Published in final edited form as:

Soc Sci Res. 2016 September ; 59: 13–22. doi:10.1016/j.ssresearch.2016.04.016.

Opportunities and Challenges of Big Data for the Social Sciences: The Case of Genomic Data

Hexuan Liu^{1,3,4} and Guang Guo^{1,2,3}

¹Department of Sociology, the University of North Carolina at Chapel Hill

²Carolina Center for Genome Sciences, the University of North Carolina at Chapel Hill

³Carolina Population Center, the University of North Carolina at Chapel Hill

⁴School of Criminal Justice, the University of Cincinnati

Abstract

In this paper, we draw attention to one unique and valuable source of big data, genomic data, by demonstrating the opportunities they provide to social scientists. We discuss different types of large-scale genomic data and recent advances in statistical methods and computational infrastructure used to address challenges in managing and analyzing such data. We highlight how these data and methods can be used to benefit social science research.

Keywords

genomic data; gene-environment interaction

INTRODUCTION

The era of big data characterized by the explosion of information is upon us. For example, more than 300 million photos are uploaded to Facebook every day;¹ approximately 12 billion Google queries are conducted each month;² about two billion products were purchased from Amazon marketplace sellers in 2014.³ Such big data not only generate economic value, but also promise new scientific discoveries. One of the best-known examples is the research of Ginsberg *et al.* (2009) that sought to predict influenza epidemics using 50 million Google search queries. This research led to the development of Google Flu Trends, a web tool that can detect flu outbreaks seven to ten days before they are reported by the Centers for Disease Control and Prevention (CDC).⁴ As a more recent example, Preis *et*

Corresponding author: Hexuan Liu, CB#3210, Department of Sociology, University of North Carolina, Chapel Hill, NC 27500, Phone: (919)-259-2403, hexuanl@live.unc.edu.

¹“The Top 20 Valuable Facebook Statistics.” (<https://zephoria.com/social-media/top-15-valuable-facebook-statistics/> accessed April 7, 2015)

²“By the Numbers: 60 amazing Google Search Statistics and Facts.” (<http://expandedramblings.com/index.php/by-the-numbers-a-gigantic-list-of-google-stats-and-facts/> accessed April 7, 2015)

³“By the Numbers: 80+ amazing Amazon Statistics.” (<http://expandedramblings.com/index.php/amazon-statistics/2/> accessed April 7, 2015)

⁴“Google Uses Searches to Track Flu’s Spread.” (http://msl1.mit.edu/furdlog/docs/nytimes/2008-11-11_nytimes_google_influenza.pdf accessed April 9, 2015)

al. (2013) used online search data to predict “early warning signs” of changes in the stock market. Their test runs identified the financial crisis in 2008.

Social scientists have embraced enthusiastically the use of big data. In 2011, sociologists Sott Golder and Michael Macy published a paper in *Science* on changes in individuals’ moods over time. Using 500 million public Twitter messages, they showed that people tend to have better moods when they wake up in the morning than later during the day, and the morning peak of positive moods is delayed by two hours on weekends (Golder and Macy 2011). At the 2014 annual meeting of the American Sociological Association, at least seven sessions were held focusing on big data.

In this article, we draw attention to one important type of big data—genomic data—and their implications for the social sciences. Advances in genomics are among the most spectacular scientific achievements over the past few decades. These advances have dramatically improved our understanding in a wide range of areas such as biology, health, medicine, and human nature. Over the past two decades, the development in genomic technology has resulted in a phenomenal reduction in the cost of genomic data collection. As an example, when the Human Genome Project was accomplished in 2003 (IHGSC 2001, 2004), sequencing a single human genome for the first time cost \$3 billion, and thousands of biologists and geneticists from the United States, the United Kingdom, Japan, France, Germany, and China spent 13 years on the project. By 2015, the cost of sequencing one individual’s genome has dropped to less than \$1,000, and the task can be done in hours.⁵ Such genomic revolution has led to a paradigm shift from investigation of the functions of single genes to analyzing the role and structure of the whole genome in hundreds and thousands of individuals. DNA sequencing data represent only one type of genomic data. Other types of genomic data include epigenomic data and genome-wide gene expression data. These data, when combined with traditional social science data, could lead to advances in social science that were unimaginable even a decade ago.

This paper is organized as follows. We first outline opportunities that genomic data provide for the social sciences. We then introduce different types of genomic data that have been or will be available and their potential contributions to social science research. We devote the rest of the paper to introducing statistical methods for analyzing one type of genomic data that are best-known and have generated many valuable findings: the genome-wide genotype data.

OPPORTUNITIES OF GENOMIC DATA FOR THE SOCIAL SCIENCES

In the social sciences it is commonly assumed that human beings are homogeneous at birth and that differences across individuals are attributed to social, cultural, and environmental influences. This assumption has been challenged by rapid development in molecular genetics. In recent decades, considerable effort and resources have been devoted to discovering genetic causes of human diseases (Visscher *et al.* 2012). The National Institutes

⁵“Ten years ago today, it was revealed that the human genome had been decoded. A medical revolution beckoned. So what happened next?” (<http://www.independent.co.uk/news/science/ten-years-ago-today-it-was-revealed-that-the-human-genome-had-been-decoded-a-medical-revolution-beckoned-so-what-happened-next-2011016.html> accessed April 7, 2015)

of Health (NIH)'s Catalog of Published Genome-Wide Association Studies (GWAS) of early 2015 includes more than 2,000 publications that have established associations between thousands of genetic loci and human diseases as well as other traits (Hindorff *et al.*).

Reviewing evidence from behavior genetics based on biometrical analyses, Freese (2008) noted that most of social science outcomes at the individual level are genetically influenced to some extent, and that genetic effects on the outcomes must be mediated through a chain of biological and psychological mechanisms.

Many social science outcomes such as cognitive development, educational attainment, occupational status, binge drinking, and substance abuse are likely to be influenced by numerous interacting genetic and socioenvironmental factors. Incorporating genomic measures will help social scientists better understand the complex interplay among these socioenvironmental and genetic factors. In the following, we outline specific ways social science research may benefit from incorporating genomic information (see Belsky and Israel 2014; Belsky *et al.* 2013c; Boardman *et al.* 2012a; Boardman *et al.* 2013; Boardman *et al.* 2014; Boardman *et al.* 2015; Conley *et al.* 2015; Conley and Rauscher 2013; Conley *et al.* 2013a; Conley *et al.* 2013b; Domingue *et al.* 2015; Domingue *et al.* 2014a; Domingue *et al.* 2014b; Guo *et al.* 2008a; Guo *et al.* 2015a; Guo *et al.* 2015b; Guo *et al.* 2008b; Li *et al.* 2015; Liu and Guo 2015; Liu *et al.* 2015; Mitchell *et al.* 2015; Mitchell *et al.* 2014; Mitchell *et al.* 2013; Perry Forthcoming; Pescosolido *et al.* 2008; Shanahan *et al.* 2008; Simons *et al.* 2011). Genomic sciences are still under rapid development and new types of genomic data have been produced all the time. Therefore, the ways in which social science research may benefit from genomic advances are likely to be extended considerably in the future.

First, studies of gene-environment interactions probably represent the most important opportunities for social scientists. Gene-environment interaction refers to the interdependence between an environmental effect and a genotypic effect. Gene-environment interaction implies that an environmental influence is sensitive to the effect of a genotype and vice versa. Ignoring gene-environment interactions forces us to estimate only an average genetic effect (averaged over all environments) or an average environmental effect (averaged over all genotypes), thus potentially dismissing genetic, environmental or both effects. For example, suppose we estimate a model in which body mass index (BMI) is predicted by variants in the *FTO* gene and educational attainment. Gene-environment interaction is present when the effect of *FTO* on BMI depends on education or when the effect of education depends on *FTO*. Frayling *et al.* (2007) reported that individuals who carry particular variants of the *FTO* gene were found to weigh, on average, 1.2 kg more than those who do not carry such variants. This effect of 1.2 kg is obtained without considering the environment. The effect may be smaller than 1.2 kg (or even absent) for some individuals under certain social conditions but greater for others.

Findings from gene-environment interaction studies can be used in the development of an intervention strategy if there is evidence for exogenous environmental influences (Conley 2009; Fletcher and Conley 2013; Guo *et al.* 2015b). The strategy removes or adjusts influences of social exposures resulted from genetic propensities (e.g., alcoholics cluster due to their shared genetic propensities to drinking). The strategy is based on the idea that

genotypes are fixed, but social exposures might be alterable. A variety of gene-environment interaction models have been proposed, theoretically discussed and empirically tested in the social sciences (Boardman *et al.* 2013; Conley *et al.* 2013b; Daw *et al.* 2013; Guo *et al.* 2015b; Liu and Guo 2015; Mitchell *et al.* 2015; Shanahan and Hofer 2005).

Second, genomic data could help social science researchers obtain evidence for a causal argument. At least two strategies have been developed for this purpose. The more obvious strategy is to use genomic data for isolating socioenvironmental effects from genetic and other biological confounders. Socioenvironmental effects yielded by conventional social science models are often a mixture of social and genomic effects due to genomic confounding. In such cases, socioenvironmental effects are likely to be misestimated. For example, parental influences (measured by parental education, parental occupation, and parental income) on children's educational attainment are rarely purely environmental. Because parents and children share 50% of their DNA, parental influences on children are ambiguously social and genetic. This ambiguity is the so-called gene-environment correlation (Jaffee and Price 2007; Wagner *et al.* 2013). Parental genetic effects and socioenvironmental effects are correlated and entangled. Purer social effects can be isolated when relevant genetic measures correlated with parental measures are explicitly included in the analysis (Conley *et al.* 2015).

As another strategy, certain genetic variants can be used as instrumental variables to establish causal relationships (Fletcher and Lehrer 2009). For example, variants in the ALDH2 (Aldehyde dehydrogenase 2 family) gene may be used to establish a causal relationship between alcohol consumption and a social science outcome such as educational attainment or occupational status. This approach takes advantage of the biological property that those with certain forms of ALDH2 have lower alcohol tolerance and higher allergy to alcohol consumption. These reactions essentially amplify the negative effects of alcohol consumption. The resulting variation in alcohol consumption can then be treated as an exogenous variable in a study on the effect of alcohol consumption. More generally, as Freese (2011):88) noted that "(T)he strict intragenerational exogeneity of the DNA sequence—that the DNA sequence does not change as a result of external events or internal development," DNA data may be explored in the design of a natural experiment.

Third, even if genomic measures are non-interactive and uncorrelated with social science measures, these genomic measures could still contribute to social science research. Non-interaction and non-correlation mean that disregarding genomic information does not bias the estimated socioenvironmental effects; nevertheless, these genomic measures will improve prediction of the outcome under study. As more and more different types of genomic data and more and more measures within a type of genomic data are discovered and utilized, the predictability of human genome for human outcomes will increase. Improved predictability can enrich and deepen our understanding of social science models (Freese 2011).

TYPES OF GENOMIC DATA

The role of genes in human traits has been traditionally investigated based on studies of twins, adoptees, or other family data. Such studies have been adopted in social science research to examine genetic and environmental contributions to social outcomes or to illuminate crucial biological mechanisms through which social context shapes individual outcome (e.g., Boardman *et al.* 2010; Boardman *et al.* 2012b; Guo and Stearns 2002; Nielsen 2006; Nielsen and Roos 2015; Nielsen 2008; Turkheimer *et al.* 2003). In twin/family studies, genetic variants at the molecular level are not observed, and genetic and environmental contributions are estimated as latent variables based on relatedness among genetic relatives. Also, twin/family studies rely on critical statistical assumptions. These assumptions are questioned and violation of these assumptions may lead to biases in the estimates of genetic and environmental influences (Goldberger 1979).

With the availability of candidate gene data in the late 1990s and the first ten years of the 21st century, many studies have been carried out linking human traits with DNA variants. Candidate genes allow social science researchers interested in gene-environment interactions to examine variations in the environmental influences on individuals with different genotypes (e.g., Caspi *et al.* 2003; Caspi *et al.* 2002; Daw *et al.* 2013; Guo *et al.* 2008b; Guo *et al.* 2007; Mitchell *et al.* 2015; Mitchell *et al.* 2011; Simons *et al.* 2011). This candidate gene approach, however, has been criticized because its findings are often not replicated in subsequent studies, and the reliability of this approach has become a concern (Charney and English 2012; Risch *et al.* 2009). The need to produce more robust and replicable findings called for genome-wide methods with more comprehensive genetic variant coverage and more conservative gene-selection thresholds (Caspi *et al.* 2010; Duncan and Keller 2011).

A major data revolution has occurred in genomic studies since the middle of the first decade of the 21st century. During the period, advances in genomic sciences and technology have produced a dazzling range of genomic data. In this section we describe four major types of data that are already generated and analyzed routinely in the field of genomics: genome-wide genotype data (i.e., GWAS data), DNA sequencing data, epigenomic data, and gene expression data. Genome-wide genotype data have been the most familiar to social scientists (e.g., Boardman *et al.* 2014; Conley *et al.* 2015; Domingue *et al.* 2014b; Guo *et al.* 2015a). Although the usefulness of other types of genomic data to the social sciences remains less examined, making these data known will help draw attention to them and get their usefulness investigated.

Genome-wide Genotype Data

The human genome includes approximately 3 billion DNA base pairs (e.g., A-T or C-G) and about 1/300 of them vary across individuals (IHGSC 2004). Genome-wide genotype data use tag single-nucleotide polymorphisms (SNPs) to capture most of the DNA variation across the human genome. Such data typically measure 100,000–2,500,000 SNPs for each individual. These data are analyzed to identify DNA variants associated with specific phenotypes in the population (Hirschhorn and Daly 2005). Measuring genome-wide genotype data has been becoming increasingly less expensive over the years, rendering it feasible to large-scale social science surveys. The Health and Retirement Study (HRS), for

example, has genotyped more than two million genetic variants from each of about 20,000 respondents who provided DNA samples.⁶ Genome-wide genotype data of a similar scale have been collected in the National Longitudinal Study of Adolescent to Adult Health (Add Health)⁷. All of these data have been or will be publicly available through the NIH database of Genotypes and Phenotypes (dbGaP) (<http://www.ncbi.nlm.nih.gov/gap>). Together with longitudinally tracked health and social outcomes, as well as social contexts, these genome-wide genotype data are poised to make major contributions to the social sciences.

Social science research can benefit from genome-wide genotype data in various ways. Most importantly, such data can be used to assess interactions of the social environment and genes. In contrast to candidate gene studies that focus on one or a few genetic variants, studies based on genome-wide genotype data can provide a more comprehensive picture of gene-environment interactions by incorporating information from the whole genome. Genome-wide genotype data have been successfully used in assessing gene-environment interactions in medical research (Garcia-Closas *et al.* 2013; Wu *et al.* 2012). Similarly, social scientists can take advantage of such data to study social outcomes such as personality, delinquency, and educational attainment. In addition, genotype-wide genotype data can be used to isolate social environmental effects from genetic confounders. Conley *et al.* (2015), for example, estimated parental influence on children's educational attainment by controlling a polygenic score constructed using genome-wide genotype data. This approach can be extended to estimate "pure" environmental influences on other outcomes of interest to social scientists.

DNA Sequencing Data—DNA sequencing is the process of establishing the precise order of all base pairs within a DNA molecule. Current genome-wide genotype data typically cover common variants (the frequency of each alternative form of the same genetic variant is greater than 5%). Yet rare variants (the frequency of at least one alternative form of the same genetic variant is below 5%) largely outnumber common variants and may significantly contribute to human phenotypes (Altshuler *et al.* 2010a). Data on rare variants were limited due to the high cost of DNA sequencing. Recent technological advances made it possible to sequence a genome faster and at a lower cost. Heretofore, through sequencing, the Human Genome Project, the SNP Consortium, the International HapMap Project, and the 1000 Genome Project have collectively identified approximately 40 million common and rare variants (Altshuler *et al.* 2010a; Altshuler *et al.* 2012; Altshuler *et al.* 2010b; Consortium 2001; Consortium 2004; Consortium 2005; Frazer *et al.* 2007; Sachidanandam *et al.* 2001).

Rare variants may also play an important part influencing complex social outcomes. If rare variants are dismissed, the genetic contribution to the phenotype of interest can be substantially underestimated. Take educational attainment as an example. While the heritability of educational attainment is estimated to be around 40% (Branigan *et al.* 2013), all measured common variants together explain only 2% of the total variation in educational attainment (Rietveld *et al.* 2013a). Such a discrepancy might be, at least partially, due to unmeasured rare variants. Although methods to handle sequencing data are still in their

⁶Health and Retirement Study GWAS Data. (<http://hrsonline.isr.umich.edu/gwas> accessed July 12, 2015)

⁷Add Health GWAS data (<http://www.cpc.unc.edu/projects/addhealth/design/wave4> accessed July 12, 2015)

infancy, such data will likely prove valuable for the social sciences. They can provide a rich database to assess genetic contribution and thus improve our understanding of gene-environment interactions and genetic confounding.

Genome-wide DNA Methylation Data—Social scientists are particularly excited about the opportunities provided by epigenetic data (Landecker and Panofsky 2013; Shanahan 2013). The exceptional interest in epigenetics derives from that epigenetic mechanisms could alter how genes are expressed without modifying the underlying DNA sequence. Epigenetics is often considered a bridge that connects nature and nurture. Epigenetic processes are highly interactive with environment: environmental variation may routinely change epigenetic patterns, which can, in turn, affect phenotypes. Epigenetics is therefore poised to make major breakthroughs in understanding how genes are regulated and expressed in relation to environmental exposures and life course experiences.

DNA methylation is an epigenetic mechanism used by cells to control gene expression. By adding methyl groups to DNA, methylation modifies the function of the DNA. As a consequence, the same genetic variant may express distinct phenotypes depending on the state of epigenetics. Beach *et al.* (2014), for example, found that cumulative socioeconomic disadvantage was associated with methylation patterns in promoter regions that may lead to mental and physical health risks in African American young adults. In another study, Yehuda *et al.* (2015) showed that holocaust exposure affected FK506 binding protein 5 (FKBP5) methylation that was observed in exposed parents as well in their offspring.

Recent development in genomic sciences has made it possible to measure epigenetic markers across the whole genome in large population-based samples (Rakyan *et al.* 2011). Genome-wide DNA methylation profiling, for example, can measure 500,000 to 1 million methylation sites per individual (Michels *et al.* 2013). An integration of such genome-wide DNA methylation data with longitudinal socioenvironmental measures will enable social scientists to improve our understanding on how epigenetic mechanisms mediate effects of the social environment on a phenotype over time (Mitchell *et al.* 2016; Notterman and Mitchell 2015; Shanahan and Hofer 2011).

Gene Expression Data—Genes affect phenotypes only when they are expressed, that is, when genetic information is used in syntheses of functional gene products such as enzymes and hormones. There are several steps in the process of gene expression, including the transcription, RNA splicing, translation, and post-translational modification of a protein. Gene expression is typically measured by the amount of messenger RNA (mRNA) produced by a gene. Various techniques can be used to detect gene expression level, including differential display, northern/southern blots, DNA microarray, serial analysis of gene expression (SAGE), and RNA sequencing (RNA-seq). RNA-seq, currently the most advanced technique, for example, can detect the presence and quantity of RNA in a biological sample at a given moment in time.

Gene expression has been shown to be associated with the social environment (Cole 2013). Findings based on animal experiments have provided evidence that the social environment regulates gene expression. In an assessment of genetic and environmental influences on

aggressive behavior of rhesus monkeys, the expression of *MAOA* is found to be sensitive to early social experiences (Newman *et al.* 2005). Findings from another study show that a female macaque's ranking within her social environment affects the expression of nearly 700 genes (Tung *et al.* 2012).

There are increasing human studies concerning the relationship between socioeconomic status (SES) and gene expression. Cole *et al.* (2007), for example, found that individuals who experienced chronic social isolation (i.e., loneliness) and those who experienced consistent social integration systematically differed in the expression of more than 200 genes in white blood cells. In a more recent study, Knight *et al.* (2016) showed that low SES is associated with increased expression of stress-related gene expression profiles in hematopoietic stem cell transplant recipients. These human studies are mostly based on observational designs and the associations between variations in the social environment and gene expression may not be causal (see Conley 2009; Fletcher and Conley 2013). Nevertheless, such studies may provide additional insights that help understand and explain the complex relationships among the social environment, genes, and phenotypes.

Advances in high-throughput RNA-seq technologies now allow researchers to survey expression of genes throughout the whole genome (Wang *et al.* 2009). Large social science surveys may collect whole-genome expression data when they become more affordable. These data in conjunction with genome-wide genotype data and social science measures can reveal new insights to important research questions, for example, which genes of interest are subject to social regulation, how the social environment provokes the dynamics, and what social, psychological and biological mechanisms mediate the effects.

ANALYTICAL CHALLENGES: GENOME-WIDE GENOTYPE DATA AS AN EXAMPLE

Analyzing genomic data is often quite different from traditional social science data analysis. Because of their massive sample size and high dimensionality, conventional statistical methods and data processing packages are often inadequate. In this section, we demonstrate recent statistical and computational advances in handling genome-wide genotype data. We focus on genome-wide genotype data because these data were made available the earliest, because geneticists and social scientists have had the most time to develop appropriate methods for such data, and because these data are already widely available for social scientists.

Genome-wide Association

The standard genome-wide association approach typically involves two procedures. First, each genetic variant is used to predict the trait in a statistical model. Because the prediction involves a large number of genetic variants, it is likely that some small p-values are due to chances. To address the multiple testing issue, a conservative p-value threshold (e.g., 5×10^{-8} or smaller) is used to select associative variants. Second, even extremely small p-values do not completely rule out all possible false-positives (i.e., variants that do not contribute to the

phenotype are claimed as associative ones), thus replication using independent data is required to establish validity of the results.

The genome-wide association approach has been extended to study gene-environment interactions (Cornelis *et al.* 2012; Mukherjee *et al.* 2012; Thomas *et al.* 2012). In contrast to the traditional GWAS approach that assumes homogeneous genetic effects across samples, the genome-wide gene-environment interaction (GWGEI) approach allows genetic effects to vary across levels of the environmental indicator or vice versa. The first social science research using the GWGEI approach is the study of Boardman *et al.* (2014). Using more than 260, 000 SNPs, this study examined the interaction between each SNP and educational status (measured by whether the respondent has a college degree) on BMI. However, no consistent gene-environment interaction patterns were reported.

Although the GWGEI approach is useful to identify genetic variants whose effects on traits vary under different environmental conditions, it only estimates the interaction for one genetic variant and one environmental indicator at a time. An extremely stringent p-value threshold (e.g., 10^{-10}) is required to minimize false positive findings (Hutter *et al.* 2013). Using such stringent standards may result in failures to identify interactions with moderate or small effects (i.e., Type II error) without a sufficiently large sample (Boardman *et al.* 2014).

Polygenic Scores

Unlike the GWGEI approach that focuses on one genetic variant at a time, the polygenic score approach takes into account multiple variants simultaneously. The first successful polygenic score analysis using genome-wide genotype data was conducted in a study of schizophrenia (Purcell *et al.* 2009), which found few individual variants associated with the outcome, but a large number of variants together significantly predicted schizophrenia. The polygenic score approach has also been applied for other health-related traits such as height (Thorleifsson *et al.* 2010; Wood *et al.* 2014), BMI (Locke *et al.* 2015; Speliotes *et al.* 2010), and cardiovascular risk (Simonson *et al.* 2011).

There are two approaches to construct polygenic scores based on findings of extant GWAS: “top-hits” and “whole-genome.” The “top-hits” approach calculates polygenic scores by summing the number of risk alleles across genetic variants that pass a genome-wide significance threshold (p -values $< 5 \times 10^{-8}$) (e.g., Belsky *et al.* 2013c; Domingue *et al.* 2014a; Liu and Guo 2015; Qi *et al.* 2012). The “whole-genome” approach is more liberal. In the “whole-genome” approach, polygenic scores are constructed using all or a large number of measured genetic variants (i.e., more liberal or less stringent p-value thresholds), assuming that these variants have moderate or small effects on the outcome (e.g., Conley *et al.* 2015; Domingue *et al.* 2015).

Belsky and Israel (2014) summarized three applications of polygenic scores in social science research. First, polygenic scores can be employed to investigate developmental processes. As an example, Belsky *et al.* (2012) examined associations between body weights at different life stages from birth to young adulthood and polygenic scores based on GWAS. They found that children had similar weights at birth in spite of their genetic predisposition,

but those with higher genetic risks grew at a high rate than those with lower genetic risks, and variations in the growth mediated genetic influences on obesity at later life. Similarly, polygenic scores have also been used in investigations of developmental characteristics of smoking and asthma (Belsky *et al.* 2013a; Belsky *et al.* 2013d). Second, polygenic scores allow social scientists to assess the complex interplay between social context and multiple genetic variants on complex traits. For example, Demerath *et al.* (2013) constructed a polygenic score based on 32 obesity-related SNPs [based on findings of Speliotes *et al.* (2010)], and examined birth-year variation in the genetic association with obesity-related traits. They showed that the genetic association with BMI for males born in 1970s was three times as great as that for those born in 1930s. In another study, using polygenic scores based on the same 32 SNPs, Liu and Guo (2015) provided evidence that cumulative socioeconomic advantage over the life course compensated for the genetic influence on BMI in middle and late adulthood, whereas cumulative socioeconomic disadvantage amplifies such genetic influence. Third, polygenic scores provide a way to study genetic and environmental contributions to social outcomes. Using genomic data from HRS, Domingue *et al.* (2014b) examined genetic and educational contributions to human homogeneity (individuals tend to marry those who are similar to them). They showed significant genetic correlation between married couples, but the strength of the genetic correlation is only about 10% of the assortative mating by education levels.

In recent years, there are increasingly GWAS on social outcomes, such as personality (Moor *et al.* 2012), cognitive functions (Davies *et al.* 2016), educational attainment (Rietveld *et al.* 2013a), antisocial behavior (Tielbeek *et al.* 2012), and subjective wellbeing (Benjamin *et al.* 2016). Social scientists can take advantage of findings in these studies to construct polygenic scores for traits of interest.

There are also limitations in applying polygenic scores in social science research. First, polygenic scores lack molecular specificity. While polygenic scores provide an individual-level measure of genetic predisposition to a trait, they offer little purchase on specific biological mechanisms through which genetic predisposition operates. Moreover, polygenic score results can be affected by population stratification. Currently most GWAS samples are of European descent. Differences in ancestry from this population may introduce noise into polygenic scores applied in other populations. Finally, genetic correlation between the discovery sample (i.e., the sample used to discover genetic associations with the trait and estimate their effects) and the study sample (i.e., the sample that provides genotypes in the calculation of polygenic scores) may bias the polygenic score results (Wray *et al.* 2013). GWAS meta-analysis often uses samples from different sources. If the study sample is part of the discovery sample, polygenic scores in the study sample might be biased. For example, HRS is included in some GWAS discovery samples and so polygenic scores constructed based on those GWAS may have inflated effect sizes in the HRS.

Genomic-relatedness-matrix Restricted Maximum Likelihood Estimation (GREML)

GREML is another innovative approach that handles genome-wide genotype data. GREML was initially developed to tackle the “missing heritability” issue in GWAS (Yang *et al.* 2010). To illustrate, while about 80% of variance in human height is believed to be heritable,

height-related genetic variants identified in GWAS collectively explain less than 20% of observed height variation (Wood *et al.* 2014). A major explanation of the gap in the estimate of genetic contribution is that there are more genetic variants associated with the trait, but these variants cannot be identified using the traditional GWAS approach due to a lack of statistical power (Visscher *et al.* 2012). Using GREML, Yang *et al.* (2011) showed that genome-wide data account for more than 40% of the variation in human height. This method has also been used to estimate the genomic contribution to schizophrenia (Lee *et al.* 2012b), intelligence (Chabris *et al.* 2012; Davies *et al.* 2011), personality traits (Vinkhuyzen *et al.* 2012), subjective well-being (Rietveld *et al.* 2013b), and economic and political preferences (Benjamin *et al.* 2012).

GREML can be applied by social scientists to study gene-environment interactions. Guo *et al.* (2015a), for instance, conducted a GREML analysis to examine how the collective influence of SNPs across the whole human genome on BMI differs by age and historical period. They found that the genomic influence on BMI weakened with age across the life course, and the genomic influence on BMI was substantially and significantly larger after the mid-1980s than in the few decades before the mid-1980s within each age group of 21–40, 40–50, 51–60 and 60 and older. In other two studies, GREML was used to estimate the collective influence of a large number of SNPs on delinquency and violence under conditions with different levels of social control (Li *et al.* 2015; Liu *et al.* 2015). There is evidence that the collective genetic influence was greater among adolescents who lived under conditions with lower levels of social control than those who lived under conditions with higher levels of social control.

GREML has also been extended to estimate the genetic correlation namely, the extent to which different traits are influenced by common genetic variants (Lee *et al.* 2012a). Using the bivariate GREML model, Deary *et al.* (2012) estimated a highly significant genetic correlation of .62 between intelligence in adolescence (age 11) and in late adulthood (age 65–78). In a more recent study, Boardman *et al.* (2015) examined the genetic correlation between education and health measures such as BMI, depression, and self-rated health. They found that the correlation between depression and education and between self-rated health and education were largely explained by common genetic factors, while there was no evidence that the correlation between BMI and education was influenced by common genetic factors.

GREML is not without shortcomings. It assumes that all SNP effects follow a normal distribution. Violation of this assumption may bias the results (Wang *et al.* 2015). In addition, GREML requires genetically unrelated individuals. Due to common environmental effects, the inclusion of related individuals could result in a biased estimate of the genetic variance. This requirement often leads to a reduction of the effective sample size. Finally, GREML estimates are sensitive to the sample size and the number of SNPs used in the analysis. The standard errors may increase dramatically when the sample size and the number of SNPs decrease (Kumar *et al.* 2016; Yang *et al.* 2016).

Analytic and Computing Infrastructure

Big genomic data frequently requires non-generic software and powerful computing capabilities. Along with the explosion of large-scale genomic data, the computing infrastructure has been revolutionized. Specialized bioinformatics tools have been developed to manage and analyze these data. These tools can be freely downloaded and they provide researchers with great computational efficiency. PLINK (Purcell *et al.* 2007), for example, can finish a genome-wide analysis that estimates over 2,000,000 regression models using a dataset including more than 10,000 individuals in minutes. However, some genomic analysis requires such a computing power that it is impractical to process genomic data on a single computer. In such cases, we can partition big analytical tasks into smaller manageable subtasks that can be processed in parallel using computing clusters. Computational innovations (e.g., the MapReduce programming paradigm) have been developed to facilitate management and analysis of big data.

CONCLUSIONS

The increasing availability of genomic data and methods has opened up unprecedented opportunities for social scientists. Almost all human traits of interest to social scientists are complex traits influenced by both environmental and genetic factors. Various types of genomic data together with traditional social science measures will enable them to develop and test new research hypotheses and enrich existing theories.

Social scientists not only take advantage of advances in genomics, but also can make unique contributions to understanding complicated relationships among genes, phenotypes, and environmental influences. For complex traits such as smoking and aggression, effects of individual genetic variants are often small. To detect small genetic effects, data from different sources are combined to achieve sufficient statistical power. As a consequence, the large samples used in current GWAS may not represent a population of interest nor are there consistent phenotypic or environmental measures (Belsky *et al.* 2013b). Moreover, gene-environment interaction research has typically focused on proximate environmental exposures of risks—such as temperature, radiation, virus, and injury—that interact with genetic factors to influence health or behavioral outcomes (Ottman 1996). There is a recent awareness of considering group-level social and cultural processes in the investigation of gene-environment interactions (Boardman *et al.* 2013; Guo *et al.* 2008b; Liu and Guo 2015; Shanahan and Hofer 2011; Simons *et al.* 2011). Social scientists can help design nationally representative samples with more consistent measures of phenotypes and environmental exposures of interest, and develop and test gene-environment interaction hypotheses informed by the social sciences.

In summary, the human genome is a unique and valuable source millions of years in the making. Genomic data are becoming available at a phenomenal rate. It is time for social scientists to collaborate with biologists and geneticists to bring these data together with conventional social science data to advance scientific knowledge and innovation.

Acknowledgments

We would like to acknowledge valuable feedback from Arne Kalleberg, Brian Levy, and Batool Zaidi. We are grateful to the Carolina Population Center (R24 HD050924) for general support.

References

- Altshuler, David L., Durbin, Richard M., Abecasis, Gonçalo R., Bentley, David R., Chakravarti, Aravinda, et al. A Map of Human Genome Variation from Population-Scale Sequencing. *Nature*. 2010a; 467(7319):1061–1073. [PubMed: 20981092]
- Altshuler, David M., Durbin, Richard M., Abecasis, Gonçalo R., Bentley, David R., Chakravarti, Aravinda, et al. An Integrated Map of Genetic Variation from 1,092 Human Genomes. *Nature*. 2012; 491(7422):56–65. [PubMed: 23128226]
- Altshuler, David M., Gibbs, Richard A., Peltonen, Leena, Schaffner, Stephen F., Yu, Fuli, et al. Integrating Common and Rare Genetic Variation in Diverse Human Populations. *Nature*. 2010b; 467(7311):52–58. [PubMed: 20811451]
- Beach, Steven RH., Dogan, Meesha V., Brody, Gene H., Philibert, Robert A. Differential Impact of Cumulative SES Risk on Methylation of Protein–protein Interaction Pathways as A Function of SLC6A4 Genetic Variation in African American Young Adults. *Biological Psychology*. 2014; 96(1): 28–34. [PubMed: 24192273]
- Belsky, Daniel W., Israel, Salomon. Integrating Genetics and Social Science: Genetic Risk Scores. *Biodemography and Social Biology*. 2014; 60(2):137–155. [PubMed: 25343363]
- Belsky, Daniel W., Moffitt, Terrie E., Baker, Timothy B., Biddle, Andrea K., Evans, James P., et al. Polygenic Risk and the Developmental Progression to Heavy, Persistent Smoking and Nicotine Dependence: Evidence from a 4-Decade Longitudinal Study. *JAMA Psychiatry*. 2013a; 70(5):534–542. [PubMed: 23536134]
- Belsky, Daniel W., Moffitt, Terrie E., Caspi, Avshalom. Genetics in Population Health Science: Strategies and Opportunities. *American journal of public health*. 2013b; 103(Suppl 1(1)):S73–83. [PubMed: 23927511]
- Belsky, Daniel W., Moffitt, Terrie E., Sugden, Karen, Williams, Benjamin, Houts, Renate, et al. Development and Evaluation of a Genetic Risk Score for Obesity. *Biodemography and Social Biology*. 2013c; 59(1):85–100. [PubMed: 23701538]
- Belsky, Daniel W., Sears, Malcolm R., Hancox, Robert J., Harrington, HonaLee, Houts, Renate, et al. Polygenic Risk and the Development and Course of Asthma: An Analysis of Data from a Four-Decade Longitudinal Study. *The Lancet Respiratory Medicine*. 2013d; 1(6):453–461. [PubMed: 24429243]
- Belsky, Daniel W., Williams, Benjamin, Poulton, Richie, Caspi, Avshalom, Moffitt, Terrie E., et al. Polygenic Risk, Rapid Childhood Growth, and the Development of Obesity: Evidence from a 4-Decade Longitudinal Study. *Archives of Pediatrics and Adolescent Medicine*. 2012; 166(6):515–521. [PubMed: 22665028]
- Benjamin, Daniel J., Johannesson, Magnus, Peter Visscher, M., Cesarini, David, van der Loos, Matthijs JHM., et al. The Genetic Architecture of Economic and Political Preferences. *Proceedings of the National Academy of Sciences*. 2012; 109(21):8026–8031.
- Benjamin, Daniel, Bartels, Meike, Koellinger, Philipp. Genetic Associations with Subjective Well-Being Also Implicate Depression and Neuroticism. 2016. <http://biorxiv.org/content/biorxiv/early/2015/11/24/032789.full.pdf>
- Boardman, Jason D., Benjamin Domingue, W., Jason Fletcher, M. How Social and Genetic Factors Predict Friendship Networks. *Proceedings of the National Academy of Sciences*. 2012a; 109(43): 17377–17381.
- Boardman, Jason D., Blalock, Casey L., Pampel, Fred C. Trends in the Genetic Influences on Smoking. *Journal of Health and Social Behavior*. 2010; 51(1):108–123. [PubMed: 20420298]
- Boardman, Jason D., Daw, Jonathan, Freese, Jeremy. Defining the Environment in Gene-Environment Research: Lessons from Social Epidemiology. *American Journal of Public Health*. 2013; 103(1):S64–S72. [PubMed: 23927514]

- Boardman, Jason D., Domingue, Benjamin W., Blalock, Casey L., Haberstick, Brett C., Harris, Kathleen M., et al. Is the Gene-Environment Interaction Paradigm Relevant to Genome-Wide Studies? The Case of Education and Body Mass Index. *Demography*. 2014; 51(1):119–139. [PubMed: 24281739]
- Boardman, Jason D., Domingue, Benjamin W., Daw, Jonathan. What Can Genes Tell Us About the Relationship between Education and Health? *Social Science and Medicine*. 2015; 127:171–80. [PubMed: 25113566]
- Boardman, Jason D., Roettger, Michael E., Domingue, Benjamin W., McQueen, Matthew B., Haberstick, Brett C., et al. Gene-Environment Interactions Related to Body Mass: School Policies and Social Context as Environmental Moderators. *Journal of Theoretical Politics*. 2012b; 24(3): 370–388. [PubMed: 23236222]
- Branigan, Amelia R., McCallum, Kenneth J., Freese, Jeremy. Variation in the Heritability of Educational Attainment: An International Meta-Analysis. *Social Forces*. 2013; 92(1):109–140.
- Caspi, Avshalom, Braithwaite, Antony, Poulton, Richie, Sugden, Karen, Moffitt, Terrie E., et al. Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-HTT Gene. *Science* (New York, NY). 2003; 301(5631):386–389.
- Caspi, Avshalom, Hariri, Ahmad R., Holmes, Andrew, Uher, Rudolf, Moffitt, Terrie E. Genetic Sensitivity to the Environment: The Case of the Serotonin Transporter Gene and Its Implications for Studying Complex Diseases and Traits. *The American Journal of Psychiatry*. 2010; 167(5): 509–527. [PubMed: 20231323]
- Caspi, Avshalom, McClay, Joseph, Moffitt, Terrie E., Mill, Jonathan, Martin, Judy, et al. Role of Genotype in the Cycle of Violence in Maltreated Children. *Science*. 2002; 297(5582):851–854. [PubMed: 12161658]
- Chabris, Christopher F., Hebert, Benjamin M., Benjamin, Daniel J., Beauchamp, Jonathan P., Cesarini, David, et al. Most Reported Genetic Associations with General Intelligence Are Probably False Positives. *Psychological Science*. 2012; 23(11):1314–1323. [PubMed: 23012269]
- Charney, Evan, English, William. Candidate Genes and Political Behavior. *American Political Science Review*. 2012; 106(1):1–34.
- Cole, Steve W., Hawkey, Louise C., Arevalo, Jesusa M., Sung, Caroline Y., Rose, Robert M., et al. Social Regulation of Gene Expression in Human Leukocytes. *Genome Biology*. 2007; 8(9):R189–R189. [PubMed: 17854483]
- Cole, Steven W. Social Regulation of Human Gene Expression: Mechanisms and Implications for Public Health. *American Journal of Public Health*. 2013; 103(1):S84–S92. [PubMed: 23927506]
- Conley, Dalton. The Promise and Challenges of Incorporating Genetic Data into Longitudinal Social Science Surveys and Research. *Biodemography and social biology*. 2009; 55(2):238–251. [PubMed: 20183907]
- Conley, Dalton, Domingue, Benjamin W., Cesarini, David, Dawes, Christopher, Rietveld, Cornelius A., et al. Is the Effect of Parental Education on Offspring Biased or Moderated by Genotype? *Sociological Science*. 2015; 2:82–105.
- Conley, Dalton, Emily, Rauscher. Genetic Interactions with Prenatal Social Environment: Effects on Academic and Behavioral Outcomes. *Journal of Health and Social Behavior*. 2013; 54(1):109–127. [PubMed: 23475742]
- Conley, Dalton, Rauscher, Emily, Dawes, Christopher, Magnusson, Patrik KE., Siegal, Mark L. Heritability and the Equal Environments Assumption: Evidence from Multiple Samples of Misclassified Twins. *Behavior Genetics*. 2013a; 43(5):415–426. [PubMed: 23903437]
- Conley, Dalton, Rauscher, Emily, Siegal, Mark L. Beyond Orchids and Dandelions: Testing the 5-HTT “Risky” Allele for Evidence of Phenotypic Capacitance and Frequency-Dependent Selection. *Biodemography and Social Biology*. 2013b; 59(1):37–56. [PubMed: 23701535]
- Consortium, International Human Genome Sequencing. Initial Sequencing and Analysis of the Human Genome (Vol 409, Pg 860, 2001). *Nature*. 2001; 412(6846):565–566.
- Consortium, International Human Genome Sequencing. Finishing the Euchromatic Sequence of the Human Genome. *Nature*. 2004; 431(7011):931–945. [PubMed: 15496913]
- Consortium, The International HapMap. A Haplotype Map of the Human Genome. *Nature*. 2005; 437(7063):1299–1320. [PubMed: 16255080]

- Cornelis, Marilyn C., Tchetgen, Eric J., Liang, Liming, Qi, Lu, Chatterjee, Nilanjan, et al. Gene-Environment Interactions in Genome-Wide Association Studies: A Comparative Study of Tests Applied to Empirical Studies of Type 2 Diabetes. *American Journal of Epidemiology*. 2012; 175(3):191–202. [PubMed: 22199026]
- Davies G, Luciano M, McGhee K, Lopez L, Gow AJ, et al. Genome-Wide Association Studies Establish That Human Intelligence Is Highly Heritable and Polygenic. *Molecular Psychiatry*. 2011; 16(10):996–1005. [PubMed: 21826061]
- Davies G, Marioni RE, Liewald DC, Hill WD, Hagenaars SP, Harris SE, Ritchie SJ, Luciano M, Fawnsritchie C, Lyall D, Cullen B, Cox SR, Hayward C, Porteous DJ, Evans J, McIntosh AM, Gallacher J, Craddock N, Pell JP, Smith DJ, Gale CR, Deary IJ. Genome-wide Association Study of Cognitive Functions and Educational Attainment in UK Biobank (N=112151). *Molecular Psychiatry*. 2016; doi: 10.1038/mp.2016.45
- Daw, Jonathan, Shanahan, Michael, Harris, Kathleen Mullan, Smolen, Andrew, Haberstick, Brett, et al. Genetic Sensitivity to Peer Behaviors: 5httlpr, Smoking, and Alcohol Consumption. *Journal of Health and Social Behavior*. 2013; 54(1):92–108. [PubMed: 23292504]
- Deary, Ian J., Corley, Janie, Redmond, Paul, Fox, Helen C., Rowe, Suzanne J., et al. Genetic Contributions to Stability and Change in Intelligence from Childhood to Old Age. *Nature*. 2012; 482(7384):212–215. [PubMed: 22258510]
- Demerath, Ellen W., Towne, Bradford, Choh, Audrey C., Johnson, William, Curran, Joanne E., et al. The Positive Association of Obesity Variants with Adulthood Adiposity Strengthens over an 80-Year Period: A Gene-by-Birth Year Interaction. *Human Heredity*. 2013; 75(2–4):175–185. [PubMed: 24081233]
- Domingue, Benjamin W., Belsky, Daniel W., Conley, Dalton, Harris, Kathleen, Boardman, Jason D. Polygenic Influence on Educational Attainment: New Evidence from the National Longitudinal Study of Adolescent to Adult Health. *AERA Open*. 2015; 1(3):1–13. [PubMed: 28164148]
- Domingue, Benjamin W., Belsky, Daniel W., Harris, Kathleen Mullan, Smolen, Andrew, McQueen, Matthew B., et al. Polygenic Risk Predicts Obesity in Both White and Black Young Adults. *PloS one*. 2014a; 9(7):e101596. [PubMed: 24992585]
- Domingue, Benjamin W., Fletcher, Jason, Conley, Dalton, Boardman, Jason D. Genetic and Educational Assortative Mating among Us Adults. *Proceedings of the National Academy of Sciences*. 2014b; 111(22):7996–8000.
- Duncan, Laramie E., Keller, Matthew C. A Critical Review of the First 10 Years of Candidate Gene-by-Environment Interaction Research in Psychiatry. *American Journal of Psychiatry*. 2011; 168(10):1041–1049. [PubMed: 21890791]
- Fletcher, Jason M., Conley, Dalton. The Challenge of Causal Inference in Gene-Environment Interaction Research: Leveraging Research Designs from the Social Sciences. *American journal of public health*. 2013; 103(1):S42–S45. [PubMed: 23927518]
- Fletcher, Jason M., Lehrer, Steven F. Using Genetic Lotteries within Families to Examine the Causal Impact of Poor Health on Academic Achievement. *National Bureau of Economic Research*; 2009.
- Frayling, Timothy M., Timpson, Nicholas J., Weedon, Michael N., Zeggini, Eleftheria, Freathy, Rachel M., et al. A Common Variant in the Fto Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity. *Science*. 2007; 316(5826):889–894. [PubMed: 17434869]
- Frazer, Kelly A., Ballinger, Dennis G., Cox, David R., Hinds, David A., Stuve, Laura L., et al. A Second Generation Human Haplotype Map of over 3.1 Million Snps. *Nature*. 2007; 449(7164):851–861. [PubMed: 17943122]
- Freese, Jeremy. Genetics and the Social Science Explanation of Individual Outcomes. *The American Journal of Sociology*. 2008; 114(S1):S1–S35.
- Freese, Jeremy. Integrating Genomic Data and Social Science. *Politics and the Life Sciences*. 2011; 30(2):88–92. [PubMed: 22702426]
- Garcia-Closas, Montserrat, Rothman, Nathaniel, Figueroa, Jonine D., Prokunina-Olsson, Ludmila, Han, Summer S., et al. Common Genetic Polymorphisms Modify the Effect of Smoking on Absolute Risk of Bladder Cancer. *Cancer Research*. 2013; 73(7):2211–2220. [PubMed: 23536561]

- Ginsberg, Jeremy, Smolinski, Mark S., Brilliant, Larry, Patel, Rajan S., Mohebbi, Matthew H., et al. Detecting Influenza Epidemics Using Search Engine Query Data. *Nature*. 2009; 457(7232):1012–1014. [PubMed: 19020500]
- Goldberger, Arthur S. Heritability. *Economica*. 1979; 46(184):327–347.
- Golder, Scott A., Macy, Michael W. Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength across Diverse Cultures. *Science*. 2011; 333(6051):1878–1881. [PubMed: 21960633]
- Guo G, Tong YY, Cai TJ. Gene by Social Context Interactions for Number of Sexual Partners among White Male Youths: Genetics-Informed Sociology. *American journal of sociology*. 2008a; 114(3):S36–S66.
- Guo, Guang, Liu, Hexuan, Wang, Ling, Shen, Haipeng, Hu, Wen. The Genome-Wide Influence on Human Bmi Depends on Physical Activity, Life-Course, and Historical Period. *Demography*. 2015a; 52(5):1651–1670. [PubMed: 26319003]
- Guo, Guang, Li, Yi, Cai, Tianji, Wang, Hongyu, Duncan, Greg J. Peer Influence, Genetic Propensity, and Binge Drinking: A Natural Experiment and a Replication. *American Journal of Sociology*. 2015b; 121(3):914–954.
- Guo, Guang, Roettger, Michael E., Cai, Tianji. The Integration of Genetic Propensities into Social-Control Models of Delinquency and Violence among Male Youths. *American Sociological Review*. 2008b; 73(4):543–568.
- Guo, Guang, Roettger, Michael E., Shih, Jean C. Contributions of the Dat1 and Drd2 Genes to Serious and Violent Delinquency among Adolescents and Young Adults. *Human Genetics*. 2007; 121(1): 125–136. [PubMed: 17120049]
- Guo, Guang, Stearns, Elizabeth. The Social Influences on the Realization of Genetic Potential for Intellectual Development. *Social Forces*. 2002; 80(3):881–910.
- Hindorf, LA., MacArthur, J (European Bioinformatics Institute), Morales, J (European Bioinformatics Institute), Junkins, HA., Hall, PN., et al. A Catalog of Published Genome-Wide Association Studies. www.genome.gov/gwastudies
- Hirschhorn, Joel N., Daly, Mark J. Genome-Wide Association Studies for Common Diseases and Complex Traits. *Nature Reviews Genetics*. 2005; 6(2):95–108.
- Hutter, Carolyn M., Mechanic, Leah E., Chatterjee, Nilanjan, Kraft, Peter, Gillanders, Elizabeth M., on behalf of the NCI Gene-Environment Think Tank. Gene-Environment Interactions in Cancer Epidemiology: A National Cancer Institute Think Tank Report. *Genetic Epidemiology*. 2013; 37(7):643–657. [PubMed: 24123198]
- Jaffee SR, Price TS. Gene-Environment Correlations: A Review of the Evidence and Implications for Prevention of Mental Illness. *Molecular Psychiatry*. 2007; 12(5):432–442. [PubMed: 17453060]
- Knight, Jennifer M., Rizzo, J Douglas, Logan, Brent R., Wang, Tao, Arevalo, Jesusa MG., et al. Low Socioeconomic Status, Adverse Gene Expression Profiles, and Clinical Outcomes in Hematopoietic Stem Cell Transplant Recipients. *Clinical Cancer Research*. 2016; 22(1):69–78. [PubMed: 26286914]
- Kumar, Siddharth Krishna, Feldman, Marcus W., Rehkopf, David H., Tuljapurkar, Shripad. Limitations of GCTA as A Solution to the Missing Heritability Problem. *Proceedings of the National Academy of Sciences*. 2016; 113(1):E61–E70.
- Landecker, Hannah, Panofsky, Aaron. From Social Structure to Gene Regulation, and Back: A Critical Introduction to Environmental Epigenetics for Sociology. *Annual Review of Sociology*. 2013; 39:333–357.
- Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR. Estimation of Pleiotropy between Complex Diseases Using Single-Nucleotide Polymorphism-Derived Genomic Relationships and Restricted Maximum Likelihood. *Bioinformatics*. 2012a; 28(19):2540–2542. [PubMed: 22843982]
- Lee, S Hong, DeCandia, Teresa R., Ripke, Stephan, Yang, Jian, Sullivan, Patrick F., et al. Estimating the Proportion of Variation in Susceptibility to Schizophrenia Captured by Common Snps. *Nature Genetics*. 2012b; 44(3):247–U35. [PubMed: 22344220]
- Li, Yi, Liu, Hexuan, Guo, Guang. Does Marriage Moderate Genetic Effects on Delinquency and Violence? *Journal of Marriage and Family*. 2015; 77(5):1217–1233. [PubMed: 26549892]

- Liang, Liming, Willis-Owen Saffron, AG., Laprise, Catherine, Wong Kenny, CC., Davies, Gwyneth A., et al. An Epigenome-Wide Association Study of Total Serum Immunoglobulin E Concentration. *Nature*. 2015; 520(7549):670–674. [PubMed: 25707804]
- Liu, Hexuan, Guo, Guang. Lifetime Socioeconomic Status, Historical Context, and Genetic Inheritance in Shaping Body Mass in Middle and Late Adulthood. *American Sociological Review*. 2015; 80(4):705–737. [PubMed: 27231400]
- Liu, Hexuan, Li, Yi, Guo, Guang. Gene by Social-Environment Interaction for Youth Delinquency and Violence: Thirty-Nine Aggression-Related Genes. *Social Forces*. 2015; 93(3):881–903.
- Liu, Yun, Aryee, Martin J., Padyukov, Leonid, Fallin, M Daniele, Hesselberg, Espen, et al. Epigenome-Wide Association Data Implicate DNA Methylation as an Intermediary of Genetic Risk in Rheumatoid Arthritis. *Nature biotechnology*. 2013; 31(2):142–147.
- Locke, Adam E., Kahali, Bratati, Berndt, Sonja I., Justice, Anne E., Pers, Tune H., et al. Genetic Studies of Body Mass Index Yield New Insights for Obesity Biology. *Nature*. 2015; 518(7538):197–U401. [PubMed: 25673413]
- Michels, Karin B., Binder, Alexandra M., Dedeurwaerder, Sarah, Epstein, Charles B., Greally, John M., et al. Recommendations for the Design and Analysis of Epigenome-Wide Association Studies. *Nature Methods*. 2013; 10(10):949–955. [PubMed: 24076989]
- Mitchell C, Brooks-Gunn J, Garfinkel I, McLanahan S, Notterman D, et al. Family Structure Instability, Genetic Sensitivity, and Child Well-Being. *American Journal of Sociology*. 2015; 120(4):1195–1225.
- Mitchell C, Hobcraft J, McLanahan SS, Siegel SR, Berg A, et al. Social Disadvantage, Genetic Sensitivity, and Children's Telomere Length. *Proceedings of the National Academy of Sciences of the United States of America*. 2014; 111(16):5944–5949. [PubMed: 24711381]
- Mitchell, Colter, McLanahan, Sara, Brooks-Gunn, Jeanne, Garfinkel, Irwin, Hobcraft, John, et al. Genetic Differential Sensitivity to Social Environments: Implications for Research. *American Journal of Public Health*. 2013; 103(1):S102–S110. [PubMed: 23927507]
- Mitchell, Colter, Notterman, Daniel, Brooks-Gunn, Jeanne, Hobcraft, John, Garfinkel, Irwin, et al. Role of Mother's Genes and Environment in Postpartum Depression. *Proceedings of the National Academy of Sciences of the United States of America*. 2011; 108(20):8189–8193. [PubMed: 21576482]
- Mitchell, Colter, Schneper, Lisa M., Notterman, Daniel A. DNA Methylation, Early Life Environment, and Health Outcomes. *Pediatric Research*. 2016; 79(1):212–219. [PubMed: 26466079]
- de Moor MHM, Costa PT, Terracciano A, Krueger RF, de Geus EJC, et al. Meta-analysis of Genome-wide Association Studies for Personality. *Molecular Psychiatry*. 2012; 17(3):337–349. [PubMed: 21173776]
- Mukherjee, Bhramar, Ahn, Jaeil, Gruber, Stephen B., Chatterjee, Nilanjan. Testing Gene-Environment Interaction in Large-Scale Case-Control Association Studies: Possible Choices and Comparisons. *American Journal of Epidemiology*. 2012; 175(3):177–190. [PubMed: 22199027]
- Newman, Timothy K., Syagailo, Yana V., Barr, Christina S., Wendland, Jens R., Champoux, Maribeth, et al. Monoamine Oxidase a Gene Promoter Variation and Rearing Experience Influences Aggressive Behavior in Rhesus Monkeys. *Biological Psychiatry*. 2005; 57(2):167–172. [PubMed: 15652876]
- Nielsen, François. Achievement and Ascription in Educational Attainment: Genetic and Environmental Influences on Adolescent Schooling. *Social Forces*. 2006; 85(1):193–216.
- Nielsen, François, Micah Roos, J. Genetics of Educational Attainment and the Persistence of Privilege at the Turn of the 21st Century. *Social Forces*. 2015
- Nielsen, François. The Nature of Social Reproduction: Two Paradigms of Social Mobility. *Sociologica*. 2008:1–35.
- Notterman, Daniel A., Mitchell, Colter. Epigenetics and Understanding the Impact of Social Determinants of Health. *Pediatric Clinics of North America*. 2015; 62(5):1227–1240. [PubMed: 26318949]
- Ottman, Ruth. Gene-Environment Interaction: Definitions and Study Designs. *Preventive Medicine*. 1996; 25(6):764–770. [PubMed: 8936580]

- Perry, Brea L. Gendering Genetics: Social and Biological Contingencies in the Protective Effects of Social Integration for Men and Women. *American Journal of Sociology*. Forthcoming.
- Pescosolido, Bernice A., Perry, Brea L., Scott Long, J., Martin, Jack K., Jr, Nurnberger, John I., et al. Under the Influence of Genetics: How Transdisciplinarity Leads Us to Rethink Social Pathways to Illness. *American Journal of Sociology*. 2008; 114(S1):S171–S201.
- Preis, Tobias, Moat, Helen Susannah, Eugene Stanley, H. Quantifying Trading Behavior in Financial Markets Using Google Trends. *Scientific Reports*. 2013; 3(1684):1–6.
- Purcell, Shaun M., Wray, Naomi R., Stone, Jennifer L., Visscher, Peter M., O'Donovan, Michael C., et al. Common Polygenic Variation Contributes to Risk of Schizophrenia and Bipolar Disorder. *Nature*. 2009; 460(7256):748–752. [PubMed: 19571811]
- Purcell, Shaun, Neale, Benjamin, Todd-Brown, Kathe, Thomas, Lori, Ferreira, Manuel AR., et al. Plink: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*. 2007; 81(3):559–575. [PubMed: 17701901]
- Qi Q, Chu AY, Kang JH, Jensen MK, Curhan GC, et al. Sugar-Sweetened Beverages and Genetic Risk of Obesity. *New England Journal of Medicine*. 2012; 367(15):1387–96. [PubMed: 22998338]
- Rakyan, Vardhman K., Down, Thomas A., Balding, David J., Beck, Stephan. Epigenome-Wide Association Studies for Common Human Diseases. *Nature Reviews Genetics*. 2011; 12(8):529–541.
- Rietveld, Cornelius A., Agrawal, Arpana, Eriksson, Johan G., Albrecht, Eva, Alizadeh, Behrooz Z., et al. Gwas of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment. *Science*. 2013a; 340(6139):1467–1471. [PubMed: 23722424]
- Rietveld, Cornelius A., Krueger, Robert F., Bartels, Meike, Collier, David A., Benjamin, Daniel J., et al. Molecular Genetics and Subjective Well-Being. *Proceedings of the National Academy of Sciences*. 2013b; 110(24):9692–9697.
- Risch, Neil, Herrell, Richard, Lehner, Thomas, Liang, Kung-Yee, Eaves, Lindon, et al. Interaction between the Serotonin Transporter Gene (5-HTTLPR), Stressful Life Events, and Risk of Depression: A Meta-Analysis. *The Journal of the American Medical Association*. 2009; 301(23):2462–2471. [PubMed: 19531786]
- Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, et al. A Map of Human Genome Sequence Variation Containing 1.42 Million Single Nucleotide Polymorphisms. *Nature*. 2001; 409(6822):928–933. [PubMed: 11237013]
- Shanahan, Michael J. Social Genomics and the Life Course: Opportunities and Challenges for Multilevel Research. In: Waite, L., editor. *Changing Social Processes and Social Contexts in an Aging Society*. National Academy of Science; 2013. p. 255–76.
- Shanahan, Michael J., Hofer, Scott M. Social Context in Gene-Environment Interactions: Retrospect and Prospect. *The Journals of Gerontology Series B, Psychological Sciences and Social Sciences*. 2005; 60(Spec 1 (Special Issue 1)):65–76.
- Shanahan, Michael J., Hofer, Scott M. Molecular Genetics, Aging, and Well-Being: Sensitive Period, Accumulation, and Pathway Models. In: R. H. B. a. L. K. George. , editor. *Handbook of Aging and the Social Sciences*. 7th. New York: Elsevier; 2011. p. 135–147.
- Shanahan, Michael J., Vaisey, Stephen, Erickson, Lance D., Smolen, Andrew. Environmental Contingencies and Genetic Propensities: Social Capital, Educational Continuation, and Dopamine Receptor Gene *Drd2*. *American Journal of Sociology*. 2008; 114(Suppl 1):S260–S286.
- Simons, Ronald L., Lei, Man Kit, Beach, Steven RH., Brody, Gene H., Philibert, Robert A., et al. Social Environment, Genes, and Aggression: Evidence Supporting the Differential Susceptibility Perspective. *American Sociological Review*. 2011; 76(6):883–912.
- Simonson, Matthew A., Wills, Amanda G., Keller, Matthew C., McQueen, Matthew B. Recent Methods for Polygenic Analysis of Genome-Wide Data Implicate an Important Effect of Common Variants on Cardiovascular Disease Risk. *BMC Medical Genetics*. 2011; 12(1):146–146. [PubMed: 22029572]
- Speliotes, Elizabeth K., Mägi, Reedik, Buchanan, Thomas A., Randall, Joshua C., Vedantam, Sailaja, et al. Association Analyses of 249,796 Individuals Reveal 18 New Loci Associated with Body Mass Index. *Nature Genetics*. 2010; 42(11):937–48. [PubMed: 20935630]

- Thomas, Duncan C., Lewinger, Juan Pablo, Murcray, Cassandra E., James Gauderman, W. Invited Commentary: Ge-Whiz! Ratcheting Gene-Environment Studies up to the Whole Genome and the Whole Exposome. *American Journal of Epidemiology*. 2012; 175(3):203–207. [PubMed: 22199029]
- Thorleifsson G, Palmer CN, Johnson T, Luan J, Park JH, et al. Hundreds of Variants Clustered in Genomic Loci and Biological Pathways Affect Human Height. *Nature*. 2010; 467(7317):832–838. [PubMed: 20881960]
- Tielbeek, Jorim J., Medland, Sarah E., Benyamin, Beben, Byrne, Enda M., Heath, Andrew C., Madden, Pamela AF., Martin, Nicholas G., Wray, Naomi R., Verweij, Karin JH. Unraveling the Genetic Etiology of Adult Antisocial Behavior: A Genome-Wide Association Study. *PloS One*. 2012; 7(10)
- Tung, Jenny, Barreiro, Luis B., Johnson, Zachary P., Hansen, Kasper D., Michopoulos, Vasiliki, et al. Social Environment Is Associated with Gene Regulatory Variation in the Rhesus Macaque Immune System. *Proceedings of the National Academy of Sciences of the United States of America*. 2012; 109(17):6490–6495. [PubMed: 22493251]
- Turkheimer, Eric, Haley, Andreana, Waldron, Mary, D'Onofrio, Brian, Gottesman, Irving I. Socioeconomic Status Modifies Heritability of Iq in Young Children. *Psychological Science*. 2003; 14(6):623–628. [PubMed: 14629696]
- Vinkhuyzen AAE, Luciano M, Payton A, Horan M, Ollier W, et al. Common Snps Explain Some of the Variation in the Personality Dimensions of Neuroticism and Extraversion. *Translational psychiatry*. 2012; 2(4):e102. [PubMed: 22832902]
- Visscher, Peter M., Brown, Matthew A., McCarthy, Mark I., Yang, Jian. Five Years of Gwas Discovery. *American Journal of Human Genetics*. 2012; 90(1):7–24. [PubMed: 22243964]
- Wagner, Brandon, Li, Jiang, Liu, Hexuan, Guo, Guang. Gene-Environment Correlation: Difficulties and a Natural Experiment-Based Strategy. *American Journal of Public Health*. 2013; 103(Suppl 1 (Journal Article)):S167–73. [PubMed: 23927502]
- Wang, Ling, Shen, Haipeng, Liu, Hexuan, Guo, Guang. Mixture Snps Effect on Phenotype in Genome-Wide Association Studies. *BMC Genomics*. 2015; 16(3):1–11. [PubMed: 25553907]
- Wang, Zhong, Gerstein, Mark, Snyder, Michael. Rna-Seq: A Revolutionary Tool for Transcriptomics. *Nature Reviews Genetics*. 2009; 10(1):57–63.
- Wood AR, Esko T, Yang J, Vedantam S, Pers TH, et al. Defining the Role of Common Variation in the Genomic and Biological Architecture of Adult Human Height. *Nature Genetics*. 2014; 46(11): 1173–1186. [PubMed: 25282103]
- Wray, Naomi R., Yang, Jian, Hayes, Ben J., Price, Alkes L., Goddard, Michael E., et al. Pitfalls of Predicting Complex Traits from Snps. *Nature reviews Genetics*. 2013; 14(7):507–515.
- Wu, Chen, Kraft, Peter, Zhai, Kan, Chang, Jiang, Wang, Zhaoming, et al. Genome-Wide Association Analyses of Esophageal Squamous Cell Carcinoma in Chinese Identify Multiple Susceptibility Loci and Gene-Environment Interactions. *Nature Genetics*. 2012; 44(10):1090–1097. [PubMed: 22960999]
- Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, et al. Common SNPs Explain a Large Proportion of the Heritability for Human Height. *Nature Genetics*. 2010; 42(7):565–569. [PubMed: 20562875]
- Yang, Jian, Hayes, M Geoffrey, Hill, William G., Landi, Maria Teresa, Alonso, Alvaro, et al. Genome Partitioning of Genetic Variation for Complex Traits Using Common Snps. *Nature Genetics*. 2011; 43(6):519–525. [PubMed: 21552263]
- Yang, Jian, Hong Lee, S., Wray, Naomi R., Goddard, Michael E., Visscher, Peter. Commentary on 'Limitations of GCTA as A Solution to the Missing Heritability Problem'. *BioRxiv*. 2016. <http://biorxiv.org/content/early/2016/01/20/036574.abstract>
- Yehuda, Rachel, Daskalakis, Nikolaos P., Bierer, Linda M., Bader, Heather N., Klengel, Torsten, Holsboer, Florian, Binder, Elisabeth B. Holocaust Exposure Induced Intergenerational Effects on FKBP5 Methylation. *Biological Psychiatry*. 2015. <http://dx.doi.org/10.1016/j.biopsych.2015.08.005>