are roughly 10 million SNPs on the human genome – on average, about one instance per 300 base pairs – on which a mutation (i.e., a single nucleotide being different) commonly occurs in humans. Thus, only at every 300 base pairs as we move along a DNA strand is there a possible genetic difference between two random individuals.

Genetic researchers use terminology such as “rs15260(A;C)” to indicate someone with a sequence of A and then C, at location rs15260 of their genome, which is a specific position on a chromosome at which a particular SNP appears. At this location, most people might have the nucleotide "A" on one strand and “A” on the other, and a small subgroup might have an alternative base pair (i.e., CC or AC). No distinction is made between AC or CA and researchers do not distinguish if you inherited the A from your mother or father. Each variant of a SNP that has been observed in humans is called an “allele”. In other words, alleles describe variant forms of given gene that are found at the same place on a homologous chromosome. The most common allele on a given SNP is known as the major allele, and the less common allele is sometimes called the minor allele or the risky allele. For example, a frequently studied SNP is called “TaqI DRD2” and is located at chromosomal position rs1800497. This SNP was originally believed to play a role in determining the density of dopamine receptors in the brain. These receptors play a key role in transmitting signals across regions of the brain, and fewer receptors meant signals would take longer to be received. Across individuals, there are genetic variants in this SNP known as A1A1, A1A2 and A2A2, where A1 is the risky allele. That is, carriers of the Taql DRD2 A1 allele have significant loss of dopamine receptor density in the brain, and this is often hypothesized to be linked to poor outcomes.

In practice, researchers often refer to the number of possible genetic variants of a given SNP by the number of risky (i.e., low-frequency) alleles that the SNP contains. In other words, rather than referring to someone as having variant “rs1800497(A1;A1)”, researchers will refer to this person as having “two risky alleles for the Taq1 DRD2 gene”. At present, it is believed that only a very small minority of all the known SNPs play important roles in influencing the function and structure of the human body. These roles could be selectively advantageous or disadvantageous (of which the latter possibility is the source of the term “risky”), and the genetic material accounting for these SNPs takes up less than 0.1 per cent of the human genome. It is this genetic material that is most frequently targeted by current research aiming to explain the genetic underpinnings of observed differences across human individuals in socioeconomic outcomes.

Collecting molecular genetic data

To obtain genetic data, most academic data collections as well as commercial direct-to-consumer companies that provide genetic reports use a procedure called a buccal smear. The survey participant or consumer is provided with a vial and a small brush or cotton swab that is used to collect a sample of cells from the inside surface of the cheek. These cells are then placed into the vial and sent to a laboratory for sequencing. While buccal smears are popular, genetic tests can also be performed on samples of blood, hair, skin, and amniotic fluid, among other tissues.

While there are many potential SNPs, researchers’ ability to measure genetic factors is constrained by the type of test being used by the laboratory. Most tests used take measures not of the whole genome, but of just a targeted section. As a matter of terminology, a ‘genetic test’ examines a targeted section of DNA whose genes have a known function such as producing a protein, whereas a ‘genomic test’ investigates large sections of genetic material and information where there is often no specific genetic target.[iv] Often the target of genetic test is a region of the genome (often called an ‘exon’) that codes for the production of specific proteins In other words, scientists use DNA microarrays to measure the expression levels of large numbers of genes simultaneously. Suppose the area of a specific exon contains 600 nucleotides. Every three nucleotides correspond to one codon and one amino acid. So, this area would contain 600/3=200 codons and the resulting protein contains 200 amino acids. The size of a protein is often expressed as its molecular mass and the genetic test measures the amino acids produced that are translated from a triplet of nucleotides at that point, which is roughly how one recovers the SNP variant since the DNA sequence of a gene determines the amino acid sequence of the resulting protein. This form of sequencing limits the locations to where sequencing occurs to roughly one per cent of a person's genome where exons are present. Many sequencing protocols focus only on the exome, which means they measure only the roughly 2% of genes that codes for protein. In other words, the exome is the collection of all exons. While methods to undertake whole genome sequencing are also available they are both more costly and time consuming.

Numerous tests and microarrays have been developed. At present, most microarrays used to measure SNPs in datasets social scientists employ focus exclusively on exons. In general, the quality of a given genetic test depends on the average number of times each base pair in the genome is read during the test's sequencing process. Sequencing determines the precise order of nucleotides and testing methods must identify which of the four nucleotides (A, C, G or T) is located at a specific point in a strand of DNA. Since reading each base pair in every chromosome can become quite expensive in terms of time and resources, many data sets used to produce research published in both scientific and social science disciplines report an imputed SNP. To impute a SNP, geneticists rely on what is known as 'linkage disequilibrium'. Two SNPs on the genome are said to be in linkage disequilibrium when the patterns observed in their alleles are related in a population. High linkage disequilibrium – perhaps counterintuitively to the ears of economists – means that the SNPs' particular allele patterns are almost always inherited together. High linkage disequilibrium thus means that by having accurate information on the alleles present in neighboring SNPs, one can take a well-informed guess of the alleles on a given SNP an individual has inherited.

The fact that many SNPs are imputed is generally ignored in subsequent empirical analyses (i.e., researchers typically treat the data as if it were measured without any error). The genetic data available to researchers is a product of the type of sequences that have been assessed (whether the whole genome or particular SNPs) that determine the content of the data, and the number of steps and activities involved in the data collection process that determines the overall quality level of the data. Decisions about the type and quality level of the genetic sequencing conducted in laboratories are often influenced by budgetary and other considerations not directly related to a social scientist's likely research agenda, and these decisions jointly affect which SNPs are available for analysis and the potential degree of estimation error that afflicts the data provided.

In the next section, we briefly summarize the molecular genetic research conducted by social scientists, primarily utilizing data on SNPs, that has generated important new empirical and theoretical insights. These insights are limited at the moment but will likely grow at a rapid rate as the declining cost of genetic sequencing will enable more social scientists to augment their data collections with genetic information.

New insights from social science research using genetic information may have substantial economic value, potentially delivering a second wave of innovation and windfall gains on the heels of the first. The Human Genome project spurred a revolution in biotechnology spending and innovation around the world. Results from the Human Genome project are claimed in Battelle Technology Partnership Practice (2011) to have generated US$796 billion in economic activity in the US alone. Revolutionary advances in DNA sequencing technologies have enabled rapid, low cost determination of individuals' DNA sequences, increasingly marketed direct to consumers (e.g., by companies such as ancestry.com and 23andme). The growing field of pharmacogenomics examines how genetic variation affects an individual's response to a drug, and scientists have been able to leverage findings from this field, together with techniques founded in molecular biology and genetics, spawning the creation of new goods such as new crop varieties. The continued application of genomics in this area could lead to agricultural products with specific nutritional content, or products of a specific size or texture that could yield economic benefits by lowering the cost of shipping. However, some envisioned applications have raised ethical concerns in the scientific community about the control of access to data on genetic markers, as well as public concerns about the appropriate use of genomic information.

Social science research using genetic data

Social scientists who utilize molecular genetic data as a source of explanatory variables in their analyses must decide how these measures should be introduced. Some researchers create a single measure, such as the count of risky alleles for a SNP. This formulation imposes the assumption of linearity in the effects of risky alleles on the outcome under consideration, while having the advantage of leading to a sparser set of covariates than many alternative formulations. Other researchers create a set of indicator variables, each of which flags a particular variant of the risky allele relative to the base category. To illustrate, let us return to the Taq1DRD2 gene that has variants A1A1, A1A2 and A2A2, where A1 is the risky allele. Using the first approach mentioned above, an individual's genetic material present on the Taq1 DRD2 SNP is coded as zero (for A2A2 individuals), one (for A1A2 individuals), or two (for A1A1 individuals). Using the second approach, this same variation is captured with a vector of dummy variables for having the A1A1 variant and the A1A2 variant, with estimated effects of each variant compared to the most common (A2A2) variant that serves as the base category.

One advantage of the second approach is that there is no need to make a functional form assumption about the way that the genetic information affects the outcome of interest. A simple statistical test can be employed to determine whether the estimated effects indicate that the linear restriction implied by the first approach is warranted. A second advantage of the indicator variable approach relates to the interpretation of genetic effects. Genetic markers are immutable characteristics that are fixed at conception. Greiner and Rubin (2011) point out that immutable characteristics can be viewed as treatments, since many outcomes are determined through the mediation of perception rather than directly from those immutable characteristics, and perceptions are not immutable. In this "potential outcome" framework, a specific SNP variant could be viewed as a legitimate treatment: any difference in outcomes between two individuals who are identical in all other characteristics except for their genetic sequence on that specific SNP could be attributed to that genetic difference – with the exact mechanism generating the difference in outcomes remaining unspecified and open to further inquiry. In contrast, using the number of risky alleles as the approach to capturing the genetic information would not only restrict the mechanism to operate solely through the count of risky alleles, but would also require maintaining the assumption of continuous treatment effects: the (single) estimated marginal effect on the outcome would be identical when moving zero to one risky alleles as when moving from one to two risky alleles. On balance, discretizing genetic information in empirical work through the creation of large dummy variable arrays may be preferable, at least at this stage in our understanding of the link between genetic information and socioeconomic outcomes.

Existing work by social scientists falls under one of the following three areas: reporting associations; recovering causal effects; and exploring the interactions of genes with the environment that are relevant to outcome generation. We next summarize some of the key developments in each of these areas.

From candidate genes to genome-wide studies

Initially, social scientists who incorporated molecular genetic data into their research programs focused on a handful of genetic markers. These markers, called 'candidate genes,' were generally those occupying specific pre-selected regions of our DNA. These locations were not selected randomly, but rather were chosen as the target of genotyping either because of being suspected of being directly involved in generating the outcome, or because the types of protein encoded by the candidate gene(s) located there may logically suggest that those genes could influence the outcome being investigated. A candidate gene study ex ante selects one or more specific genetic markers – essentially a set of tested SNPs – to investigate. Candidate gene studies address the question of whether the specific SNPs studied are associated with outcomes of interest to social scientists. These associations may arise if the test SNP is directly associated with the outcome, or if it is indirectly associated with the outcome because of linkage disequilibrium, whereby another SNP whose occurrence is correlated with the test SNP is the one that directly affects outcomes. Most researchers producing candidate gene studies assume they are identifying a direct association, and further work would be required to rule out indirect channels. As an example, some early work (e.g., Zhong, Israel, Xue, Ebstein, & Chew (2009) and Dreber et al. (2009)) tried to provide a biological micro-foundation to utility maximization by examining the associations across people of patterns on particular SNPs with measures of economic primitives collected in the laboratory, such as risk aversion and time discounting. The idea of indirect effects is that intergenerationally transmitted traits affecting risk aversion and time discounting may arise from schooling decisions made by parents and not from predetermined characteristics inherited from one's parents.

Candidate gene studies were easy and quick to undertake, making them seductive to researchers. Numerous researchers generating these early studies do not explicitly explain how their putative candidate genes were chosen. Many studies appear to have been carried out mainly due to data availability, with ex post justification rather than a clear theoretical rationale provided for the exercise. Further, concerns regarding proper scientific practice in this area have emerged. Many early studies lack statistical power, yielding potentially false-positive results. As the number of studies using this approach increased, it became apparent that many of the early results could not be replicated in analyses undertaken with other samples. For example, Chabris et al. (2013) illustrate several points about the limits of candidate gene studies by trying to replicate previously identified candidate genes using data from three independent longitudinal studies. Their results are disappointing from a replication perspective, since they found fewer significant associations than a traditional power analyses would have predicted ex ante.

To help ensure that evidence generated using the method of candidate genes would be credible, in 2012 the academic journal *Behavior Genetics* adopted strict standards for the publication of candidate gene studies (Hewitt 2012). To be considered for publication, any candidate gene study must be well powered and make corrections in statistical inference for multiple testing, and any new finding must be accompanied by a replication. These higher publication standards have meant that conducting candidate gene studies is relatively less appealing today than undertaking research that seeks to identify the associations of characteristics or behavior with measures of SNP variation across the full genome.

Studies using information on SNPs across the genome are typically designed as a data mining exercise, requiring no prior knowledge, imperfect as it might be, of which genes are likely to be related to the outcome of interest. Over the last decade, these studies have involved increasingly larger sample sizes that are generally constructed by combining multiple datasets. Each of the data sets being combined is required to have measures of the same set of common SNPs as well as the target outcome measure. The sampling criteria across the pooled data sets are often not identical, with social science surveys that have well developed sampling frames being pooled with voluntary response samples such as those from 23andMe (where participants must elect to send a sample to 23andMe for genotyping) and case-control studies, in which data is collected separately on groups who differ in some outcome. Despite the reduction in external validity of any findings that this may produce, proponents argue that drawing genetic information about narrow demographic groups from the large pooled data sets can better support the detection of robust evidence linking outcomes to genetic variation, even when the associations are modest in practical terms.

The standard approach taken in a large-scale genome-wide association (GWA) study begins with estimating a model on a 'training sample'. The training sample utilizes all the observations contained in most, but not all, of the assembled datasets. To convince the research community that a given GWA study has detected a true association, it is now standard that researchers examine whether the results using the training sample replicate using other datasets that were held out from the initial analysis. These additional datasets are then referred to collectively as an 'evaluation' (or 'test') data set. In spirit, this approach is in line with that of time series econometricians comparing the accuracy of alternative strategies used to calculate an economic forecast, as illustrated in the box office revenue prediction exercise discussed in Lehrer and Xie (2017).

The implementation of this type of design is supported by recent data-sharing initiatives, such as the pooling of databases collected by individual research teams under the stewardship of the Wellcome Trust Case Control Consortium (https://www.wtccc.org.uk), with the stated aim of improving the understanding of the aetiological basis of several major causes of global disease.

GWA studies can appear complicated to those familiar with conventional social science research, as the method relies on understanding work conducted in the statistical genetics literature in which scientists use different terminology than what is used in the econometrics literature. However, the development of off-the-shelf software designed to implement a GWA study promises a dramatic reduction in the barriers to entry. Not only does off-the-shelf software obviate the need to fully comprehend the underlying statistical genetics literature, HapMap-based genotyping platforms further facilitate the genome-wide approach by enabling theory-blind data mining across the genome in search of possible sources of variation relevant to a particular outcome.[v]

In practice, an important ingredient in any GWA study is how to choose focal SNPs optimally, based on HapMap data, to maximize the regional genomic variation that the researcher will subsequently be able to use in identifying genetic effects. Intuitively, since markers located near each other are often inherited jointly, increasing the independent variation available in the data drawn from each SNP, thereby avoiding the covariance produced by linkage disequilibrium, means selecting SNPs that are reasonably far apart on the genome. Researchers often select one SNP among a highly correlated set of SNPs to include in a specification. While this may sound promising, there is a trade-off: in order to make a valid claim about the source of the genetic effects found, the researcher must choose from among the SNPs located in a particular chromosomal region the one that is in fact responsible for the outcome, either on its own or through interaction with the environment. If the wrong SNP is selected, the researcher may falsely attribute an association between the included SNP and the outcome to the influence of genes on the included SNP, rather than to genes on an omitted SNP whose genetic information is strongly associated with that appearing on the included SNP, due to linkage disequilibrium (a particular form of omitted variable bias). The challenge posed by linkage disequilibrium varies across racial and ethnic groups due to both their population size and their migration history. For example, the mean size of regions of strongly associated SNPs, sometimes called haplotype blocks that are defined algorithmically in a ubit of measurement called kilobases, where 1 kilobase is equal to 1000 base pairs of DNA. For populations of European or Asian ancestry, a halotype block, is estimated to be 22 kilobases versus only 11 kilobases in populations of recent African ancestry. Because of this, in a GWA study, researchers only consider individuals of specific ancestry.

Perhaps the best-known example of GWA is illustrated in a sequence of papers carried out by The Social Science Genetic Association Consortium (www.thessgac.org) that seeks to understand the associations between genetic markers and educational attainment. This outcome was selected since it appears in a multitude of datasets. The most recent project, Lee et al. (2018), uses data on 1.1 million individuals and identifies 1271 independent genome-wide-significant SNPs. The genome wide significant SNPs are those with p-values that are below a specific threshold for statistical significance is critical to control the number of false-positive associations. Currently, standard practice is to use a genome-wide significance p-value threshold of 5*10E-8 to judge whether a SNP is significantly associated with the outcome under consideration. Since there are more possible hypotheses of significant association than data points (the authors could potentially include data on 7.1 million SNPs), one must make corrections for multiple testing. While the use of a genome wide significant p-value is often credited to the original HapMap studies, this would not be considered optimal to hold the level of committing a type one error from a multiple testing perspective. In Lee et al. (2018), a Bonferroni corrected p-value threshold would be lower and is needed to hold the overall type 1 error rate at the desired level, taking the value of 1.25*10E-8. The Bonferroni procedure is often viewed as being quite conservative in the multiple testing literature since it divides the overall significance level by the number of hypotheses undertaken in the study. If the Bonferroni threshold were applied to the Lee et al. (2018) results, only 1024 of the 1271 SNPs would have been judged statistically significant.[vi]

Lee at al. (2018) presents a marked extension of an earlier GWA study of educational attainment conducted by the SSGA consortium. Prior work in this stream of research investigating the genetic basis of educational attainment appeared in Rietveld et al. (2013a), who combined data on 42 cohorts consisting of over 100 000 individuals, and Rietveld et al. (2014) who further expanded that dataset. Okbay et al. (2016) was the third study, conducting a GWA of roughly 300 000 people, and finding 74 SNPs associated with educational attainment. Akin to prior GWA of educational attainment, in Okbay et al. (2016) only one trait/outcome was considered, whereas Lee et al. (2018) use a recent methodological extension to classical GWA to examine how the same set of genetic markers are associated with multiple cognitive traits. The most striking finding from the series of papers completed by the SSGA consortim appears in the portion of Okbay et al. (2016) that conducted a replication with a test dataset involving 110 000 individuals from the UK Biobank, 72 of the initially identified 74 SNPs remain significantly associated with educational attainment. The authors conduct numerous robustness checks of their main analyses where they ensure common support is imposed across samples by excluding dissimilar individuals and also consider alternative sets if control variables to capture any potentially unobserved confounding genetic differences across the samples used in the main analysis. Further, they utilize the latest quality control protocols being applied in the medical genetics literature (Winkler et al. 2014) and carefully account for population stratification, defined formally in the next section, to ensure that similar people are being compared across the combined datasets. This line of research holds the potential to help us map the molecular basis for educational attainment, although the economic significance of each of the individual 74 SNPs is found to be quite small. Further, in aggregate, the 74 SNPs identified in Okbay et al. (2016) explain only 0.43 per cent of the variation in educational attainment across individuals in the sample.

The idea of using larger sample sizes to detect the true association between genetic variation and a disease, assuming such an association exists, is motivated by statistical power considerations. To retain power, sample sizes must increase with the following: higher odds of a type one error emerging, as more SNPs are included and more association tests performed; higher odds of measurement error in either the outcome or the explanatory variables; smaller magnitude of the genetic effect; lower frequency of the risky allele; and increased importance of omitted factors, including heterogeneity in the underlying association being estimated, caused for example by multiple genes that contribute to the disease, ancestry differences across population subsets, or gene-gene or gene-environment interactions.

One of the main outputs from GWA studies that proponents suggest could be useful for social scientists is what is known as a polygenic score. A polygenic score is constructed as a weighted sum of the individual risky alleles for each SNP used in the study that is reliably related to a particular trait or outcome, where each allele is weighted by its effect size as estimated from a GWA study (Dudbridge 2013). In practice, different methods are used in different studies to construct polygenic scores and there does not appear to be a consensus emerging on the most appropriate way. The alternative methods primarily differ on the weighting schemes and on which SNPs should be included in the calculation. In all methods, the underlying idea is that based on GWA study results, we can apply weights that indicate relative importance in generating the focal outcome to the measure of genetic information present on each SNP. The resulting polygenic score can be used by researchers to exploit the joint predictive power of many SNPs within an estimating equation to predict a focal outcome.

As an explanatory variable, a polygenic score will explain more variation in outcomes than any set of individual SNPs and can also accommodate the possibility of combined genetic influence. From an econometric perspective the score is a generated regressor, an issue that most analysts using polygenic scores in their models ignore. From a behavioral perspective, the score is just a linear combination of different candidate causal factors, and implicitly makes assumptions about the relative substitutability of those factors (i.e., of the genetic variation present in different SNPs) within the total effect-generation mechanism. From a more policy oriented or therapeutic standpoint, polygenic scores provide a way of identifying individuals at high risk for certain outcomes.

GWA studies have given rise over time to polygenic scores that can predict a significant amount of variation in important outcomes. For example, the polygenic score constructed in Lee et al. (2018) from their GWA can explain 11 per cent of the variation in educational attainment of participants of The National Longitudinal Study of Adolescent to Adult Health (http://www.cpc.unc.edu/projects/addhealth), and 13 per cent of the variation in educational attainment using data from The Health and Retirement Study (http://hrsonline.isr.umich.edu/)). These figures represent marked increases in the predictive power of generated polygenic scores relative to those calculated from earlier GWA studies, including Okbay et al. (2016). To provide more context for how well their score can predict educational attainment, Lee et al. (2018) show that their constructed score does a better job of predicting educational attainment than household income but is a worse predictor than either mother's or father's education. More concretely, in specifications that control for all demographic variables jointly, the score's incremental R-squared is 4.6 per cent.

Papageorge and Thom (2017) and Barth, Papageorge, & Thom (2018) each present an early application of polygenic scores in labor economics. The former study, using the Health and Retirement Study (HRS), presents evidence that the polygenic score that predicts educational attainment is also associated with higher wages, but only among individuals with a college education. Further, suggestive evidence is provided that the genetic gradient in wages has steepened in more recent birth cohorts, which the authors suggest is consistent with interactions between technological change and labor market ability (what might be termed good-gene-biased technological change). In the latter study, evidence is presented using the same dataset that the polygenic score from Lee et al. (2018) can predict wealth at retirement. The authors suggest that the polygenic score is a proxy variable for one's ability to navigate complex financial choices.

At present, research using GWA is more favored within the research community than that using the candidate gene approach. On the one hand, without strong prior hypotheses, the agnosticism of a GWA study regarding theories of outcome determination, and its survey of the entire genome, hold appeal. It is a pure empirical exercise. On the other hand, the specific empirical specification used in a GWA study is not innocent: any specification imposes strong assumptions on how genetic factors are linked to the outcome under consideration, including the absence of gene-by-environment interactions. In practice, these assumptions are often implicit and not well justified as being reasonable in the empirical application. This runs counter to standard practice in most social science research using conventional methods. Hence, we predict that as the methodology becomes better understood, more criticism will emerge from social scientists about this aspect of GWA research.

More generally, it is hard to see how evidence from GWA studies contributes to existing literatures in the social sciences that are informed by an underlying behavioral model. For example, consider the earlier GWA of educational attainment. A voluminous literature in multiple disciplines examines how individuals make a sequence of education choices (e.g., what courses to choose in high school, whether to apply to college, which college to attend, what major to select, whether to persist in higher education, and so on) and generally postulate that faced with imperfect information about their options, individuals, trade off the expected costs and benefits of each potential choice. In most situations, individuals themselves have imperfect (if any) knowledge of their own genetic code when making these decisions. If individuals differ in terms of the genetic markers that can explain educational attainment, do these differences change how benefits or costs are assessed in ways that should be accommodated in our behavioral models – or is genetic information already encapsulated implicitly in these models, through what is presently known as preference-based heterogeneity? Making this distinction is important for social science modelers. Put differently, if one believes they have the genes for education and success as calculated by the polygenic score of Lee et al. (2018) does this influence their effort on studying, hiring a tutor or making decisions to persist? Do genes affect constraints, and if so which ones; or does their influence directly affect utility as speculated in several candidate gene studies.

Further, studies in the social sciences often seek to understand the scenarios, such as particular environments or types of samples, in which significant effects emerge. In the case of a large scale GWA study, the odds of finding a significant effect are higher when a specific genetic variant has a similar effect in all samples, which themselves are likely differ in terms of environmental and sample characteristics. This means that the variation in environment and sample type that the social scientist would normally use to identify effect strength across the whole population does not play a role in effect identification in a GWA study. Indeed, the method is likely to mask effects that are highly heterogeneous across context or sample. Second, genetic researchers in the scientific literature are generally interested mainly in gauging the total amount of variance in outcomes that the included genetic information explains (e.g., in calculating the R^2 statistic for the full set of genetic factors included in the specification), and point estimates are generally not the focus. This differs from the orientation of most studies in the social sciences, where researchers often examine the sign, magnitude and statistical significance of key explanatory covariates included in a baseline preferred specification and complement their baseline results by investigating their sensitivity to alternative specifications. In summary, while GWA techniques transplanted into social science from the scientific literature hold the potential to uncover associations between genetics and traits and behaviors of interest to social scientists, there are dimensions of methodological tension between GWA and more conventional approaches that will take time to resolve.

Moving beyond association: Using genetic markers to estimate causal effects

Angrist and Pischke (2010) recently argued that what they term the credibility revolution in empirical economics has spawned many research studies over the last thirty years that increasingly exploit plausibly exogeneous variation to identify causal effects that are of interest to both academic and policy audiences. Whatever its source, this undeniable trend has spurred significant controversy within the profession as to whether causal effects (as opposed to structural parameters that can be tied to an underlying behavioral model) are of prime interest, and whether researchers are choosing projects based on the availability and plausibility of identifying variation, irrespective of the importance of the research question being addressed. [vii] Parallel debates have occurred within economics and epidemiology regarding whether studies that use genetic data are thereby capitalizing on a source of exogenous variation with which to identify the impact of specific health conditions on socioeconomic outcomes. Using genetic information as a source of exogenous identifying variation was first introduced in economics by Ding, Lehrer, Rosenquist, & Audrain-McGovern. (2009), who essentially used candidate genes as instruments to understand the impact of health outcomes on academic performance using instrumental-variable techniques.[viii] This empirical approach requires the researcher to assume that the genetic instruments are not only correlated with health outcomes,[ix] but that they only influence academic outcomes through their influence on health.

The main empirical finding from this study is that depression and obesity each lead to approximately a one standard deviation reduction in academic performance. This deterioration is shown to differ by gender: young women's academic performance is found to be more adversely affected than young men by negative physical and mental health conditions. Lastly, using genetic instruments, the separate estimated impacts of inattention and hyperactivity on academic performance differ sharply in magnitude and sign. The instrumental variable estimates are substantially larger in magnitude relative to OLS estimates that ignore the endogeneity of health. The differential effect of inattention and hyperactivity is not observed if one does not decompose the diagnosis of ADHD into being clinically inattentive (AD) or clinically hyperactive / impulsive (HD). These results indicate that there are only poor health consequences from AD and the authors speculate that this may arise since parents, peers and teachers may be more likely to respond with extra investments to a child with HD than AD.

Concerns regarding the plausible exogeneity of the genetic instrument used in Ding et al. (2009) have emerged. These concerns have focused on population stratification, pleitropy, and dynastic effects. We expand on these concerns in the next two paragraphs.

The concern about population stratification is based on the existence of subtle genetic differences between groups of individuals that are not accounted for in the model, and the resultant possibility that the gene being used as the source of exogenous variation is correlated with a missing genetic marker related to group membership that is itself driving the results. Pleiotropy is the phenomenon of a single genetic variant influencing multiple traits. Pleiotropy is likely to be widespread in the human genome and was first pointed out as a concern in Conley (2009). More recent work has shown that if pleiotropy arises because the SNP instrument(s) influences one trait, which in turn influences another (known as 'vertical pleiotropy'), then instrumental variables strategies can still be used. However, if pleiotropy arises due to the SNP instruments influencing two traits through independent pathways ('horizontal pleitropy') then there is a greater chance that the instrument is invalid due to violation of the exclusion restriction (meaning in the context of Ding et al. (2009) that the genetic marker in fact belongs in the outcome equation itself, rather than only influencing outcomes via health). Hemani, Bowden, & Davey Smith (2018) presents a recent review of methods that researchers can employ to assess whether horizontal pleiotropic associations are a concern, and many of these tools nicely complement the work of Conley, Hansen, & Rossi (2012) that Ding et al. (2009) encourage researchers to use as a form of sensitivity analysis in studies using genetic instruments.

Dynastic effects relating to the line of heredity present obvious challenges to the use of genetic information as instrumental variables, since, where unobserved, such effects may confound the estimates. For example, consider using a genetic marker for a particular poor health condition in children as an instrument to identify the effect of that health condition on children's academic performance. Without more detailed data on parental diagnoses as well as parental genes, we cannot use the intended instrument to separate out the portion of academic performance that is uniquely due to the child's condition. The IV effect estimated may include the impact of family environments provided by the parents whose own poor health, which partly fed into those environments, can be explained by the same genes that were chosen as instruments for the children. Despite this concern that the instrument in such a case violates the exclusion restriction, in fact impacting academic performance through channels other than child health, we suggest that recovered estimates are still of policy relevance since individuals are in general not randomly assigned to families, and policymakers are generally interested in the total impact of these disorders.

At a fundamental level, one will never know whether a specific candidate gene is a valid instrument, since one cannot randomly assign genes to humans or create human equivalents to knock-out mice. A knock-out mouse is a genetically modified mouse in which researchers have inactivated, or 'knocked out', an existing gene by replacing it or disrupting it with an artificial piece of DNA. By causing a specific gene to be inactive in the mouse and observing any differences in the mouse from normal behavior or physiology, researchers can infer the probable function of that gene. Considering that such experiments are impossible to conduct in humans due to ethical concerns, Ding et al. (2009) suggest that researchers using genes as instruments should apply Conley et al. (2012)'s proposed 'local to zero approximation' method that methodically tests the sensitivity of the results obtained to the degree of plausibility of the identifying assumptions.

The analysis in Ding et al. (2009) points out an additional concern that applies more broadly to many studies seeking to understand the causal effect of poor health on academic and labor market outcomes – including those that do not use genetic instruments. Comorbid health conditions, defined as conditions in which two or more disorders or illnesses occur in the same person (whether simultaneously or sequentially), are frequently observed in humans. Ding et al. (2009) show that failing to account for comorbid diagnoses would result in biased estimates of the causal effect of specific health diagnoses on socioeconomic outcomes. The authors also suggest that comorbidity can strongly influence whether genetic instruments satisfy the exclusion restriction criteria that must be satisfied to justify their use as instruments.

The issue of comorbidity has broad implications for genetic studies in many subfields. Recently, Brickell et al. (2018) provide evidence of a strong association between a polygenic score predictive of a diagnosis of ADHD (R^2 = 0.83–1.69%) and a broad range of childhood psychiatric symptoms among children aged nine to twelve in Sweden. This suggests that common genetic risk variants associated with ADHD also influence a general genetic liability towards broad psychopathology in childhood, hinting strongly at common genetic bases for multiple co-occurring psychiatric conditions. This type of finding reinforces the challenge that comorbidity poses in disentangling the underlying genetic basis of disease and reinforces the need to consider carefully the justification for particular specifications of empirical models in studies that use genetic data.

The challenge of comorbidity is also a motivation for calculating polygenic scores. Most definitions of health are based strictly on the presence or absence of symptoms, and when similar symptoms are shared by multiple health conditions, diagnosis of the true underlying condition may be delayed. However, by comparing a patient's polygenic scores for different conditions, doctors may be better able to judge the relative likelihood of multiple candidate comorbid conditions underlying the presenting patient's symptoms. As the predictive accuracy of polygenic scores continues to increase, genetic markers could not only increase the palette of options for pursuing instrumental variable strategies, but also help solve both econometric and diagnostic challenges created by comorbidity.

In the epidemiological literature, the use of genetic information as a source of identifying variation is termed Mendelian randomization. Mendelian randomization was first proposed in Katan (1986) and applied empirically in Davey Smith (2003). To fully merit the term 'randomization', Mendelian randomization would require the absence of dynastic effects. However, genes are inherited by design from one's parents, and those parents also transmit environments and behavioral traits across generations. Lehrer and Ding (2017) suggest that epidemiological studies relying on what is claimed as Mendelian randomization would be more accurately described as following a 'Mendelian encouragement' design. Even in the presence of dynastic effects, genetic markers still encourage certain traits and behaviors, and their influence on outcomes at the population level may still be significant despite obvious ways for individuals not to comply behaviorally with their genetic assignments.

Within the epidemiology literature, important methodological developments have addressed a variety of concerns relating to the use of SNP variations as instruments. One set of extensions, proposed in Bowden, Davey Smith, & Burgess. (2015), uses Egger regression analysis, a tool initially developed in Egger et al. (1997) to detect small study bias in meta-analyses. Bowden et al. (2015) demonstrate that this alternative estimator can recover consistent parameter estimates even if all genetic instrumental variables are invalid by invoking an assumption about the strength of the instruments in the first-stage independent of their direct effects on the outcome. While this approach can mitigate problems owing to weak instruments, it does not deliver an estimate interpretable as a local average treatment effect (cf. Imbens and Angrist (1994) for the assumptions required to merit this interpretation), and it is unclear how exactly how to interpret the estimates recovered. While these alternative strategies developed in the epidemiological literature can recover a consistent parameter estimate under certain assumptions, the recovered estimate does not have a direct causal interpretation akin to that of the conventional instrumental variables estimate if there is treatment effect heterogeneity. Other recent innovations face similar challenges related to interpretation. Examples include Gage et al. (2017)'s 'bidirectional Mendelian randomization', proposed to deal with potential reverse causation (i.e., causal effects of the outcome on the endogenous regressor) and Rees, Wood, & Burgess (2017)'s 'multivariable Mendelian randomization', designed to deal with a finite set of ex ante known possible pleiotropic characteristics of the SNP based genetic instruments.

A final emerging approach to recovering causal effects involves using polygenic scores as instruments. As noted above, these scores can explain far more variation in the variable being instrumented than can variations in individual SNPs. However, because polygenic scores contain information drawn from multiple SNPs, some of which may independently affect the focal outcome, the likelihood of violating of the exclusion restriction assumption is increased. Further, many of the gene variants used to construct the polygenic score could have pleiotropic effects that lead to indirect influence on the focal outcome. Conley (2018) elucidates further concerns with recent proposed extensions to instrumental variable methods that involve polygenic scores as the source of identifying variation. Debates about polygenic scores extend beyond considerations of their validity as instruments in support of causal identification. Purcell et al. (2009) list concerns about their usefulness, whereas Belsky et al. (2012, 2013) provide empirical examples illustrating their potential benefits.

Within economics, a final variant on the instrumental variable strategy that exploits genetic inheritance within full biological siblings was introduced by Fletcher and Lehrer (2009a, 2009b, 2011). Fletcher and Lehrer coin the term 'genetic lottery' to motivate the use of an instrumental variable estimator for family fixed effects based on genetic information. Assuming a genetic lottery operates on humans is intuitively what is meant by the term 'Mendelian randomization'. Controlling for family fixed effects as instrumented by genetic information removes any (fixed) dynastic effects on outcomes that are shared between full biological siblings.

This strategy exploits variation in genetic inheritance within families and can also be used to test an important empirical model known as the family fixed effects estimator.[x] Researchers using this estimator to estimate a casual effect essentially compare siblings (or twins) with different exposure to a particular treatment. This estimator is popular since it allows researchers to eliminate all family level correlates of treatment likelihood and can be carried out when there is no quasi-experiment to exploit. However, researchers must assume that all within-family decisions related to treatment take-up are exogenous. The genetic lottery approach of Fletcher and Lehrer relaxes that assumption, allowing for endogenous take-up decisions within families. The estimates recovered through their approach also enable a formal specification test of the assumption that the family fixed effects estimator on its own fully solved the endogeneity problems in a given study employing that estimator. Researchers interested in conducting such a specification test are strongly encouraged to use a bootstrapped Hausman test in place of the traditional Hausman test used in Fletcher and Lehrer (2011), since neither the conventional family fixed effects estimator nor the Fletcher and Lehrer (2011) genetic lottery IV estimator is efficient under the null hypothesis of the Hausman test. In each of the applications they examine, with the traditional test Fletcher and Lehrer reject that the family fixed effects estimator does not fully solve the endogeneity problem in health when estimating its effects on academic and early labor market outcomes. The genetic lottery approach offers a new research design for researchers in the social sciences.

In summary, there is a rich and growing toolbox of genetically informed methods being developed in the epidemiological literature for estimating how outcomes of interest to social scientists are generated. Each new tool relies on a different set of identifying assumptions to support its use in causal inference. These methods are becoming increasingly available to research communities across the biomedical and social sciences and have potential applications beyond estimating causal effects including in testing the validity of the assumptions maintained when conducting conventional analysis. Genetic data can also help us understand new dimensions of individual difference, and its causes and effects.

Gene-Environment Interactions

Genetic markers can be used to explore the possibility of heterogeneous responses, both to policies and to (economic or therapeutic) treatments. Heterogeneous responses often imply an interaction of genetic and environmental influences (henceforth 'G*E') in producing outcomes. Researchers across a multitude of disciplines champion the importance of G*E effects, particularly for early childhood education. The consensus view is that it is highly unlikely that genes are destiny: environmental exposure appears to change how genes are expressed and therefore their scope for influencing outcomes.[xi]

At present, studies in the social sciences that estimate G*E effects are mainly confined to examining the effects of adding genetic interaction terms to existing research designs. A subset of these research designs exploits plausibly exogenous variation in environments; a second subset follows Fletcher and Lehrer (2011) in exploiting variation in genes within families; and a third subset is more exploratory in nature, aiming to elucidate which possible channels of influence may be most promising to investigate in further research. We discuss each of these strategies, in reverse order.

Pinning down causal G*E effects requires either exogenous variation in environmental factors or an econometric strategy that can discover breakpoints in the relationships between genetic factors and outcomes which, in turn, can be exploited for the identification of causal effects. Rosenquist et al. (2015) follow the latter approach. Motivating their study is the observation that many gene-environment interaction studies examine within-birth-cohort differences among individuals with varying environmental exposures occurring within a similar time-period of data collection. This research design, while valuable, relies upon the assumption that the environmental variation of interest is not linked with any other genetic and/or environmental factors that also affect the outcome. Using between-birth-cohort differences, on the other hand, allows for the testing of hypotheses related to time varying changes in the whole of the environment affecting the population. The authors use the longitudinal offspring sample of the Framingham Heart Study collected between 1971 and 2008. This data was collected in one small geographic area, thereby reducing any biases due to un-observables that might explain sorting across regions based on environmental conditions. They restrict the sample to be between the ages of 30 and 63 to ensure there are no differences in the age support across birth cohorts. After all, on average people get heavier as they age, and the authors do not want to mistake an age effect for a cohort effect. The main analysis applies the threshold regression estimator of Hansen (1999) to determine whether there is a structural break, of unknown timing, in the relation between genes and body mass index (BMI) using variation in both genes and outcomes across cohorts. The selected breakpoint is based on the model that best fits the data, using a grid-search algorithm.

Specifically, Rosenquist et al. (2015) test whether the well documented association between a particular SNP variant (located at rs993609) and BMI varies across birth cohorts, time period in which the data was collected, and/or the lifecycle. The analysis can be viewed as an examination of how trajectories of obesity across the life cycle vary across birth cohorts in ways that are explained by genetic inheritance. Put differently, the analysis is designed to disentangle the extent to which historical versus contemporaneous environmental factors interact with genetic features, The SNP variant studied is known as the FTO gene, first christened drily by Peters, Ansmeier, & Ruther (1999) 'the fatso gene'. This gene has been well studied.   Frayling et al. (2007) present evidence that on average, one copy of the risky variant of this SNP produces up to three and a half extra pounds of weight. Two copies of the gene lead to seven extra pounds — and increase a person's risk of becoming obese by 50 percent. Yet, there is a great deal of variation in the magnitude of this association as estimated in different studies.

The main finding of Rosenquist et al. (2015) is that there is a robust change in the relationship between the FTO risky allele and BMI across birth cohorts, with an observed inflection point for those born after 1942. This result is robust to the inclusion of family fixed effects. The threshold regression estimator allows Rosenquist et al. (2015) to statistically test for the presence of a structural break in the relationship of genes to obesity. Their result suggests that in samples containing individuals born prior to 1942, having one or two copies of the risky allele would not lead to the addition of a statistically significant amount of weight. This statement can be made with confidence based on specification tests of the unrestricted model that controls for gene*cohort, gene*time, and gene*age effects. These tests provide evidence that gene*time effects are not statistically significant once the other two sources of variation are accounted for. Only if one were to ignore gene*cohort effects would it seem that G*E effects are due to chronological timing (i.e., to events unfolding through time that affected the strength of the relationship between FTO variations and BMI, regardless of cohort or age). Upon reflection, this result is unsurprising since environments are highly correlated over the lifecycle for most individuals and so, once cohort and age effects are controlled, there is limited variation remaining in experienced environmental conditions that might affect the strength of genetic influences on outcomes.

Understanding which specific historical influences alter the penetrance of genetic variants across cohorts is not considered in the study. There are many environmental changes between birth cohorts hypothesized to be responsible for the rise in obesity including the rate of change in sedentary lifestyles, urban design, occupational shifts, dietary modifications (e.g. growth and penetrance of fast food restaurants) , and social effects among other environmental changes that have been hypothesized to be linked with obesity. The authors suggest that their finding may explain the low replication rates of the findings produced in many GWA studies, since GWA studies often pool together different datasets that are collected in different periods of time, failing to account for the possibility that genetic associations may differ across birth cohorts due to variation across cohorts in prevailing environmental factors. The authors suggest that GWA researchers may wish to counter this problem by controlling for environmental stratification in addition to the more commonly employed population stratification across the pooled datasets.

Rosenquist et al. (2015)'s control for gene*age effects may also help to explain the low replicability of GWA results concerning other outcomes. For example, Oliynyk (2018) examines how the age of the sample can influence the findings of a GWA study concerning late onset human diseases that have a large genetic component. Many common diseases, such as dementia, fit into this category and a better understanding of them is critical to policy making as the population in many developed countries continues to age. The evidence suggests that for diseases that show high cumulative incidence together with high initial heritability, samples that balance the age and birth cohorts of case and control observations may be inferior to samples that combine the youngest possible cases with the oldest possible controls, if our objective is to use these samples to gain the maximum discovery power available from GWA studies. Such studies show the importance of understanding heterogeneity in G*E effects across both age and birth cohort dimensions for improving the potency of the GWA method in detecting and correctly interpreting genetic associations.

Beyond generating methodological insights of benefit to researchers applying GWA approaches, G*E studies conducted through a social scientific lens can help determine whether policies, such as sin taxes, have different impacts on people according to their genetic predisposition to risky behaviors. If this is the case, then some policies may place a disproportionate burden on individuals with specific genetic dispositions.

Social scientists exploring G*E effects can also improve our ability to gauge human progress. An interesting example of a G*E study of this type exploits an experiment due to history. Rimfield et al. (2018) use the independence of Estonia following the fall of the Soviet Union to ask whether there was a difference in the genetic determinants (based on SNPs and polygenic scores) of educational attainment and occupation before versus after the collapse of the USSR. DNA differences are found to explain twice as much variation in educational attainment and occupational status in the post-Soviet era compared with the Soviet era. This change in the extent of genetic influence in the Estonian population is interpreted by the authors as illustrating an increased importance of the meritocratic dimension of selection into both education and occupation following the shift from a communist to a capitalist society.

Another example of a study that compares the genetic component of response to a policy across cohorts is found in Okbay et al. (2016a). These authors compare cohorts before and after a suite of schooling reforms in Sweden that, most importantly, extended mandatory schooling from seven to nine years. The authors find that the association between educational attainment and the polygenic score they constructed from their own GWA study is roughly half as large among Swedish individuals in the later cohort compared to the earlier cohort, suggesting that the Swedish reforms reduced the importance, in terms of educational attainment, of having won the ‘genetic lottery’.

Much other work by social scientists evaluating G*E effects does not explicitly consider the endogeneity of the environmental variables that are to at least some extent selected by the individual. Some of this work is more structural in nature, intending to shed new light on an underlying behavioral model. For example, Biroli (2015) integrates genetic factors into the canonical model of health production due to Grossman (1972), allowing genetic variants to differentially affect both the health production function and preferences related to the incentives surrounding health investment faced by individuals. Using data from both the Framingham Heart Study and Avon Longitudinal Study of Parents and Children, he finds evidence that genetic factors do change both the production function of BMI and the level of investments in health that are optimal for an individual.

The unbiasedness of the coefficients estimated in the empirical analysis of Biroli (2015) requires the assumption that caloric intake is an exogenous environmental factor, and not a behavioral choice. The endogeneity of environmental variables is also not considered in Hatemi’s (2013) exploration of G*E effects from proximate events such as losing a job, suffering a major financial loss, or getting a divorce on the short-term change in attitudes towards economic policy. The underlying logic is that such events should change attitudes toward policy such that they stay aligned with the maximization of self-interest, and that the degree of adjustment in their attitudes that individuals experience when such events occur may be moderated by genetic factors. The analysis presents associations that are suggestive of different responses by those with different genetic markers, with individuals who lost a job more likely to oppose policies that may have caused the change in their economic situation when they have specific genetic markers. Future work exploring G*E effects flowing from environmental shocks would clearly be more credible if the identifying variation in the environmental variables were plausibly exogenous.

Rather than exploiting variation in environmental variables, Thompson (2014) exploits within-family variation in genetic inheritance (i.e., in an individual’s draw from the genetic lottery) to explore differences in the relation between household income and children’s educational outcomes across households with different variants of the MAOA (monoamine-oxidase A) gene, located at rs1465108, that encodes an enzyme partially responsible for the metabolism of several neurotransmitters. Results indicate that the impact of income on outcomes is stronger for those with rarer variants. Conley and Rauscher (2013) advise caution on the back of implementing a research design that aims to capture G*E effects by exploiting within-family variation. They explore how genetic traits moderate the relationship between birthweight and several outcomes, including high school GPA, by exploiting birthweight differences within twins. The sole statistically significant G*E effect reported has a sign that is the opposite of what had been suggested by prior scientific research.

Credible evidence of G*E effects may help policymakers target the delivery of policies to those who would benefit most. However, as we discuss in the next section, genetic data will also pose challenges for policymakers who have been slow to develop regulatory policies related to how genetic data can be used.

Can genetic research findings inform public policy?

With a growing evidence base emerging, it is natural to ask how policymakers should leverage many of the important genetic discoveries surveyed above. Answers to this question depend upon what sort of policy is being considered. For example, as an increasing number of studies connect DNA variation with individual-specific predictors of socioeconomic status such as intelligence and personality traits, the ethical, legal and social implications of using the findings produced in scientific studies are likely to loom large.

Ding and Lehrer (2017) argue that when discussing genetics and public policy, attention should not be focused upon the question of whether a specific outcome or trait is primarily a function of genes that are immutable. As with many policies that target environmental influences, the question that policymakers must continue to ask is whether the available evidence suggests that a proposed policy would pass a conventional cost-benefit test. To illustrate this argument, they draw on an example in Goldberger (1979) that clearly points out that there is an ethically defensible role for public policy when a problem has its root in genetic factors. Goldberger uses the example of poor eyesight: even if poor eyesight were strictly a result of genetic inheritance, policymakers could provide glasses to those afflicted. Not doing so would be inefficient and even arguably less ethically sound than doing so.

Genetic factors, while themselves immutable, offer policymakers the ability to personalize how policies are delivered to individuals. This personalization also opens a new set of challenges to designing effective policy, since issues related to privacy and discriminatory treatment based on genetic characteristics may emerge. Any cost-benefit test applied to the use of genetic markers in designing public policy should weigh the consequences of issuing a broad mandate (a one-size-fits-all policy approach) versus targeting policy to those with specific characteristics.

There is perhaps no area where genetic data is currently playing a larger role than personalized medicine and the pharmaceutical industry. Pharmacogenomic tests can already identify whether a breast cancer patient will respond to the drug Herceptin, whether an AIDS patient should take the drug Abacavir, or what the correct dose of the blood-thinner Warfarin should be for a person with a specific genetic marker profile. Proponents of using genetic data in health policy suggest that by tailoring recommendations to each person’s DNA, health care professionals will be able to work with individuals to focus efforts on the specific strategies — from diet to high-tech medical surveillance — that are most likely to maintain health for each individual.

The potential of precision medicines that create a better match between patients and medications is tantalizing for policymakers. In early 2015, the White House announced a ‘bold new research effort to revolutionize how we improve health and treat disease,’ and launched a Precision Medicine Initiative with a $215 million initial investment in 2016. The United States is not alone in its interest in this space. Other countries, including France and China, have recently announced major public investments ranging from the equivalent of several hundreds of millions to several billions of U.S. dollars over the coming years. While this funding promotes innovation policy in the biotechnology and pharmaceutical industries, Stern, Alexander, & Chandra (2017) point out that it also changes many of the economic incentives that pharmaceutical manufacturers face in the drug development process. These changes may have important unintended consequences.

First, studies aiming to develop precision medicine approaches will mechanically target smaller patient populations than more traditional approaches to health research. The products discovered are likely to include those with large expected clinical benefits to small patient populations selected on specific genetic dimensions, with these benefits unlikely to be as significant, or present at all, in the population at large. For example, in January 2018, the FDA approved Luxturna to treat Leber congenital amaurosis (LCA), an inherited eye disorder. Targeting a variant of the RPE65 gene, this is the first gene therapy to gain FDA approval, and the one-off treatment is priced in the United States at slightly over $400 000 per eye. This high price arises since the marginal customer is expected to have a greater willingness to pay for Luxturna than for a more conventional therapy, as the drug has been proven to be more efficacious than conventional therapies within a smaller patient population. The higher price also helps to justify the fixed costs of drug development. The increase in funding for personalized medicine may cause drug manufacturers to shift their attention to subsets of products that are effective for smaller populations and are able to command high(er) prices, ceteris paribus.

Drug price and market size are naturally related. For example, Acemoglu and Linn (2004), Kyle and McGahan (2012) and Dubois, de Mouzon, Morton, & Seabright, (2015) each present evidence that market size for a particular drug is associated with the number of firms conducting research in the area of that drug. Further, targeting patient sub-populations with particular biomarkers often allows manufacturers to more easily qualify for an ‘orphan drug’ designation through the Orphan Drug Act of 1983. To receive this designation, a company must argue that it is focusing on developing a therapy for a disease sub-population of fewer than 200 000 patients. Being awarded this designation has been found to deliver powerful financial incentives for pharmaceutical firms. If the FDA approves a new molecular entity (i.e., a drug) to treat an ‘orphan condition’, the innovating firm receives tax credits equaling 50 per cent of clinical trial expenses and an extra two years of marketing exclusivity. These benefits appear to be enticing: in 2015, 47 per cent of new drugs approved by the FDA were orphan drugs.

Not only does this policy environment for orphan drugs support higher prices for longer, but since the market for precision medicines is small, price competition through follow-on entry (i.e., by firms producing generic or biosimilar drugs) may not develop. Existing evidence from the European Union in Morton, Stern, & Stern et al. (2018) and Berndt and Trusheim (2015) shows that even after the exclusivity periods granted to orphan drugs end, there may not be a large enough market to stimulate the development of biosimilar follow-on drugs, weakening the potential for price competition. The consequences of a decline in price competition for health treatments will arguably be more severe in single-payer health systems that face a balanced budget requirement, in which the increased funding that is required to provide precision medical treatments is often financed via reductions in government expenditures on other programs.

Perhaps of greater concern is that biomarkers may lead pharmaceutical companies to engage in genetically informed price discrimination. For example, a second gene therapy drug approved by the FDA in May 2018 is known as Kymriah. The pharmaceutical company that developed the drug is following a practice that they term ‘indication-based pricing.’ This means that the price varies according to the therapeutic application. In this case, for those with large B-cell lymphoma, the price is $373 000 for the one-time treatment, while the price for using the treatment to treat pediatric leukemia is $475 000. Identifiable biomarkers which include but are not limited to specific SNP variants can become an important tool for facilitating price discrimination, as they can be used to segment the drug market into identifiable subgroups that differ based on not only the expected efficacy of the product, but (and because of that) by willingness to pay for the product.

Whether the price of a drug should be aligned to its clinical value in each approved indication is a profit-maximizing strategy that proponents suggest helps ensure social benefits firm drug developments. However, it could increase health spending which society would have to cover either through taxes or insurance costs. Beyond pricing, the FDA noted in 2016 that a shift in focus towards precision medicines will likely result in targeting only a subset of genetic markers. Specifically, the FDA-NIH (2016) speculate that attention will be paid to developing drugs that are tied to markers that are either predictive of therapeutic sensitivity or could be used for diagnosis and prognosis. They write that markers can be 'used to identify individuals who are more likely than similar individuals without the biomarker to experience a favorable or unfavorable effect from exposure to a medical product'. This could shift attention away from markers that were not previously investigated for association perhaps due to being omitted due to linkage disequilibrium or not being measured in the microarray, as well as from developing drugs that have small genetic components.

On a more positive note, Budish, Roin, & Williams (2015) suggest that increases in the speed of clinical trials, for example due to larger expected effect sizes, may provide an incentive for pharmaceutical manufacturers to target drugs for different conditions, thus potentially bringing more innovations to the market. This could counteract some of the negative incentives from creating drugs that target smaller patient populations. Chandra, Garthwaite, & Stern (2018) provide evidence that genetic markers are associated with the length of clinical trials for cancer drugs. Precision based medicines are found to be developed from trials that are on average six to seven months shorter in duration than those designed to test non-precision medicines. The authors speculate that this decrease in trial duration occurs since a therapeutic effect is easier to detect due to the greater putative efficacy of genetically targeted drugs in the targeted subpopulation.

Despite the great potential of developing precision treatments, the above examples show that this new direction in medical research is not a free lunch. There are unintended consequences of shifting the attention of pharmaceutical companies towards precision medicines newly conceivable in part based on scientific findings from molecular genetic discoveries. These consequences relate to companies' decisions about which therapies to develop, how to price new drugs, and how to design and implement clinical trials. More policy attention is needed to reduce the negative consequences that may develop in response to the growing body of evidence documenting the genetic determinants of health and disease outcomes.

A second important policy area relates to the possible misuse of genetic data. This includes not just the potential promotion of eugenics-style initiatives, but also the potential for genetic data on inherited predispositions to influence decisions in the workplace or in relation to insurance coverage. For example, genetic data and research findings based on it may lead to differential treatment, or genetic discrimination, by health insurers. The most obvious possibility in this space would be an insurer's refusal to give coverage to an individual who has a genetic variation that raises his odds of developing a specific health disorder.

In the United States, the 2008 Genetic Information Nondiscrimination Act (GINA) attempts to address the need to regulate how genetic information is used, most notably protecting against discrimination in health insurance provision and employment. However, GINA does not apply to either life insurance or long-term care insurance, or to employers of fewer than 15 employees. More challenging is that GINA places the burden on victims of genetic discrimination to prove that their information was misused.

Last, policy that more carefully considers privacy is likely needed to regulate how genetic testing is undertaken. Two important facets of privacy relate to the scope of parties to whom and context in which genetic test results should be returned, and to the question of whether direct-to-consumer genealogic testing companies should be required to undertake a subject verification process before proceeding with testing. On the latter, there are currently few hurdles (if any) to companies proceeding with testing once a request has been made and money has changed hands. It would be quite possible to send someone else's tissue sample for testing and receive a full report on that person's genetic profile. Variants of this idea appear in numerous TV shows and films in scenes where characters try to creatively obtain DNA information by analyzing materials a person may have touched. One popular example on detective shows is collecting a water bottle from a suspect who was being interrogated.

Results from genetic tests are often difficult to interpret. Many individual genetic factors have very small effects, and arming people with knowledge of these predispositions without providing the appropriate context and qualifiers could worsen outcomes. Adverse reactions by individuals to poorly communicated test results may lead to unintended consequences, e.g., due to an over-response to the presence of a genetic predisposition (which is merely correlational), possibly leading to patient-demanded medical care that may be unnecessary and ineffective. As an example considering social/education policy, suppose a parent has knowledge of polygenic scores for educational attainment of their two children. We assume that there is a difference in their genetic scores such as 35%. As discussed above, the available evidence suggests that the effects of each risk allele used in the calculation of the polygenic score is very small in magnitude. However, not understanding the small effect that corresponds to weeks and not years of education, may encourage the parent to make investment decisions on how to assign inputs (tutoring or time helping with homework) that can accentuate rather than mitigate this genetic difference.

On a more positive note, consider the following example. Suppose that genetic screening can reliably predict complex learning disorders that are a function of many genes, each with a very small effect. If a single polygenic score is calculated from an ensemble of markers that have well validated significant (if individually small) effects, this score can be interpreted as a measure of an individual's risk for a specific disorder or trait, which, in many situations, may take psychologists years to diagnose. Armed with knowledge of whether their child or employee is at an elevated risk for poor learning outcomes, parents and employers will be able to make different investments years prior to receiving a formal diagnosis through conventional means. Since these investments may affect how the underlying genes are expressed and thereby alter the risk of observing the outcome, regulations that effectively support this type of communication and limit unintended consequences may assist in raising welfare.

Ding and Lehrer (2017) point out that beginning in April 2017, the US Food and Drug Administration allowed the genetic testing firm 23andMe to sell reports with qualifiers showing customers whether they have an increased genetic risk of developing certain diseases and conditions. The number of conditions is limited, and this policy reversed a decision in 2013 that forced 23andMe to stop communicating the results of health-related traits. The authors also suggest that the Stanford Cancer Institute's decision support information web- based interface, available at http://brcatool.stanford.edu/, is an example of an effective mechanism that communicates sensitive personal information with appropriate safeguards. Regulations may be needed in these areas, particularly in light of our current limited understanding of how genetic markers operate.

Developing regulations regarding how test results are returned, to whom they are returned, and for what purposes they are returned may also aid social scientists wishing to collect molecular genetic data from participants in research studies. Higher compliance rates may result if participants are given information at point of consent about the purpose of data collection and how the data will be shared within the research community. Without such disclosure, the use of molecular genetic data by other researchers for reasons that were not apparent at the time of data collection – something quite likely given the data sharing infrastructure in this area – may constitute a violation of individual privacy, on top of the usual data security concerns. As an example of this issue, consider how law enforcement in Sacramento, California were able to track down and arrest the "Golden State Killer", Joseph James DeAngelo. The lead investigator submitted DNA collected years ago from one of the crime scenes to an open source genealogy website called GEDmatch, thereby narrowing the field to a small pool of potential suspects. Since the site is open source, no court records were needed to access the DNA records on the GEDmatch web site, but privacy advocates argue that most individuals who submit their DNA to these companies are unaware that they are effectively sharing their DNA with law enforcement.

In summary, policy making in the brave new world of genetic information does not require a wholesale transformation of how policies are developed. The speed at which concerns regarding genetic data can be effectively integrated into policy design is tied partly to improvements in scientific understanding of how genetic markers operate, but even more strongly to the speed with which this developing knowledge is conveyed to stakeholders so that a social consensus on optimal policy can emerge.

Conclusions and Future Directions

Heritability plays a role in generating nearly every socioeconomic and health outcome. This feature has long been ignored by social scientists due in part to data availability, and by policymakers who often fall victim to thinking that the fixity of one's genetic code at conception implies that there is nothing we can do to improve outcomes, even if we knew an individual's complete genetic code. However, heredity is not destiny, and much work is needed to clarify what is meant by a genetic predisposition and what policy levers are revealed by knowing in which people such predispositions lie. Social scientists can contribute to this work by translating the revolutionary advances in genetics and genomics to reach both policy audiences and the broader academic community. Great care must be taken in these translations to elucidate the assumptions imposed in the underlying analyses, to ensure that our developing knowledge is used appropriately to develop effective policy.

Genes influence not only health and disease, but also human traits and behaviors. Science is only beginning to unravel the complicated pathways leading from genes through the environment to outcomes, and there are numerous avenues via which social scientists can enter this area to generate new insights.

Much of the current research in the social sciences that uses genetic data draws heavily on the literature from molecular genetics and other non-social sciences. Knowledge and protocols from the social sciences can assist in expanding the evidence base. For example, researchers in population studies and sister fields have substantial experience with data collection and issues related to pooling data from different sources. Much of the current literature in genetics reviewed above does not consider sampling issues or explicitly discuss the external validity of findings. Social scientists familiar with data manipulation and imputation would also be well placed to advise upon the consequences of using imputed versus actual SNP data, and other matters relating to data quality. Further, methodologists may be able to develop methods to improve the efficiency of estimation based on data collected in alternative manners.

From a more theoretical perspective, many scientific studies that draw on genetic data are silent on the topic of underlying behavioral models, yet many outcomes including those in health and education are likely a function of a sequence of individual decisions, genetic factors, and the interactions of the two. For example, conventional GWA protocols ignore the interaction of the environment with genetic factors, frequently assuming linear effects on outcomes of the simple count of alleles. The calculation of polygenic scores from GWA estimates and the subsequent use of these scores in outcome prediction ignore issues related to uncertainty and estimation error. By incorporating an underlying behavioral model, researchers could be explicit about the assumptions being imposed in such exercises, and the evidence from integrating genetic markers into an existing conventional analysis could then be used to further refine the behavioral model. Work in this direction holds the potential to advance our understanding of genetic mechanisms in a logically consistent and statistically valid framework.

There is also considerable scope for methodological developments driven by social scientists to take the analysis of genetic data far beyond the use of off-the-shelf software. For example, applying new econometric tools to uncover and understand heterogeneity in genetic effects holds much promise. These tools can draw from the expanding literature on treatment effect heterogeneity and may be quite useful in particular for G*E analyses. Lehrer (2016) also points out that there may be a serious identification challenge in current G*E analysis wherein the same data is used to describe both situations where exposure to an environmental factor that predicts a behavior is conditional upon a person's genotype, and situations when the genotype's direct effect on the behavior is moderated by some environmental effect. For example, suppose that a gene affects a risky health behavior that is also cue-conditioned (i.e., faced with a given environment, an individual is more likely to engage in the behavior – e.g., smoking in a night club is more likely than in other locations). It may be the case that in addition to interacting with smoking to produce disease, the gene also directly leads those who possess it to visit night clubs. While statistically separating these pathways is desirable, better communication protocols are also required to help policy audiences understand what is being identified in any given analysis. Lehrer (2016) suggests that researchers use the terminology 'G*E responses' to refer to situations where exposure to an environmental factor that in turn predicts behavior is conditional upon a person's genotype, and the term 'G*E modifications' to refer to differential genetic reactions to environmental factors. Personalized medicine and many policies that would target individuals by genotype may be best guided by information about G*E modifications, whereas G*E responses may be more interesting for researchers to study if they are interested in understanding the underlying behavioral reasons for observed heterogeneity in estimated environmental effects on outcomes across the population.

With improved methodological tools, more credible evidence from rigorous G*E studies may lead to the reshaping of social science theories. Many policies and programs have been observed to have heterogeneous effects on individuals with different demographic and socioeconomic characteristics that are consistent with an underlying theory. For example, with information on genetic markers that associate with addiction, there is the possibility to better understand why changes in sin taxes affect decisions on the intensive extensive margins of substance use differently for different individuals.

Rather than characterizing individuals as simply being ‘rational’ or ‘impulsive’ or ‘behavioral’, researchers may be able to pinpoint individual biological characteristics that can explain the underlying heterogeneity in choice behavior. Similarly, as Biroli (2015) illustrates, one can ask whether heterogeneity due to genetic inheritance affects calories burnt and/or calories consumed, thereby helping to shape future theories about the development of obesity by explaining why behavioral heterogeneity may arise. Many simple economic models predict treatment effect heterogeneity based on individual characteristics (see Lehrer et al. (2016) for an example, in the form of a static labor supply model) and treat genetic influence as predetermined at conception. Future theoretical work could relax these assumptions by exploring whether arming individuals with knowledge of their genetic make-up, e.g., through receiving results from genetic testing companies, shapes decisions in various realms, including insurance coverage up-take or risky behaviors. The existence of direct-to-consumer genetic testing introduces a new source of an information asymmetry, known to the individual but unknown to insurers, that may affect individual decisions and thereby expand the scope of treatment effect heterogeneity.

Beyond statistical issues and theoretical modelling, perhaps the area where social scientists’ increased involvement with genetic data may prove to be most valuable relates to knowledge translation activities for the benefit of policy audiences. We argue that the social benefits of using genetic information are tied to how that information is communicated. Findings from research that is well designed and robust should be clearly communicated in a way that neither oversimplifies nor overstates the role of genetic factors. The problematic consequences of inadequate communication are well known. Findings from previous research in behavioral genetics have not always been well communicated, with the unfortunate example of the analysis of Herrnstein and Murray, whose book The Bell Curve (1994) led to significant controversy. Given this history and the real potential for recurrence as new findings emerge from genetic studies, it is of the utmost importance not only to gather sufficient scientifically valid information about the genetic factors underpinning outcomes, yielding more definitive scientific insight, but to communicate these insights in a way that avoids misunderstanding and stigmatization when considering the implications of such research for individuals and society.

References

1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature*, 526, 68–74.

Acemoglu, D., & Linn, J. 2004. Market size in innovation: Theory and evidence from the pharmaceutical industry. The Q*uarterly Journal of Economics,* 119(3), 1049-1090.

Almond, D., Chay, K. Y., & and Lee, D. S. (2005). The Costs of Low Birth Weight. *The Quarterly Journal of Economics*, 120(3), 1031-1083.

Angrist, J. D., & Pischke, J-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives,* 2(1), 3–30.

Barth, D. J., Papageorge, N., & Thom, K. (2018). Genetic endowments and wealth inequality. *NBER Working paper w24642.*

Behrman, J. R. (2016). In: Komlos, J. & Rashad, I. (eds.) Twin studies in economics in the *Oxford handbook of economics and human biology*. Oxford University Press, p. 385-404.

Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica,* 80(6), 2369-2429.

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(5), 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 Pediatric and Adolescent Medicine*, 166(6), 515–521.

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

Benjamin, D. J., Chabris, C. F., Glaeser, E. L., Gudnason, V., Harris, T. B., Laibson, D. I., Launer, L., & Purcell, S. (2007). Genoeconomics. In: Weinstein M, Vaupel JW, Wachter KW (eds) *Biosocial Surveys, Committee on population, division of behavioral and social sciences and education*. The National Academies Press, Washington

Berndt, E. R., & Trusheim, M. R. (2015). Biosimilar and biobetter scenarios for the US and Europe: What should we expect? In *Biobetters*, pp. 315-360. Springer New York, 2015.

Biroli, P. (2015). Genetic and economic interaction in the formation of human capital: the case of obesity. Mimeo. University of Zurich

Bowden, J., Davey Smith, G., & Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology,* 44(2), 512–525.

Battelle Technology Partnership Practice (2011). The U.S. Biopharmaceuticals Sector: Economic Contribution to the Nation available at http://phrma-docs.phrma.org/sites/default/files/pdf/2011_battelle_report_on_economic_impact.pdf

Brickell, I., Larsson, H., Lu, Y., Pettersson, E., Chen, Q., Kuja-Halkola, R., Karlsson, R., Lashey, B. B., Lichenstein, P., & Maartin, J. (2018). The contribution of common genetic risk variants for ADHD to a general factor of childhood psychopathology, *Molecular Psychiatry,* in press, https://doi: 10.1038/s41380-018-0109-2.

Budish, E., Roin, B. N., & Williams, H. (2015). Do firms underinvest in long-term research? Evidence from cancer clinical trials. *The American Economic Review,* 105(7), 2044-2085.

Chabris, C. F., Lee, J. J., Benjamin, D. J., Beauchamp, J. P., Glaeser, E. L., Borst, G., Pinker, S., & Laibson, D. I. (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.

Chandra, A., Garthwaite, C., & Stern, A.D. (2018). Characterizing the Drug Development Pipeline for Precision Medicines forthcoming *in Economic Dimensions of Personalized and Precision Medicine*, E. Berndt, D. Goldman & J. Rowe, (eds.) University of Chicago Press.

Conley D., & Zhang, S. (2018). The promise of genes for understanding cause and effect. *Proceedings of the National Academy of Sciences,* 115(2), 5626–5628. https://doi.org/10.1073/pnas.1805585115.

Conley, D., & Rauscher, E. (2013). Genetic interactions with prenatal social environment: effects on academic and behavioral outcomes. *Journal of Health and Social Behavior,* 54(1), 109–127. DOI: 10.1177/0022146512473758.

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. DOI: 10.1080/19485560903415807

Conley, T. G., Hansen, C. B., & Rossi, P. E. (2012). Plausibly exogenous. *Review of Economics and Statistics,* 94(2), 260–272.

Currie, J., & Thomas, D. (1995). Does Head Start Make a Difference? American Economic *Review*, 85(3), 341-364.

Deming, D. (2009). Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start." *American Economic Journal: Applied Economics*, 1(3), 111-134.

Ding, W. & Lehrer, S. F. (2017), What is the role for molecular genetic data in public policy*? IZA World of Labor* 395. Available at https://wol.iza.org/articles/what-is-the-role-for-molecular-genetic-data-in-public-policy/long.

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), 578–597.

Ding, W., Lehrer, S.F., Rosenquist, N.J., & Audrain-McGovern, J. (2006). The impact of poor health on education: new evidence using genetic markers, *National Bureau of Economic Research Working Paper Series No. 12304.*

Dreber, A., Apicella, C. L., Eisenberg, D. T. A., Garcia, J. R., Zamore, R. S., Lum, J. K., & Campbell, B. (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.

Dubois, P., de Mouzon, O., Morton, F. M. S., & Seabright, P. (2015). Market size and pharmaceutical innovation. *The RAND Journal of Economics,* 46(4), 844-871.

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLoS Genetics* https://doi.org/10.1371/journal.pgen.1003348

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal,* 315, 629–634.

FDA-NIH Biomarker Working Group. (2016). BEST (Biomarkers, EndpointS, and other Tools) Resource. Silver Spring (MD): Food and Drug Administration (US); Bethesda (MD): National Institutes of Health. Available at https://www.ncbi.nlm.nih.gov/books/NBK326791/.

Figlio, D., Guryan, J., Karbownik, K., & Roth, J. (2014). The Effects of Poor Neonatal Health on Children's Cognitive Development. *American Economic Review*, 104(12), 3921- 3955.

Fletcher, J. M., & Lehrer, S.F (2011). Genetic lotteries within families. *Journal of Health Economics*, 30(4), 647–659.

Fletcher, J. M., & 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., & 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, https://doi.org/10.2202/1558-9544.1180.

Frayling, T. M., Timpson, N. J., Weedon, M. N., Zeggini, E., Freathy, R. M., …, Hattersley, A. T. & McCarthy, M. I. (2007). A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science,* 316(5826), 889-894.

Gage, S. H., Jones H. J., Burgess, S., Bowden, J., Davey-Smith, G., Zammit, S. & Munafò M. R. (2017). Assessing causality in associations between cannabis use and schizophrenia risk: a two- sample Mendelian randomization study. *Psychological Med*icine, 47(5), 971–980.

Goldberger, A. S. (1979). Heritability. *Economica* 46(184), 327–347.

Greiner, J., & Rubin, D. (2011). Causal effects of perceived immutable characteristics. *The Review of Economics and Statistics,* 93(3), 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(2), 345–368.

Hatemi, P. K. (2013). The influence of major life events on economic attitudes in a world of gene-environment interplay. *American Journal of Political Science*, 57(4), 987–1000.

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.

Hemani, G., Bowden, J., & Davey Smith, G. (2018). Evaluating the potential role of pleiotropy in Mendelian randomization studies, *Human Molecular Genetics*, in press, https://doi.org/10.1093/hmg/ddy163.

Herrnstein, R. J., & Murray, C. (1994). *The bell curve*. The Free Press, New York

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

Huber, E., Donnelly, P. M., Rokem, A., & Yeatman, J. D. (2018) Rapid and widespread white matter plasticity during an intensive reading intervention. *Nature Communications,* 9, 2260, doi:10.1038/s41467-018-04627-5.

Imbens, G. W., & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica,* 62(2), 467-475.

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

Keane, M. (2010). A structural perspective on the experimentalist school. *Journal of Economic Perspectives*, 24(1), 47-58.

Kyle, M. K., & McGahan, A. M. (2012). Investments in pharmaceuticals before and after TRIPS. *Review of Economics and Statistics*, 94(4), 1157-1172.

Lee, J. J., Wedow, R., Okbay, A., Kong, E., ..., Turley, P., Visscher, P. M., Benjamin, D. J. & Cesarini, D. (2018). Gene discovery and polygenic prediction from a 1.1-million-person GWAS of educational attainment. *Nature Genetics*, in press, https://doi: 10.1038/s41588-018-0147-3.

Lehrer, S. F., & Ding, W. (2017). Are genetic markers of interest for economic research? *IZA Journal of Labor Policy* 6:2, https://doi.org/10.1186/s40173-017-0080-6.

Lehrer, S. F., & Xie, T. (2017). Box office buzz: Does social media data steal the show from model uncertainty when forecasting for Hollywood? *Review of Economics and Statistics*, 99(5), 749-755.

Lehrer, S. F. (2016). In: Komlos J, Rashad I (eds) Biomarkers as inputs in the Oxford handbook of economics and human biology. Oxford University Press, 339-365

Lehrer, S. F., Pohl, V. R., & Song, K. (2016). Targeting Policies: Multiple Testing and Distributional Treatment Effects, *National Bureau of Economic Research Working Paper Series No. 22950.*

Mendel, G. J. (1866). Versuche über Pflanzen-Hybriden [Experiments Concerning Plant Hybrids]. In *Verhandlungen des naturforschenden Vereines in Brünn* [Proceedings of the Natural History Society of Brünn] IV (1865): 3–47.

Mukherjee, S. (2017). *The Gene: An Intimate History,* Scribner Press.

Okbay, A., Beauchamp, J. P., Fontana, M. A., Lee, J. J., Pers, T. H., Rietveld, C. A., Turley, P., …, Visscher, P. M., Esko, T., Koellinger, P. D., Cesarini, D., & Benjamin, D. J. (2016). Genome-wide association study identifies 74 loci associated with educational attainment. *Nature,* 533, 539–542, https://doi:10.1038/nature17671.

Oliynyk, R. T. (2018). Age-related late-onset disease heritability patterns and implications for genome-wide association studies, *biorxiv*, https://doi.org/10.1101/349019.

Papageorge, N. W., & Thom, K. (2017). Genes, Education, and Labor Market Outcomes: Evidence from the Health and Retirement Study. *IZA Discussion Paper dp10200*

Peters, T., Ansmeier, K., & Ruther, U. (1999). Cloning of Fatso (Fto), a novel gene deleted by the Fused toes (Ft) mouse mutation. *Mammalian Genome,* 10(10), 983-986.

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.

Rees, J. M. B., Wood, A. M., & Burgess, S. (2017). Extending the MR- Egger method for multivariable Mendelian randomization to correct for both measured and unmeasured pleiotropy. *Statistics in Medicine,* 36, 4705–4718.

Rietveld, C. A., Esko, T., Davies, G., Pers, T. H., Turley, P. A., Beben, B., Chabris, C. F., Emilsson, V., Johnson, A. D., Lee, J. J., de Leeuw, C., Marioni, R. E., Medland, S. E., Miller, M. B., Rostapshova, O., Van der Lee, S. J., Vinkhuyzen, A. A. E., Amin, N., Dalton, C., Derringer, J., van Duijn, C. M., Fehrmann, R., Franke, L., Glaeser, E. L., Hansell, N. K., Hayward, C., Iacono, W. G., Ibrahim-Verbaas C. A., Jaddoe, V., Karjalainen J, Laibson, D., Lichtenstein, P., Liewald, D. C., Magnusson, P. K. E., Martin, N. G., McGue, M., McMahon, G., Pedersen, N. L., Pinker, S., Porteous, D. J., Posthuma, D., Rivadeneira, F., Smith, B. H., Starr, J. M., Tiemeier, H., Timpson, N. J., Trzaskowski, M., Uitterlinden, A. G., Verhulst, F. C., Ward ME, Wright, M. J., Smith, G. D., Deary, I.J., Johannesson, M., Plomin, R., Visscher, P. M., Benjamin, D. J., Cesarini, D., & Koellinger P. D. (2014). Common genetic variants associated with cognitive performance identified using proxy-phenotype method. *Proceedings of the National Academy of Sciences* 111(38), 13790–13794, https://doi: 10.1073/pnas.1404623111.

Rietveld, C. A., Medland, S. E., Derringer, J., Yang, J., Esko, T., Martin, N. W., Westra, H. J., Shakhbazov, K., …, Conley, D., Davey-Smith, G., Franke, L., Groenen, P. J. F., Hofman, A., Johannesson, M., Kardia, S. L. R., Krueger, R. F., Laibson, D., Martin, N. G., Meyer, M. N., Posthuma, D., Thurik, A. R., Timpson, N. J., Uitterlinden, A. G., van Duijn, C. M., Visscher. P. M., Benjamin, D. J., Cesarini, D. & Koellinger, P. D. (2013). GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science*, 340(6139), 1467–1471, https://doi:10.1126/science.1235488.

Rimfield, K., Krapohl, E., Trzaskowski, M., Coleman, J. R. I., Selzam, S., Dale, P. S., Esko, T., Metspalu, A., & Plomin, R. (2018). Genetic influence on social outcomes during and after the Soviet era in Estonia, *Nature Human Behaviour,* 2, 269–275.

Rosenquist, J.N., Lehrer, S. F., Malley, A. J. O., Zaslavsky, A. M., Smoller, J. W., & 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.

Sacerdote, B. (2007). How Large are the effects from changes in family environment? A study of Korean American adoptees. *The Quarterly Journal of Economics*, 122(1), 119–157.

Scott Morton, F. M., Stern, A. D. & Stern, S. (2018) "The impact of the entry of biosimilars: evidence from Europe." *Forthcoming in the Review of Industrial Organization*.

Schmutz, J., Wheeler, J., Grimwood, J.,Dickson, M., Yang, J., Caoile, C., Bajorek, E., Black, S., Chan, Y. M., Denys, M., Escobar, J., Flowers, D., Fotopulos, D., Garcia, C., Gomez, M., Gonzales, E., Haydu, L., Lopez, F., Ramirez, L., Retterer, J., Rodriguez, A., Rogers, S., Salazar, A., Tsai, M., & Myers, R. M. *(2004).* Quality assessment of the human genome sequence*. Nature, 429(6990): 365–368.*

Sims, C. A. (2010). But economics is not an experimental science. *Journal of Economic Perspectives,* 24(1), 47-58.

Smith, G. D., & Ebrahim, S. (2003). 'Mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology,* 32(1), 1–22.

Stern, A. D., Alexander, B. M., & Chandra, A. (2017). Innovation incentives and biomarkers. *Clinical Pharmacology & Therapeutics*, 103(1), 34-36. doi: 10.1002/cpt.876. Epub 2017 Oct 16.

Taubman, P. (1976). The determinants of earnings: genetics, family, and other environments: a study of white male twins. *American Economic Review* 66(5), 858–870.

Thompson, O. (2014.) Economic background and educational attainment the role of gene-environment interactions. *Journal of Human Resources*, 49(2), 263–294.

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

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), https://doi.org/10.1371/journal.pone.0008516.

Glossary

Allele: Allele is used to describe variant forms (i.e. where the base pairs differ) of a given gene. Each person inherits two alleles for each gene, one from each parent. If the two alleles are the same, the individual is homozygous for that gene. If the alleles are different, the individual is heterozygous.

Amino Acid: Amino acids are a set of 20 different molecules used to build protein. In exome sequencing, microarrays provide information on the amino acid sequence of proteins and by using this information one can learn the sequence of genes.

Base: Each unit of DNA is made up of one of four different bases (Adenine (A), Cytosine (C), Thymine (T), and Guanine (G)) that is attached to sugar (deoxyribose).

Base Pair: The strands of DNA inherited from each parent are joined together in a specific manner (A base pairs with T and C base pairs with G) by hydrogen bonds. Knowledge of one side of DNA gives information of the base on the other side.

Chromosome: The chromosome located in the nucleus of each cell is how DNA is stored. Humans have 23 pairs of chromosome and the length of each of these strands of DNA varies between 48 million to 250 million bases.

Codon: A codon is a trinucleotide sequence of DNA that corresponds to a specific amino acid. The genetic code describes the relationship between the sequence of DNA bases (A, C, G, and T) in a gene and the corresponding protein sequence that it encodes.

DNA: DNA stands for deoxyribonucleic acid and it is made up of the four bases. The DNA molecule consists of two strands that wind around one another to form a shape known as a double helix. The sequence of the bases in DNA influence how proteins are assembled and numerous outcomes.

DNA sequencing: Techniques used by laboratories to determine the exact sequence of bases (A, C, G, and T) in a DNA molecule.

Double Helix: DNA is made up of two strands that are twisted together in a shape that is known as the double helix. The structure was discovered by Watson and Crick (1953).

Exome: The genome can be simplified into two components: parts that code for protein and parts that don’t. The part that codes for protein (~2% of the total genome), is also known as the exome.

Exon: Just like the genome is broken into parts that code for protein (exome) and the parts that don’t, genes are broken into parts that code for protein and parts that don’t. Exon refers to the part of a gene that codes for amino acids that subsequently combine to make proteins.

Gene: A gene is a piece of DNA that generally varies between a few hundred base pairs to many thousand base pairs. Genes are arranged on the chromosome and provide instructions to build proteins.

Genetic Map: This provides the relative locations of genes on each chromosome. The map uses the concept of linkage disequilibrium since the closer two genes are to each other on the chromosome, the greater the probability that they will be inherited together.

Genetic Marker: A DNA sequence with a known physical location on a chromosome.

Genome: In humans, the genome consists of 23 pairs of chromosomes, found in the nucleus, as well as a small chromosome found in the cells' mitochondria. Each set of 23 chromosomes contains approximately 3.2 billion bases of DNA sequence.

Genotype: A term used to refer to the two alleles inherited for a particular gene. This is the version of a DNA sequence an individual has.

HapMap: A map describing common patterns of genetic variation among individuals. It provides information on the location of DNA variations from combinations of alleles or to a set of single nucleotide polymorphisms (SNPs) on each chromosome.

Imputation: A statistical method to predict which of the four bases (A, C, G or T) is located at a position that was not sequenced by using information on bases located close by on the chromosome.

Linkage: This provides information on how associated DNA sequences are on the same chromosome. Two genes that tend to be transmitted together we say are linked to each other.

Nucleotide: A building block of RNA and DNA. The bases used in DNA adenine (A), cytosine (C), guanine (G), and thymine (T).

Polygenic Trait: A characteristic or outcome that is affected by many, many different genes

Protein: Proteins are complex molecules involved in many critical functions of the body ranging from the production of antibodies to the transportation of substances, structure and sending messages. Each protein is composed of a chain of amino acids.

Polymorphism: A term indicating the location of variation in a DNA sequence across individuals.

Sequencing: The process of reading the bases in DNA. Whole genome sequencing is comprehensive, whereas other methods may only sequence a few bases. The sequence of bases along DNA provides instructions to assemble protein and RNA molecules.

Single Nucleotide Polymorphisms (SNPs): A polymorphism involving variation in a single base pair.

Variant: A single difference in the DNA between two people and is also known as single nucleotide polymorphism (SNP) and on occasion allele.

Whole Genome Sequencing (WGS): The process of reading every single of the 3.2 billion bases in the DNA of an individual.

Endnotes

[i] The remaining 98 per cent of the human genome is often referred to as 'non-coding DNA'. While it does support a large variety of functions that are crucial to the survival of an organism (e.g., regulating when proteins are made, and controlling the packaging of DNA within the cell), the exact role of this remaining 98 per cent remains less understood than that of the two per cent of our DNA that encodes protein-production instructions.

[ii] The size of the human genome is quite large relative to the genome for either *E. coli* (a bacterium that lives in the human gut) and a fruit fly that are, respectively, approximately five million and 123 million base pairs in length. However, the human genome is much shorter than the genome for other living things, such as the loblolly pine tree, which is roughly 23 billion base pairs in length.

[iii] A draft of the entire human genome sequence was first made available in 2001, but it was only finalized on April 14, 2003. A major quality assessment in Schmutz et al. (2004) of the human genome sequence was published on May 27, 2004, indicating that in over 92% of samples taken, the sequence exceeded 99.99% accuracy, which was within the intended goal. A 'Gold Standard' version of the human genome sequence excluding one chromosome was then released in October 2004. The full sequence of the last chromosome was published in the journal *Nature* in May 2006.

[iv] Indeed, at present the most common routine for sequencing an individual human's genome involves generating a 'draft' sequence for the tested individual and comparing it to the representative human genome, viewed as a 'reference' human genome sequence.

[v] An important but often overlooked weakness of much of the off-the-shelf GWA software available today is that its measurement of the variation in each SNP is strictly in the form of counts of the number of risky alleles.

[vi] Note that in most of the GWA reported in Lee et al. (2018) each specification included up to 10 000 SNPs if data from 23andMe was utilized.

[vii] See Angrist and Pischke (2010) for more discussion, and comments by Keane (2010) and Sims (2010) that provide a critique of this shift.

[viii] This paper was first presented at a conference in 2003, and an early version appears as a NBER working paper (Ding, Lehrer, Rosenquist, & Audrain-McGovern (2006)).

[ix] As discussed earlier, there are likely significant advantages of using arrays of binary indicators for different genetic variations as instruments, relative to using variables that count the number of alleles. The results of models using binary indicator arrays are more flexible, easier to interpret, and enable the researcher more easily to investigate which particular variants are driving the identification. While using arrays that comprehensively dummy out all observed genetic variations will increase the number of instruments and could lead to a many-instrument problem (Hausman, Newey, Woutersen, Chao, & Swanson 2012), new strategies have been proposed

in Belloni and Cherenozhukov (2102) that use the least absolute selection and shrinkage operator ("Lasso") to reduce the number of instruments.

[x] The family fixed effects estimator is a workhorse estimator used in behavioral genetics as well as in family and population economics. For example, this strategy has been used to examine the longer-run effects of early childhood education programs (see e.g. Currie & Thomas (1995), Deming (2009)) and the causes and consequences of early-life health indicators (see e.g. Almond et. al. (2005), Figlio et. al. (2014)), among other research questions.

[xi] A fascinating recent study by Huber, Donnelly, Rokem, & Yeatman (2018) shows that altering a child's educational environment through a targeted intervention program can induce rapid, large-scale changes in the properties of the brain's white matter tissue. This is policy relevant, since white matter properties are often held to underlie variation in performance and to causally influence individual learning trajectories. Individual differences in white matter properties likely reflect the joint influence of genetics and environment. If underlying genetic differences predestine certain individuals to struggle with learning, then understanding the way these genetic differences translate into different effects of learning interventions may suggest optimal ways to allocate different interventions across students to ensure that all children have the opportunity to start schooling with an equal level of (genetically customized) preparation.