INTEGRATION OF SOCIOLOGY WITH GENOMIC DATA IN STUDIES OF SOCIAL STRATIFICATION AND DELINQUENCY

Hexuan Liu

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Sociology.

Chapel Hill
2016

Approved by:

Guang Guo

Kathleen M. Harris

François Nielsen

Michael J. Shanahan

Claire Yang

ABSTRACT

Hexuan Liu: Integration of Sociology with Genomic Data in Studies of Social Stratification and Delinquency
(Under the direction of Guang Guo)

This dissertation demonstrates how social theories and genomics can be integrated to improve our understanding of sociological issues. I conduct the following studies combining genomic data and conventional sociological measures.

First, I investigate the interaction of social environment and genetic factors on delinquency and violence. Using data from the National Longitudinal Study of Adolescent to Adult (Add Health), I find that adverse social environments are associated with greater genetic risk of delinquency and violence, while favorable social environments are associated with smaller genetic risk of delinquency and violence.

Second, I assess genetic and environmental contributions to socioeconomic stability and mobility over the life course. I examine genetic and environmental influences on socioeconomic achievement at different life stages, taking advantage of the genome-wide data in HRS. I provide evidence that both genetic and environment factors make significant contributions to stability and mobility of socioeconomic achievement over the life span.

Third, I investigate genetic and environmental influences on educational attainment across generations. In this study, I conceptualize a model of multigenerational influences and test the model using educational measures from three generations in conjunction with genome-wide

data in HRS. I find significant genetic correlations in educational attainment across three generations. This suggests genetic factors play an important role in stabilizing intergenerational educational attainment in the U.S. Also, I provide evidence that about half of parent's genetic influence on children's education can be ascribed to genetic transmission and the other half is medicated by parents' own education.

To my father, mother and Qing

ACKNOWLEDGEMENTS

I would like to thank those who helped me complete this dissertation. First and foremost I want to thank my advisor Dr. Guang Guo for his guidance and support throughout the entire process of this research and writing the dissertation. I would also like to thank Dr. Kathleen Harris, Dr. François Nielsen, Dr. Michael Shanahan, and Dr. Claire Yang for their insightful comments and suggestions. My deepest gratitude goes to my beloved parents and wife. Their love, understanding, and support accompanied me and encouraged me throughout my study. Finally, I would like to extend my best regards to all of those who helped and supported me in various ways throughout my PhD study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

1.1 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 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) notes 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 this chapter, I 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 2016; 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) notes 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).

1.2 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.

1.2.1 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.[1] Genome-wide genotype data of a similar scale have been collected in the National Longitudinal Study of Adolescent to Adult Health (Add Health)[2]. 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.

---

[1] Health and Retirement Study GWAS Data. (http://hrsonline.isr.umich.edu/gwas accessed July 12, 2015)

[2] Add Health GWAS data (http://www.cpc.unc.edu/projects/addhealth/design/wave4 accessed July 12, 2015)

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, estimate 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.

1.2.2 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 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.

1.2.3 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, find that cumulative socioeconomic disadvantage is 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) show that holocaust exposure affects FK506 binding protein 5 (FKBP5) methylation that is 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).

1.2.4 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, find 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) show 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.

1.3 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, I demonstrate recent statistical and computational advances in handling genome-wide genotype data. I 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.

1.3.1 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 examines the interaction between each SNP and educational status (measured by whether the respondent has a colleague 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).

1.3.2 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) summarize 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) examine associations between body weights at different life stages from birth to young adulthood and polygenic scores based on GWAS. They find 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) construct a polygenic score based on 32 obesity-related SNPs [based on findings of Speliotes *et al.* (2010)], and examine birth-year variation in the genetic association with obesity-related traits. They show 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) provide evidence that cumulative socioeconomic advantage over the life course compensates 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) examine genetic and educational contributions to human homogeny (individuals tend to marry those who are similar to them). They show 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 HRS.

1.3.3 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) show that genome-wide data account for more than 40% of the variation in human height.

GREML can be applied by social scientists to study gene-environment interactions. Guo *et al.* (2015a), for instance, conduct 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 find 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 is 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 is greater among adolescents who live under conditions with lower levels of social control than those who live under conditions with higher levels of social control.

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).

1.3.4 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.

REFERENCES

Altshuler, David L., Richard M. Durbin, Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, *et al.* 2010a. "A Map of Human Genome Variation from Population-Scale Sequencing." *Nature* 467(7319):1061-1073.

Altshuler, David M., Richard M. Durbin, Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, *et al.* 2012. "An Integrated Map of Genetic Variation from 1,092 Human Genomes." *Nature* 491(7422):56-65.

Altshuler, David M., Richard A. Gibbs, Leena Peltonen, Stephen F. Schaffner, Fuli Yu, *et al.* 2010b. "Integrating Common and Rare Genetic Variation in Diverse Human Populations." *Nature* 467(7311):52-58.

Beach, Steven R.H., Meesha V. Dogan, Gene H. Brody, and Robert A. Philibert. 2014." 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* 96 (1): 28-34.

Belsky, Daniel W. and Salomon Israel. 2014. "Integrating Genetics and Social Science: Genetic Risk Scores." *Biodemography and Social Biology* 60(2):137-155.

Belsky, Daniel W., Terrie E. Moffitt, Timothy B. Baker, Andrea K. Biddle, James P. Evans, *et al.* 2013a. "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, Daniel W., Terrie E. Moffitt, and Avshalom Caspi. 2013b. "Genetics in Population Health Science: Strategies and Opportunities." *American journal of public health* 103 Suppl 1(1):S73-83.

Belsky, Daniel W., Terrie E. Moffitt, Karen Sugden, Benjamin Williams, Renate Houts, *et al.* 2013c. "Development and Evaluation of a Genetic Risk Score for Obesity." *Biodemography and Social Biology* 59(1):85-100.

Belsky, Daniel W., Malcolm R. Sears, Robert J. Hancox, HonaLee Harrington, Renate Houts, *et al.* 2013d. "Polygenic Risk and the Development and Course of Asthma: An Analysis of Data from a Four-Decade Longitudinal Study." *The Lancet Respiratory Medicine* 1(6):453-461.

Belsky, Daniel W., Benjamin Williams, Richie Poulton, Avshalom Caspi, Terrie E. Moffitt, *et al.* 2012. "Polygenic Risk, Rapid Childhood Growth, and the Development of Obesity: Evidence from a 4-Decade Longitudinal Study." *Archives of Pediatrics and Adolescent Medicine* 166(6):515-521.

Benjamin, Daniel. J. , Magnus Johannesson, Peter Visscher, M., David Cesarini, Matthijs. J. H. M. van der Loos, *et al.* 2012. "The Genetic Architecture of Economic and Political Preferences." *Proceedings of the National Academy of Sciences* 109(21):8026-8031.

Benjamin, Daniel, Meike Bartels, and Philipp Koellinger. 2016. "Genetic Associations with Subjective Well-Being Also Implicate Depression and Neuroticism." http://biorxiv.org/content/biorxiv/early/2015/11/24/032789.full.pdf

Boardman, Jason D. , Benjamin Domingue, W., and Jason Fletcher, M. . 2012a. "How Social and Genetic Factors Predict Friendship Networks." *Proceedings of the National Academy of Sciences* 109(43):17377-17381.

Boardman, Jason D., Casey L. Blalock, and Fred C. Pampel. 2010. "Trends in the Genetic Influences on Smoking." *Journal of Health and Social Behavior* 51(1):108-123.

Boardman, Jason D., Jonathan Daw, and Jeremy Freese. 2013. "Defining the Environment in Gene-Environment Research: Lessons from Social Epidemiology." *American Journal of Public Health* 103(1):S64-S72.

Boardman, Jason D., Benjamin W. Domingue, Casey L. Blalock, Brett C. Haberstick, Kathleen M. Harris, *et al.* 2014. "Is the Gene-Environment Interaction Paradigm Relevant to Genome-Wide Studies? The Case of Education and Body Mass Index." *Demography* 51(1):119-139.

Boardman, Jason D., Benjamin W. Domingue, and Jonathan Daw. 2015. "What Can Genes Tell Us About the Relationship between Education and Health?" *Social Science and Medicine* 127:171-80.

Boardman, Jason D., Michael E. Roettger, Benjamin W. Domingue, Matthew B. McQueen, Brett C. Haberstick, *et al.* 2012b. "Gene–Environment Interactions Related to Body Mass: School Policies and Social Context as Environmental Moderators." *Journal of Theoretical Politics* 24(3):370-388.

Branigan, Amelia R., Kenneth J. McCallum, and Jeremy Freese. 2013. "Variation in the Heritability of Educational Attainment: An International Meta-Analysis." *Social Forces* 92(1):109-140.

Caspi, Avshalom, Antony Braithwaite, Richie Poulton, Karen Sugden, Terrie E. Moffitt, *et al.* 2003. "Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-Htt Gene." *Science (New York, N.Y.)* 301(5631):386-389.

Caspi, Avshalom, Ahmad R. Hariri, Andrew Holmes, Rudolf Uher, and Terrie E. Moffitt. 2010. "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* 167(5):509-527.

Caspi, Avshalom, Joseph McClay, Terrie E. Moffitt, Jonathan Mill, Judy Martin, *et al.* 2002. "Role of Genotype in the Cycle of Violence in Maltreated Children." *Science* 297(5582):851-854.

Chabris, Christopher F., Benjamin M. Hebert, Daniel J. Benjamin, Jonathan P. Beauchamp, David Cesarini, *et al.* 2012. "Most Reported Genetic Associations with General Intelligence Are Probably False Positives." *Psychological Science* 23(11):1314-1323.

Charney, Evan and William English. 2012. "Candidate Genes and Political Behavior." *American Political Science Review* 106(1):1-34.

Cole, Steve W., Louise C. Hawkley, Jesusa M. Arevalo, Caroline Y. Sung, Robert M. Rose, *et al.* 2007. "Social Regulation of Gene Expression in Human Leukocytes." *Genome Biology* 8(9):R189-R189.

Cole, Steven W. 2013. "Social Regulation of Human Gene Expression: Mechanisms and Implications for Public Health." *American Journal of Public Health* 103(1):S84-S92.

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

Conley, Dalton, Benjamin W. Domingue, David Cesarini, Christopher Dawes, Cornelius A. Rietveld , *et al.* 2015. "Is the Effect of Parental Education on Offspring Biased or Moderated by Genotype?" *Sociological Science* 2:82-105.

Conley, Dalton and Emily Rauscher. 2013. "Genetic Interactions with Prenatal Social Environment: Effects on Academic and Behavioral Outcomes." *Journal of Health and Social Behavior* 54(1):109-127.

Conley, Dalton, Emily Rauscher, Christopher Dawes, Patrik K. E. Magnusson, and Mark L. Siegal. 2013a. "Heritability and the Equal Environments Assumption: Evidence from Multiple Samples of Misclassified Twins." *Behavior Genetics* 43(5):415-426.

Conley, Dalton, Emily Rauscher, and Mark L. Siegal. 2013b. "Beyond Orchids and Dandelions: Testing the 5-Htt "Risky" Allele for Evidence of Phenotypic Capacitance and Frequency-Dependent Selection." *Biodemography and Social Biology* 59(1):37-56.

Consortium, International Human Genome Sequencing. 2001. "Initial Sequencing and Analysis of the Human Genome (Vol 409, Pg 860, 2001)." *Nature* 412(6846):565-566.

—. 2004. "Finishing the Euchromatic Sequence of the Human Genome." *Nature* 431(7011):931-945.

Consortium, The International HapMap. 2005. "A Haplotype Map of the Human Genome." *Nature* 437(7063):1299-1320.

Cornelis, Marilyn C., Eric J. Tchetgen Tchetgen, Liming Liang, Lu Qi, Nilanjan Chatterjee, *et al.* 2012. "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* 175(3):191-202.

Davies, G., M. Luciano, K. McGhee, L. Lopez, A. J. Gow, *et al.* 2011. "Genome-Wide Association Studies Establish That Human Intelligence Is Highly Heritable and Polygenic." *Molecular Psychiatry* 16(10):996-1005.

Davies, G, R. E. Marioni, D. C. Liewald, W. D. Hill, S. P. Hagenaars, S. E. Harris, S. J. Ritchie, M. Luciano, C. Fawnsritchie, D. Lyall, B. Cullen, S. R. Cox, C. Hayward, D. J. Porteous, J. Evans, A. M. Mcintosh, J. Gallacher, N. Craddock, J. P. Pell, D. J. Smith, C. R. Gale, and I. J. Deary. 2016. "Genome-wide Association Study of Cognitive Functions and Educational Attainment in UK Biobank (N=112151)." *Molecular Psychiatry.* DOI:10.1038/mp.2016.45

Daw, Jonathan, Michael Shanahan, Kathleen Mullan Harris, Andrew Smolen, Brett Haberstick, *et al.* 2013. "Genetic Sensitivity to Peer Behaviors: 5httlpr, Smoking, and Alcohol Consumption." *Journal of Health and Social Behavior* 54(1):92-108.

Deary, Ian J., Janie Corley, Paul Redmond, Helen C. Fox, Suzanne J. Rowe, *et al.* 2012. "Genetic Contributions to Stability and Change in Intelligence from Childhood to Old Age." *Nature* 482(7384):212-215.

Demerath, Ellen W., Bradford Towne, Audrey C. Choh, William Johnson, Joanne E. Curran, *et al.* 2013. "The Positive Association of Obesity Variants with Adulthood Adiposity Strengthens over an 80-Year Period: A Gene-by-Birth Year Interaction." *Human Heredity* 75(2-4):175-185.

Domingue, Benjamin W, Daniel W. Belsky, Dalton Conley, Kathleen Harris, and Jason D Boardman. 2015. "Polygenic Influence on Educational Attainment: New Evidence from the National Longitudinal Study of Adolescent to Adult Health." *AERA Open* 1(3):1-13.

Domingue, Benjamin W., Daniel W. Belsky, Kathleen Mullan Harris, Andrew Smolen, Matthew B. McQueen, *et al.* 2014a. "Polygenic Risk Predicts Obesity in Both White and Black Young Adults." *PloS one* 9(7):e101596.

Domingue, Benjamin W., Jason Fletcher, Dalton Conley, and Jason D. Boardman. 2014b. "Genetic and Educational Assortative Mating among Us Adults." *Proceedings of the National Academy of Sciences* 111(22):7996-8000.

Duncan, Laramie E. and Matthew C. Keller. 2011. "A Critical Review of the First 10 Years of Candidate Gene-by-Environment Interaction Research in Psychiatry." *American Journal of Psychiatry* 168(10):1041-1049.

Fletcher, Jason M. and Dalton Conley. 2013. "The Challenge of Causal Inference in Gene-Environment Interaction Research: Leveraging Research Designs from the Social Sciences." *American journal of public health* 103(1):S42-S45.

Fletcher, Jason M. and Steven F. Lehrer. 2009. "Using Genetic Lotteries within Families to Examine the Causal Impact of Poor Health on Academic Achievement." National Bureau of Economic Research.

Frayling, Timothy M., Nicholas J. Timpson, Michael N. Weedon, Eleftheria Zeggini, Rachel M. Freathy, *et al.* 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.

Frazer, Kelly A., Dennis G. Ballinger, David R. Cox, David A. Hinds, Laura L. Stuve, *et al.* 2007. "A Second Generation Human Haplotype Map of over 3.1 Million Snps." *Nature* 449(7164):851-861.

Freese, Jeremy. 2008. "Genetics and the Social Science Explanation of Individual Outcomes." *The American Journal of Sociology* 114(S1):S1-S35.

—. 2011. "Integrating Genomic Data and Social Science." *Politics and the Life Sciences* 30(2):88-92.

Garcia-Closas, Montserrat, Nathaniel Rothman, Jonine D. Figueroa, Ludmila Prokunina-Olsson, Summer S. Han, *et al.* 2013. "Common Genetic Polymorphisms Modify the Effect of Smoking on Absolute Risk of Bladder Cancer." *Cancer Research* 73(7):2211-2220.

Ginsberg, Jeremy, Mark S. Smolinski, Larry Brilliant, Rajan S. Patel, Matthew H. Mohebbi, *et al.* 2009. "Detecting Influenza Epidemics Using Search Engine Query Data." *Nature* 457(7232):1012-1014.

Goldberger, Arthur S. 1979. "Heritability." *Economica* 46(184):327-347.

Golder, Scott A. and Michael W. Macy. 2011. "Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength across Diverse Cultures." *Science* 333(6051):1878-1881.

Guo, G., Y. Y. Tong, and T. J. Cai. 2008a. "Gene by Social Context Interactions for Number of Sexual Partners among White Male Youths: Genetics-Informed Sociology." *American journal of sociology* 114(3):S36-S66.

Guo, Guang , Hexuan Liu, Ling Wang, Haipeng Shen, and Wen Hu. 2015a. "The Genome-Wide Influence on Human Bmi Depends on Physical Activity, Life-Course, and Historical Period." *Demography* 52(5):1651-1670.

Guo, Guang, Yi Li, Tianji Cai, Hongyu Wang, and Greg J. Duncan. 2015b. "Peer Influence, Genetic Propensity, and Binge Drinking: A Natural Experiment and a Replication." *American Journal of Sociology* 121(3):914-954.

Guo, Guang, Michael E. Roettger, and Tianji Cai. 2008b. "The Integration of Genetic Propensities into Social-Control Models of Delinquency and Violence among Male Youths." *American Sociological Review* 73(4):543-568.

Guo, Guang, Michael E. Roettger, and Jean C. Shih. 2007. "Contributions of the Dat1 and Drd2 Genes to Serious and Violent Delinquency among Adolescents and Young Adults." *Human Genetics* 121(1):125-136.

Guo, Guang and Elizabeth Stearns. 2002. "The Social Influences on the Realization of Genetic Potential for Intellectual Development." *Social Forces* 80(3):881-910.

Hindorff, L. A. , J (European Bioinformatics Institute) MacArthur, J (European Bioinformatics Institute) Morales, H. A. Junkins, P. N. Hall, *et al.* "A Catalog of Published Genome-Wide Association Studies.",www.genome.gov/gwastudies.

Hirschhorn, Joel N. and Mark J. Daly. 2005. "Genome-Wide Association Studies for Common Diseases and Complex Traits." *Nature Reviews Genetics* 6(2):95-108.

Hutter, Carolyn M., Leah E. Mechanic, Nilanjan Chatterjee, Peter Kraft, and Elizabeth M. Gillanders on behalf of the NCI Gene-Environment Think Tank. 2013. "Gene-Environment Interactions in Cancer Epidemiology: A National Cancer Institute Think Tank Report." *Genetic Epidemiology* 37(7):643-657.

Jaffee, S. R. and T. S. Price. 2007. "Gene-Environment Correlations: A Review of the Evidence and Implications for Prevention of Mental Illness." *Molecular Psychiatry* 12(5):432-442.

Knight, Jennifer M., J. Douglas Rizzo, Brent R. Logan, Tao Wang, Jesusa M. G. Arevalo, *et al.* 2016. "Low Socioeconomic Status, Adverse Gene Expression Profiles, and Clinical Outcomes in Hematopoietic Stem Cell Transplant Recipients." *Clinical Cancer Research* 22(1):69-78.

Kumar, Siddharth Krishna, Marcus W. Feldman, David H. Rehkopf, and Shripad Tuljapurkar. 2016. "Limitations of GCTA as A Solution to the Missing Heritability Problem." *Proceedings of the National Academy of Sciences* 113(1):E61-E70.

Landecker, Hannah and Aaron Panofsky. 2013. "From Social Structure to Gene Regulation, and Back: A Critical Introduction to Environmental Epigenetics for Sociology." *Annual Review of Sociology* 39:333-357.

Lee, S. H., J. Yang, M. E. Goddard, P. M. Visscher, and N. R. Wray. 2012a. "Estimation of Pleiotropy between Complex Diseases Using Single-Nucleotide Polymorphism-Derived Genomic Relationships and Restricted Maximum Likelihood." *Bioinformatics* 28(19):2540-2542.

Lee, S. Hong, Teresa R. DeCandia, Stephan Ripke, Jian Yang, Patrick F. Sullivan, *et al.* 2012b. "Estimating the Proportion of Variation in Susceptibility to Schizophrenia Captured by Common Snps." *Nature Genetics* 44(3):247-U35.

Li, Yi, Hexuan Liu, and Guang Guo. 2015. "Does Marriage Moderate Genetic Effects on Delinquency and Violence?" *Journal of Marriage and Family* 77(5):1217-1233.

Liang, Liming, A. G. Willis-Owen Saffron, Catherine Laprise, C. C. Wong Kenny, Gwyneth A. Davies, *et al.* 2015. "An Epigenome-Wide Association Study of Total Serum Immunoglobulin E Concentration." *Nature* 520(7549):670-674.

Liu, Hexuan and Guang Guo. 2015. "Lifetime Socioeconomic Status, Historical Context, and Genetic Inheritance in Shaping Body Mass in Middle and Late Adulthood." *American Sociological Review* 80(4):705-737.

Liu, Hexuan, Yi Li, and Guang Guo. 2015. "Gene by Social-Environment Interaction for Youth Delinquency and Violence: Thirty-Nine Aggression-Related Genes." *Social Forces* 93(3):881-903.

Liu, Yun, Martin J. Aryee, Leonid Padyukov, M. Daniele Fallin, Espen Hesselberg, *et al.* 2013. "Epigenome-Wide Association Data Implicate DNA Methylation as an Intermediary of Genetic Risk in Rheumatoid Arthritis." *Nature biotechnology* 31(2):142-147.

Locke, Adam E., Bratati Kahali, Sonja I. Berndt, Anne E. Justice, Tune H. Pers, *et al.* 2015. "Genetic Studies of Body Mass Index Yield New Insights for Obesity Biology." *Nature* 518(7538):197-U401.

Michels, Karin B., Alexandra M. Binder, Sarah Dedeurwaerder, Charles B. Epstein, John M. Greally, *et al.* 2013. "Recommendations for the Design and Analysis of Epigenome-Wide Association Studies." *Nature Methods* 10(10):949-955.

Mitchell, C., J. Brooks-Gunn, I. Garfinkel, S. McLanahan, D. Notterman, *et al.* 2015. "Family Structure Instability, Genetic Sensitivity, and Child Well-Being." *American Journal of Sociology* 120(4):1195-1225.

Mitchell, C., J. Hobcraft, S. S. McLanahan, S. R. Siegel, A. Berg, *et al.* 2014. "Social Disadvantage, Genetic Sensitivity, and Children's Telomere Length." *Proceedings of the National Academy of Sciences of the United States of America* 111(16):5944-5949.

Mitchell, Colter, Sara McLanahan, Jeanne Brooks-Gunn, Irwin Garfinkel, John Hobcraft, *et al.* 2013. "Genetic Differential Sensitivity to Social Environments: Implications for Research." *American Journal of Public Health* 103(1):S102-S110.

Mitchell, Colter, Daniel Notterman, Jeanne Brooks-Gunn, John Hobcraft, Irwin Garfinkel, *et al.* 2011. "Role of Mother's Genes and Environment in Postpartum Depression."

Proceedings of the National Academy of Sciences of the United States of America 108(20):8189-8193.

Mitchell, Colter, Lisa M. Schneper, and Daniel A. Notterman. 2016. "DNA Methylation, Early Life Environment, and Health Outcomes." *Pediatric Research* 79(1):212-219.

Moor, M. H. M. de, P. T. Costa, A. Terracciano, R. F. Krueger, and E. J. C. de Geus et al. 2012. "Meta-analysis of Genome-wide Association Studies for Personality." *Molecular Psychiatry* 17(3):337-349.

Mukherjee, Bhramar, Jaeil Ahn, Stephen B. Gruber, and Nilanjan Chatterjee. 2012. "Testing Gene-Environment Interaction in Large-Scale Case-Control Association Studies: Possible Choices and Comparisons." *American Journal of Epidemiology* 175(3):177-190.

Newman, Timothy K., Yana V. Syagailo, Christina S. Barr, Jens R. Wendland, Maribeth Champoux, *et al.* 2005. "Monoamine Oxidase a Gene Promoter Variation and Rearing Experience Influences Aggressive Behavior in Rhesus Monkeys." *Biological Psychiatry* 57(2):167-172.

Nielsen, François. 2006. "Achievement and Ascription in Educational Attainment: Genetic and Environmental Influences on Adolescent Schooling." *Social Forces* 85(1):193-216.

Nielsen, François and J. Micah Roos. 2015. "Genetics of Educational Attainment and the Persistence of Privilege at the Turn of the 21st Century." *Social Forces* 94(2):535-561.

Nielsen, François. 2008. "The Nature of Social Reproduction: Two Paradigms of Social Mobility." Pp. 1-35 in *Sociologica*.

Notterman, Daniel A. and Colter Mitchell. 2015. "Epigenetics and Understanding the Impact of Social Determinants of Health." *Pediatric Clinics of North America* 62(5):1227-1240.

Ottman, Ruth. 1996. "Gene-Environment Interaction: Definitions and Study Designs." *Preventive Medicine* 25(6):764-770.

Perry, Brea L. 2016. "Gendering Genetics: Social and Biological Contingencies in the Protective Effects of Social Integration for Men and Women." *American Journal of Sociology* 121(6): 1655-1696.

Pescosolido, Bernice A., Brea L. Perry, J. Scott Long, Jack K. Martin, Jr John I. Nurnberger, *et al.* 2008. "Under the Influence of Genetics: How Transdisciplinarity Leads Us to Rethink Social Pathways to Illness." *American Journal of Sociology* 114(S1):S171-S201.

Preis, Tobias, Helen Susannah Moat, and H. Eugene Stanley. 2013. "Quantifying Trading Behavior in Financial Markets Using Google Trends." *Scientific Reports* 3(1684):1-6.

Purcell, Shaun M., Naomi R. Wray, Jennifer L. Stone, Peter M. Visscher, Michael C. O'Donovan, *et al.* 2009. "Common Polygenic Variation Contributes to Risk of Schizophrenia and Bipolar Disorder." *Nature* 460(7256):748-752.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, *et al.* 2007. "Plink: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." *The American Journal of Human Genetics* 81(3):559-575.

Qi, Q., A. Y. Chu, J. H. Kang, M. K. Jensen, G. C. Curhan, *et al.* 2012. "Sugar-Sweetened Beverages and Genetic Risk of Obesity." *New England Journal of Medicine* 367(15):1387-96.

Rakyan, Vardhman K., Thomas A. Down, David J. Balding, and Stephan Beck. 2011. "Epigenome-Wide Association Studies for Common Human Diseases." *Nature Reviews Genetics* 12(8):529-541.

Rietveld, Cornelius A., Arpana Agrawal, Johan G. Eriksson, Eva Albrecht, Behrooz Z. Alizadeh, *et al.* 2013a. "Gwas of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment." *Science* 340(6139):1467-1471.

Rietveld, Cornelius A., Robert F. Krueger, Meike Bartels, David A. Collier, Daniel J. Benjamin, *et al.* 2013b. "Molecular Genetics and Subjective Well-Being." *Proceedings of the National Academy of Sciences* 110(24):9692-9697.

Risch, Neil, Richard Herrell, Thomas Lehner, Kung-Yee Liang, Lindon Eaves, *et al.* 2009. "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* 301(23):2462-2471.

Sachidanandam, R., D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein, *et al.* 2001. "A Map of Human Genome Sequence Variation Containing 1.42 Million Single Nucleotide Polymorphisms." *Nature* 409(6822):928-933.

Shanahan, Michael J. 2013. "Social Genomics and the Life Course: Opportunities and Challenges for Multilevel Research." Pp. 255-76 in *Changing Social Processes and Social Contexts in an Aging Society*, edited by L. Waite: National Academy of Science.

Shanahan, Michael J. and Scott M. Hofer. 2005. "Social Context in Gene-Environment Interactions: Retrospect and Prospect." *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences* 60 Spec No 1(Special Issue 1):65-76.

—. 2011. "Molecular Genetics, Aging, and Well-Being: Sensitive Period, Accumulation, and Pathway Models." Pp. 135-147 in *Handbook of Aging and the Social Sciences, 7th Edition*, edited by R. H. B. a. L. K. George. New York: Elsevier.

Shanahan, Michael J., Stephen Vaisey, Lance D. Erickson, and Andrew Smolen. 2008. "Environmental Contingencies and Genetic Propensities: Social Capital, Educational Continuation, and Dopamine Receptor Gene Drd2." *American Journal of Sociology* 114 Suppl(1):S260-S286.

Simons, Ronald L., Man Kit Lei, Steven R. H. Beach, Gene H. Brody, Robert A. Philibert, *et al.* 2011. "Social Environment, Genes, and Aggression: Evidence Supporting the Differential Susceptibility Perspective." *American Sociological Review* 76(6):883-912.

Simonson, Matthew A., Amanda G. Wills, Matthew C. Keller, and Matthew B. McQueen. 2011. "Recent Methods for Polygenic Analysis of Genome-Wide Data Implicate an Important Effect of Common Variants on Cardiovascular Disease Risk." *BMC Medical Genetics* 12(1):146-146.

Speliotes, Elizabeth K., Reedik Mägi, Thomas A. Buchanan, Joshua C. Randall, Sailaja Vedantam, *et al.* 2010. "Association Analyses of 249,796 Individuals Reveal 18 New Loci Associated with Body Mass Index." *Nature Genetics* 42(11):937-48.

Thomas, Duncan C., Juan Pablo Lewinger, Cassandra E. Murcray, and W. James Gauderman. 2012. "Invited Commentary: Ge-Whiz! Ratcheting Gene-Environment Studies up to the Whole Genome and the Whole Exposome." *American Journal of Epidemiology* 175(3):203-207.

Thorleifsson, G., C. N. Palmer, T. Johnson, J. Luan, J. H. Park, *et al.* 2010. "Hundreds of Variants Clustered in Genomic Loci and Biological Pathways Affect Human Height." *Nature* 467(7317):832-838.

Tielbeek ,Jorim J., Sarah E. Medland, Beben Benyamin, Enda M. Byrne, Andrew C. Heath, Pamela A. F. Madden, Nicholas G. Martin, Naomi R. Wray, and Karin J. H. Verweij.2012." Unraveling the Genetic Etiology of Adult Antisocial Behavior: A Genome-Wide Association Study." PloS One 7(10): e45086. doi: 10.1371/journal.pone.0045086.

Tung, Jenny, Luis B. Barreiro, Zachary P. Johnson, Kasper D. Hansen, Vasiliki Michopoulos, *et al.* 2012. "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* 109(17):6490-6495.

Turkheimer, Eric, Andreana Haley, Mary Waldron, Brian D'Onofrio, and Irving I. Gottesman. 2003. "Socioeconomic Status Modifies Heritability of Iq in Young Children." *Psychological Science* 14(6):623-628.

Vinkhuyzen, A. A. E., M. Luciano, A. Payton, M. Horan, W. Ollier, *et al.* 2012. "Common Snps Explain Some of the Variation in the Personality Dimensions of Neuroticism and Extraversion." *Translational psychiatry* 2(4):e102.

Visscher, Peter M., Matthew A. Brown, Mark I. McCarthy, and Jian Yang. 2012. "Five Years of Gwas Discovery." *American Journal of Human Genetics* 90(1):7-24.

Wagner, Brandon, Jiang Li, Hexuan Liu, and Guang Guo. 2013. "Gene-Environment Correlation: Difficulties and a Natural Experiment-Based Strategy." *American Journal of Public Health* 103 Suppl 1(Journal Article):S167-73.

Wang, Ling, Haipeng Shen, Hexuan Liu, and Guang Guo. 2015. "Mixture Snps Effect on Phenotype in Genome-Wide Association Studies." *BMC Genomics* 16(3):1-11.

Wang, Zhong, Mark Gerstein, and Michael Snyder. 2009. "Rna-Seq: A Revolutionary Tool for Transcriptomics." *Nature Reviews Genetics* 10(1):57-63.

Wood, A. R., T. Esko, J. Yang, S. Vedantam, T. H. Pers, *et al.* 2014. "Defining the Role of Common Variation in the Genomic and Biological Architecture of Adult Human Height." *Nature Genetics* 46(11):1173-1186.

Wray, Naomi R., Jian Yang, Ben J. Hayes, Alkes L. Price, Michael E. Goddard, *et al.* 2013. "Pitfalls of Predicting Complex Traits from Snps." *Nature reviews. Genetics* 14(7):507-515.

Wu, Chen, Peter Kraft, Kan Zhai, Jiang Chang, Zhaoming Wang, *et al.* 2012. "Genome-Wide Association Analyses of Esophageal Squamous Cell Carcinoma in Chinese Identify Multiple Susceptibility Loci and Gene-Environment Interactions." *Nature Genetics* 44(10):1090-1097.

Yang, J., B Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, *et al.* 2010. "Common SNPs Explain a Large Proportion of the Heritability for Human Height." *Nature Genetics* 42(7):565-569.

Yang, Jian, M. Geoffrey Hayes, William G. Hill, Maria Teresa Landi, Alvaro Alonso, *et al.* 2011. "Genome Partitioning of Genetic Variation for Complex Traits Using Common Snps." *Nature Genetics* 43(6):519-525.

Yang, Jian, S.Hong Lee, Naomi R. Wray, Michael E. Goddard, and Peter Visscher. 2016. "Commentary on 'Limitations of GCTA as A Solution to the Missing Heritability Problem' ". BioRxiv, http://biorxiv.org/content/early/2016/01/20/036574.abstract

Yehuda, Rachel, Nikolaos P. Daskalakis, Linda M. Bierer, Heather N. Bader, Torsten Klengel, Florian Holsboer, and Elisabeth B. Binder. 2015. " Holocaust Exposure Induced Intergenerational Effects on FKBP5 Methylation." *Biological Psychiatry*. DOI: http://dx.doi.org/10.1016/j.biopsych.2015.08.005

CHAPTER 2: INTERPLAY OF SOCIAL ENVIRONMENT AND THE WHOLE GENOME ON DELINQUENCY AND VIOLENCE

2.1 Introduction

Previous studies have shown that gene-environment interplay contributes to a variety of behavioral and social outcomes (Boardman *et al.* 2012; Caspi *et al.* 2002; Fowler *et al.* 2011; Guo *et al.* 2008; Pescosolido *et al.* 2008; Shanahan *et al.* 2008; Simons *et al.* 2011). Yet these studies have typically focused on one or only a few candidate genes at a time. The aim of this research is to provide a more comprehensive view of the gene-environment interplay by incorporating genome-wide genotype data; particularly, to show how the social environment moderates genetic associations with youth delinquent and violent behaviors.

Traits determined by a single gene or allele are rare in human beings (Glazier *et al.* 2002). The vast majority of human diseases (e.g., cancer, heart disease, and diabetes) are complex traits affected by a large number of genes (Crabbe 2002; Plomin *et al.* 2001). Likewise, almost all human traits of interest to social scientists are complex, such as personality, cognition, motivation, and health behaviors. These traits are likely the consequence of many genetic and environmental factors, as well as interactions among them (Hirschhorn and Daly 2005; Lander and Botstein 1986; Lander and Schork 1994). Therefore it is important to incorporate multi-genetic and multi-environmental factors in gene-environment interaction (G×E) research on complex social outcomes

This study focuses on influences of the whole genome on delinquency. It assesses genomic contributions to youth delinquency and violence using a recently developed mixed linear model approach in genomics studies that simultaneously accounts for a large number of genetic variables in a single regression analysis (Yang *et al.* 2011b). Moreover, it investigates the extent to which the social environment moderates the genomic contribution to delinquency. To achieve this, I compare genomic contributions to delinquency and violence between individuals exposed to environments with lower levels of social control and those who were exposed to environments with higher levels of social control (e.g., low parental attachment versus high parental attachment; loose school discipline versus strict school discipline; and disadvantaged neighborhoods versus non-disadvantaged neighborhoods). This study provides evidence that the genetic risk for adolescent delinquency and violence is largely context-dependent: the genetic risk is amplified among individuals under low-social-control (LSC) conditions, but suppressed among those under high-social-control (HSC) conditions.

2.2 Background

2.2.1 Social regulation of gene expression

Genetic factors affect but do not determine human behavior, and their impact is moderated by the environment (Freese 2008; Rutter *et al.* 2006; Shanahan and Boardman 2009). Findings from both animal and human studies provide evidence that changes in environmental conditions induce gene expression to varying degrees.

Animal studies have demonstrated the influence of social experiences on gene expression using rhesus monkeys. As shown by Suomi and colleagues, infant monkeys reared by their mothers, compared to those reared by other adults, differ in the effect of *5-HTTLPR* (the serotonin transporter gene-linked polymorphic region) on cerebrospinal fluid concentrations of

5-hydroxyindoleacetic acid (Bennett *et al.* 2002), ethanol consumption among females (Barr *et al.* 2004b), and adreno-corticotropic hormone (ACTH) levels (Barr *et al.* 2004a). In an assessment of genetic and environmental influences on aggressive behavior, the expression of *MAOA* (the monoamine oxidase A) is found to be sensitive to early social experiences (Newman *et al.* 2005). Findings from a more recent study show that a female macaque's ranking within her social environment affects the expression of nearly 700 genes (Tung *et al.* 2012).

In human studies, there is increasing evidence of social regulation of gene expression. Cole *et al.* (2007) find that people who experienced chronic social isolation (e.g., loneliness) and those who experienced consistent social integration systematically differ in the expression of more than 200 genes in white blood cells. In a more recent study, Cole *et al.* (2010) identify a genetic polymorphism (*IL6* – 174G/C) that interacts with adverse socioenvironmental conditions to increase chronic inflammation and mortality among older adults. Specifically, during periods of significant life adversity, people who bear the G allele of the polymorphism tend to show increased IL6 gene transcription and increased risk of inflammation-related diseases and mortality. Additionally, there is evidence that gene expression is subject to other social conditions such as low socioeconomic status and low social support. Chen *et al.* (2009) investigate the association between low socioeconomic status (SES) and gene expression in children with asthma. Their study shows that children with asthma from a low SES background, relative to those from high SES background, tend to show overexpression of gene regulating inflammatory processes. Moreover, in a study of women with ovarian cancer, 266 genes are shown to be differentially expressed in tumors from women with low social support and high depressive symptoms (Lutgendorf *et al.* 2009).

2.2.2 Gene-environment interaction for delinquency

As discussed above, the social environment plays a crucial role in the expression of genes related to various phenotypes. Of greater importance for the purpose of this research is social regulation of the expression of the genes related to delinquency. The *environmental triggering/suppressing* perspective offers important contributions to our understanding of how the social environment and genetic factors interactively affect delinquent and violent behaviors.

There are two components to the *environmental triggering/suppressing* perspective. First, adverse environments are likely to "*trigger*" the expression of risk alleles (Shanahan and Hofer 2005). This *"triggering"* mechanism is well demonstrated in the works of Caspi and colleagues (Caspi *et al.* 2003; Caspi *et al.* 2002), and is referred to as the *diathesis stress* model (Ellis *et al.* 2011). Central to this *"triggering"* mechanism is the coaction of the risky allele and the risky environment. For example, Caspi *et al.* (2002) identify an association between the *MAOA* genotypes and antisocial behaviors, but mainly among test subjects who experience childhood maltreatment. Second, favorable environments may suppress the expression of risk alleles. Particularly, social norms and structural constraints can inhibit individuals' behavior and choices, thereby reducing genetic effects (Shanahan and Hofer 2005). As shown by Pescosolido *et al.* (2008), the association between gamma-aminobutyric acid receptor subunit alpha-2 (*GABRA2*) and alcoholism is reduced by family support. Similarly, *DRD2* (dopamine D2 receptor) is found to contribute less to delinquency among male youths who have regular meals with a parent and those who live with both biological parents, compared respectively to those who do not have regular meals with a parent and those who do not live with both biological parents (Guo *et al.* 2008).

Most of these studies focus on a single or only a few genetic variants at a time (Beaver *et al.* 2008; Caspi *et al.* 2002; Foley *et al.* 2004; Guo *et al.* 2007; Kim-Cohen *et al.* 2006; Simons *et al.* 2011; Vanyukov *et al.* 2007). However, delinquent and violent behaviors are complex human traits that can be affected by a large number of genetic factors with small to moderate effects. Therefore, it is crucial to investigate the collective contribution of multi-genetic factors to delinquency and violence.

2.2.3 Social moderators for delinquency

In this chapter, I focus on interaction of genetic variants and three important social institutions in childhood or adolescence: the family, the school, and the neighborhood. These social institutions not only contribute to inhibiting or reducing children's deviant acts, but also have long-term impact on their development of characteristics relevant to future delinquency or crime (Gottfredson and Hirschi 1990; Hirschi 1969; Sampson and Laub 1993). Of particular interest to this study are the roles of these institutions in shaping individual propensity or self-control that can have persistent influence over the life course.

Parenting factors, such as parental attachment and supervision, are the most important source of self-control. According to Gottfredson and Hirschi (1990), self-control is cultivated during early childhood through careful rearing and effective discipline, whereas low self-control is mainly attributed to ineffective parenting. That is, if the caregivers of a child neglect to monitor his/her behavior, fail to recognize his/her deviant behaviors or punish such behaviors, as a consequence, the child may lack the ability to delay gratification, be insensitive to others' needs and interests, as well as be unwilling to accept restrictions on his/her behavior, and become more likely to use forcible or violent means to achieve his/her ends. Cullen *et al.* (2008) summarize results from 13 empirically studies examining the relationship between self-control

and various dimensions of parenting factors. Twelve of the 13 studies have provided evidence that less effective parenting is associated with weaker self-control.

School is another powerful social institution in the development self-control (Gottfredson and Hirschi 1990). Because the school has a particular interest in maintaining a good educational environment, it is expected to recognize and prevent antisocial behavior and it has the authority and means to implement effective discipline. As Denise Gottfredson (2001) suggests, "schools have the potential to teach self-control and to engage informal social controls to hold youthful behavior in check." Turner *et al.* (2005) show that the influence of school socialization on self-control is more effective for children of parents who failed in their task to teach self-control. Accordingly, school socialization may work to "pick up the slack" for inadequate parenting practices. This is consistent with the study of Meldrum (2008), in which self-control is found to be significantly predicted by school monitoring, even after controlling for familial factors.

In addition to family and school, neighborhood conditions are also critical in shaping self-control. Wikström and Sampson (2003) propose that individuals with weaker self-control are more likely to be found in disadvantaged neighborhoods with weak community capital and low collective efficacy (i.e., weak social cohesion among neighbors and their expectations to achieve common good), because these neighborhoods often lack resources and services, such as time, money, and knowledge, to support familial socialization practices. Empirical studies have provided mixed support for this position. Pratt *et al.* (2004) provide evidence that self-control is predicted by neighborhood conditions. In a more recent study, Gibson *et al.* (2010) also found support for associations between neighborhood structural characteristics and self-control, but these associations became nonsignificant after taking into account individual-level characteristics.

In summary, prior studies have demonstrated associations among the social environment, delinquency, and self-control. Although they do not directly address genetic factors, these studies are consistent with the G×E interaction view that the social environment may moderate individual propensities for delinquency. From the environmental triggering/suppressing perspective, I hypothesize that *genetic risks for delinquency and violence are greater among individuals who were weakly attached to parents and schools, loosely disciplined by parents or school authorities, or lived in disadvantaged neighborhoods than those who were closely attached to their parents and schools, strictly disciplined by parents or school authorities, or lived in non-disadvantaged neighborhoods.* This study extends previous G×E research by using genome-wide genotype data.

2.3 Data and Measurement

2.3.1 Data

Data for this study come from the National Longitudinal Study of Adolescent Health (Add Health). Add Health is a longitudinal survey of U.S. adolescents in grades 7 through 12 from 1994 to 1995 (In-School, N = 90,118; Wave I, N = 20,745). The Add Health cohort was followed up in 1996 (Wave II, N = 14,738) and again from 2001 to 2002 (Wave III, N = 15,197) (Harris, Florey, and Tabor et al.2003). In total 8,266 samples (3,831 males and 4,435 females) were put into genotyping production using the Illumina Human Omni-1.0 Quad beadchip and 2311 (1,156 males and 1,155 females) using Illumina Human Omni-2.5 Quad beadchip. Genotype samples from 9,695 individuals passed quality control. There are 629,757 autosomal SNPs with a missing rate less than 10%. To minimize confounding effects of population stratification, this study focuses on the whites. Since the genomic-relatedness-matrix restricted maximum likelihood (GREML) method used to analyze the genome-wide data requires to

exclude genetically related individuals to obtain unbiased results, I removed pairs with genetic relation greater than .025. As a result, 4972 genetic unrelated individuals compose the analytical sample of this study.

2.3.2 Variable measurement

Outcome variables: serious delinquency and violence scores. The outcome variables are based on 11 items from Add Health questionnaires at Wave III: (1) deliberately damaged others' property, (2) so badly hurt someone that medical treatment was needed, (3) used a weapon to get something from someone, (4) took part in group fights, (5) pulled a knife or gun on someone, (6) shot or stabbed someone, (7) took part in fights in which self was injured, (8)stole something worth more than $50, (9) broke into a house or building to steal, (10) sold drugs, and (11) stole something worth less than $50. (Cronbach's alpha = .69). To be consistent with the delinquency literature (Hagan and Foster 2003; Hannon 2003), I divided the 11 questions into violent and nonviolent categories. The serious delinquency score is the average of all eleven items, with higher scores indicating greater delinquency. The violence score is the average of the first 7 items. I chose outcomes from a single wave because the analytic model does not allow repeated measures. Also, I used outcomes measured at Wave III and social-environmental measures from Wave I to minimize reverse causality.

Socioenvironmental variables: Parenting factors. To simplify the G×E analysis, I constructed each social-environmental variable as a dichotomous variable. I assessed *parental attachment* using two Wave I questions asking how close a respondent felt to his or her mother and father and a question concerning the respondent's feeling about how his or her parents cared about him or her (alpha = .72). If the average of a respondent's answers to three questions was greater than or equal to the sample median, for him or her, *Parental attachment* was coded as 1,

indicating high parental attachment, and 0 otherwise (indicating low parental attachment). *Parental supervision* is constructed based on seven Wave I questions asking the respondent if his or her parents allowed him or her to make their decisions about the following: the time they must be home on weekend nights; the people they hang around with; what they wear; how much television they watch; which television programs they watch; what time they go to bed on week nights; and what they eat (alpha = .63). *Parental supervision* was coded as 1 if the average of a respondent's answers to seven questions was greater or equal to the sample median (indicating strict parental supervision), and 0 otherwise (indicating loose parental supervision).

School factors. I used two Wave I measures to assess school factors: *school attachment* and *school discipline*. To measure *school attachment*, I averaged responses to three questions (alpha = .77) asking whether a respondent (rated on a scale of 1 to 5) felt close to people at school, felt like being part of the school, or felt happy at school. I then constructed a school-level attachment variable by averaging *school attachment* of all students from each school. To measure *school discipline*, I averaged school administrators' responses to eleven questions (alpha = .87) asking in their schools what happens to a student who is caught the second time fighting with another student, injuring another student, possessing alcohol, possessing an illegal drug, possessing a weapon, drinking alcohol at school, using an illegal drug at school, smoking at school, verbally abusing a teacher, physically injuring a teacher, and stealing school property (1 = no policy; 2 = verbal warning; 3 = minor action; 4 = in-school suspension; 5 = out-of-school suspension;6 = expulsion). Like *parental attachment* and *parental supervision*, *school attachment* and *school discipline* were dichotomized on the basis of the sample median (coded as 1 if the average of the items was equal to or greater than the median, indicating high school

attachment and strict school discipline, and 0 otherwise, indicating low school attachment or loose school discipline).

Neighborhood. I assessed neighborhood environment using four Wave I block level variables from the Add Health Public Contextual Database: *proportion of aged 25+ individuals with college degree or more*, *proportion of households with income less than $15,000*, *unemployment rate* and *proportion of own children under 18 years in families and subfamilies not living with both parents*. Block is a geographic area defined by the U.S. Bureau of the Census, which in 1990, averaged 452 housing units or 1,100 people (U.S. Bureau of the Census 1993). It is the lowest level of geography in sample data published by the census bureau, and therefore captures the most localized available contextual characteristics of the areas in which individuals live (Billy, Wenzlow, and Grady 1998). I recoded each of the four variables into a 0-1 indicator. For example, the *unemployment* variable was coded as 1 if the *unemployment rate* of the block where the respondent lived was lower than or equal to the sample median (indicating non-disadvantaged neighborhoods).

Control variables. I controlled for biological sex, age, and age squared for all analyses of the genomic contribution to serious delinquency and violence. In addition, to account for potential population stratification, I adjusted all the analyses for the first ten principal components (PCs) computed from PLINK 1.9 (Chang *et al.* 2015; Price *et al.* 2010; Price *et al.* 2006).

2.4 Analytical Strategy

At the first stage of the analysis, I employed a mixed linear model to estimate the genomic contribution to delinquency. The model was fit using the Genome-wide Complex Trait

Analysis (GCTA) software package, a tool based on the work of Yang *et al.* (2011b) to estimate the overall genetic variance for complex human traits.

The mixed linear model offers the substantial advantage of simultaneously accounting for a large number of genetic variants. It was developed to address the "missing heritability" issue in genome-wide association studies (GWAS) (Yang *et al.* 2010). For example, whereas 80% of variance in human height is believed to be heritable, SNPs discovered by GWAS together can explain less than 10% of observed height variation (Visscher *et al.* 2012). In contrast to single-variant association analysis where each SNP is tested against an adjusted p-value (e.g., $5x10^{-8}$ or smaller), the mixed linear model approach treats all SNP effects as random effects. Using this approach, Yang *et al.* (2011a) show common SNPs collectively explain 41.9%, 15.9%, 25.4%, and 16.8% of the total phenotypic variances in human height, body mass index (BMI), von Willebrand factor (vWF), and OT interval (QTi), whereas highly significant and well replicated SNPs identified by GWAS merely account for 10%, 1.5%, 13%, and 7%, respectively. This method has also been employed for common diseases (Lee *et al.* 2011), schizophrenia (Lee *et al.* 2012), intelligence (Chabris *et al.* 2012; Davies *et al.* 2011), personality traits (Vinkhuyzen *et al.* 2012), subjective well-being (Rietveld *et al.* 2013), economic and political behaviors (Benjamin *et al.* 2012).

The model is described by the following equation:

$$Y=X\beta + W\mu + \varepsilon, \qquad \text{(Equation 2.1)}$$

where Y is the outcome variable; β is a vector of the coefficients of fixed effects such as age, sex and other controls; μ is a vector of SNP effects with $\mu_i \sim N(0, \sigma^2_\mu)$ where i = 1,…, I, with I being the number of SNPs; ε is a vector of residual effects with $\varepsilon_j \sim N(0, \sigma^2_\varepsilon)$ where j = 1,…, J, with J being the number of individuals; W is a standardized genotype matrix with the ij^{th} element w_{ij} =

$(s_{ij} - 2p_i)/\sqrt{[2p_i(1 - p_i)]}$ where s_{ij} is the number of copies of the reference allele for the i^{th} SNP of the j^{th} individual and p_i is the frequency of the reference allele.

Yang et al. (2010) innovatively applied a previous result that has been known in animal genetics (Goddard et al. 2009). The result defines $g = W\mu$, $\mathbf{A} = WW^T/I$ and $\sigma_g^2 = I\sigma_\mu^2$. Then Equation 2 is mathematically equivalent to Equation 1:

$$Y=X\beta + g + \varepsilon, \text{ with } Var(Y) = \mathbf{A}\sigma_g^2 + I_\varepsilon\sigma_\varepsilon^2, \qquad \text{(Equation 2.2)}$$

where g is a J*1 vector of the total genetic effects of the individuals with $g \sim N(0, \mathbf{A}\sigma_g^2)$, **A** is the genetic relationship matrix (GRM) between individuals and σ_g^2 is the total genetic variance explained by the SNPs. Hence σ_g^2 can be estimated by the restricted maximum likelihood (REML) approach, depending on the GRM estimated from all SNPs. In this study, the collective genetic contribution is assessed using the proportion of total variance in the outcome explained by all SNPs, which can be expressed as $\sigma_g^2/(\sigma_g^2 + \sigma_\varepsilon^2)$.

Next, I compared the genomic contribution to delinquency between individuals under LSC conditions and those under HSC conditions. I obtained residuals of linear regression models predicting Wave III delinquency/violence using, gender, age, age squared, Wave I delinquency/violence, and the first ten principal components. I split the residuals into two strata on the basis of each constructed dichotomous socioenvironmental variable (e.g., one stratum only includes individuals under LSC conditions and the other includes those under HSC conditions), and estimated the proportion of variance in the residuals explained by the SNPs.

2.5 Results

2.5.1 Genomic contribution to delinquency and violence

Table 2.2 displays the estimates of the genomic contribution to serious delinquency and violence. As can be seen, approximately 4% of the total variance in serious delinquency and 0%

of the total variance in violence is attributable to the SNPs in the analysis. In the face of G×E, I expect a greater genetic risk for individuals exposed to LSC environments, and a weaker effect for those who were exposed to HSC environments in the sample. Next, I tested whether the genomic contribution to serious delinquency and violence differs under LSC and HSC conditions.

2.5.2 Genomic contribution under differential conditions

Table 2.3 shows the results of comparing the genomic contributions to delinquency and violence under differential conditions. Columns 1 and 3 contain estimates of the genomic contribution to serious delinquency under HSC and LSC conditions, and columns 5 and 7 contain estimates for violence. In Table 2.3, most estimates of the genomic contribution under LSC conditions are greater than those under HSC conditions (with the exception of neighborhood education and income for violence). For example, the proportion of total variance in the serious delinquency score explained by whole-genome SNPs is estimated to be 25.8% for adolescents poorly attached to school, but the proportion drops to 0% for those who were closely/moderately attached to school. To summarize, there is evidence that the genetic risks for delinquency and violence are greater for adolescents who were weakly attached to parents and school, loosely disciplined by parents or school authorities, or lived in neighborhoods with higher unemployment levels and higher single-parent household rates as opposed to those who were closely attached to their parents and school, strictly disciplined by parents or school authorities, or lived in neighborhoods with lower unemployment rates and lower single-parent household rates.

2.5.3 Gene-environment correlations

Gene-environment correlation (rG) occurs when one's exposure to an environment depends upon his or her genotype. The existence of rG may confound the G×E effects (Caspi and Moffitt 2006; Jaffee and Price 2007; Wagner *et al.* 2013). I applied bivariate mixed linear models to detect rGs between the socio-environment variables and two outcomes. Table 2.4 shows that parental supervision and serious delinquency or violence are likely to be influenced by common SNPs, while there is no consistent evidence of rG for school- and neighborhood-level socioenvironmental variables.

2.5.4 Discussion and Conclusions

In this paper I hypothesize that high social control suppresses genetic risks for youth delinquency and violence, and low social control exacerbates the same risks. I examine the influences of crucial social institutions, such as the family, the school, and the neighborhood, on the collective contribution of the whole genome to delinquency and violence. Consistent with the *environmental triggering/suppressing* perspective, I find that favorable social conditions are associated with smaller genomic contribution, whereas adverse social conditions are associated with greater genomic contribution to adolescent delinquency and violence.

This study makes several important contributions to the G×E literature. First, this is the first studies assessing genetic risk of delinquent and violence behaviors using genome-wide data. This is a crucial improvement over previous research, which normally studies one genetic factor or only a few at a time. Delinquent and violent behaviors are complex human traits, meaning they could be affected by a large number of genetic and environmental factors. It is likely that the effects of many genetic variants are too small to be detected by testing each one individually

for an association with the phenotype. However, these variants, collectively, could make a substantial contribution.

Second, I find that the genomic influence is smaller under favorable conditions than adverse conditions. These findings are consistent with results in previous G×E research based on a one or a few genetic variants (Caspi *et al.* 2002; Guo *et al.* 2008; Pescosolido *et al.* 2008). What is more, the findings highlight the influence of social control on the genetic risks of many variants at the same time. These findings illuminate one mechanism through which social control affects delinquency and violence: it is possible that the presence of social control simultaneously prevents the expression of a large number of genetic variants associated with aggression and violence. In an environment under high social control, such as high family attachment, there may be adolescents varying in their genetic propensities for delinquent behaviors; some may possess risk alleles related to delinquency. Yet, the expression of risk alleles is prevented due to strong social control. When the control is weakened, for example, parents pay less attention, the adolescent with high genetic propensities for delinquency, relative to one with low genetic propensities, may be more apt to show gene expression.

The third contribution is methodological. I test G×E involving a large number of genetic variants. This method is an extension of the mixed linear model approach (Yang *et al.* 2011b). Compared to conventional linear regression models, the key advantage of the mixed linear model is its ability to simultaneously account for a large number of genetic variants. To illustrate, in conventional linear models, one socioenvironmental factor and the 629,757 SNPs would generate 629,757 two-way interaction terms in total. Analyses dependent on such models typically do not have sufficient statistical power to produce significant results. However, in the mixed linear model, being treated as random effects, SNPs across the whole genome could be

considered simultaneously. That allows us to estimate and compare the genomic influence under differential social conditions.

Although this study provides important insights to understanding how the social environment moderates genetic influence on delinquency and violence, some limitations should be noted. The mixed linear model approach does not allow genetically related individuals and repeated measures, leading to a reduction of the effective sample size. Also, because of the relatively small sample size, I have to dichotomize the social-environmental variables (if there were more categories, the G×E analysis would require a much larger sample to have sufficient statistical power), which might result in some loss of information. The mixed linear model assumes random SNP effects are uncorrelated with covariates in the model. Violation of this assumption may bias the results. Moreover, I do not include minority samples in the analysis. Although genetic information is available for some minorities (e.g., blacks and Asians), their sample size is insufficient to achieve adequate statistical power for separate G×E analysis. Future research can use more refined measures and extend the analysis in this study to other racial populations when data become available.

Despite these limitations, this study makes important contributions to social sciences. It underscores the significance of the dialogue between the biological and social sciences. Social scientists traditionally have assumed homogeneous human nature at birth and focused on social structural influences on individuals. However, there is growing evidence that the social environment modifies gene expression (Morgan *et al.* 2002; Norman *et al.* 2012), and genetic variability, in turn, affects individuals' responses to the environment (Freese 2008). Increasingly available molecular genetic data in large-scale datasets (e.g., Add Health, the Fragile Families Study, and the Health and Retirement Study) enable social scientists to investigate how

socioenvironmental factors shape human behavior through moderating the effects of a large number of genes. The conceptual framework and methodology in this study can be expanded to study other behavioral and social consequences of the complex interplay of multi-genetic and multi-environmental factors.

REFERENCES

Barr, Christina S., David Goldman, Stephen J. Suomi, J. Dee Higley, Timothy K. Newman, *et al.* 2004a. "Rearing Condition and Rh5-Httlpr Interact to Influence Limbic-Hypothalamic-Pituitary-Adrenal Axis Response to Stress in Infant Macaques." *Biological Psychiatry* 55(7):733-738.

Barr, Christina S., Timothy K. Newman, Stephen Lindell, Courtney Shannon, Maribeth Champoux, *et al.* 2004b. "Interaction between Serotonin Transporter Gene Variation and Rearing Condition in Alcohol Preference and Consumption in Female Primates." *Archives of General Psychiatry* 61(11):1146-1152.

Beaver, Kevin M., John Paul Wright, Matt DeLisi, and Michael G. Vaughn. 2008. "Desistance from Delinquency: The Marriage Effect Revisited and Extended." *Social Science Research* 37(3):736-752.

Benjamin, Daniel. J. , Magnus Johannesson, Peter Visscher, M., David Cesarini, Matthijs. J. H. M. van der Loos, *et al.* 2012. "The Genetic Architecture of Economic and Political Preferences." *Proceedings of the National Academy of Sciences* 109(21):8026-8031.

Bennett, A. J., J. D. Higley, K. P. Lesch, A. Heils, J. C. Long, *et al.* 2002. "Early Experience and Serotonin Transporter Gene Variation Interact to Influence Primate Cns Function." *Molecular Psychiatry* 7(1):118-122.

Boardman, Jason D. , Benjamin Domingue, W., and Jason Fletcher, M. . 2012. "How Social and Genetic Factors Predict Friendship Networks." *Proceedings of the National Academy of Sciences* 109(43):17377-17381.

Caspi, Avshalom and Terrie E. Moffitt. 2006. "Gene-Environment Interactions in Psychiatry: Joining Forces with Neuroscience." *Nature Reviews Neuroscience* 7(7):583-590.

Caspi, Avshalom, Antony Braithwaite, Richie Poulton, Karen Sugden, Terrie E. Moffitt, *et al.* 2003. "Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-Htt Gene." *Science (New York, N.Y.)* 301(5631):386-389.

Caspi, Avshalom, Joseph McClay, Terrie E. Moffitt, Jonathan Mill, Judy Martin, *et al.* 2002. "Role of Genotype in the Cycle of Violence in Maltreated Children." *Science* 297(5582):851-854.

Chabris, Christopher F., Benjamin M. Hebert, Daniel J. Benjamin, Jonathan P. Beauchamp, David Cesarini, *et al.* 2012. "Most Reported Genetic Associations with General Intelligence Are Probably False Positives." *Psychological Science* 23(11):1314-1323.

Chang, Christopher C., Carson C. Chow, Laurent Cam Tellier, Shashaank Vattikuti, Shaun M. Purcell, *et al.* 2015. "Second-Generation Plink: Rising to the Challenge of Larger and Richer Datasets." *GigaScience* 4(1):7.

Chen, E., G. E. Miller, H. A. Walker, J. M. Arevalo, C. Y. Sung, *et al.* 2009. "Genome-Wide Transcriptional Profiling Linked to Social Class in Asthma." *Thorax* 64(1):38-43.

Cole, Steve W., Louise C. Hawkley, Jesusa M. Arevalo, Caroline Y. Sung, Robert M. Rose, *et al.* 2007. "Social Regulation of Gene Expression in Human Leukocytes." *Genome Biology* 8(9):R189-R189.

Cole, Steven W., Jesusa M. G. Arevalo, Rie Takahashi, Erica K. Sloan, Susan K. Lutgendorf, *et al.* 2010. "Computational Identification of Gene-Social Environment Interaction at the Human Il6 Locus." *Proceedings of the National Academy of Sciences of the United States of America* 107(12):5681-5686.

Crabbe, John C. 2002. "Genetic Contributions to Addiction." *Annual review of psychology* 53:435-462.

Cullen, Francis T, James D Unnever, John Paul Wright, and Kevin M Beaver. 2008. "Parenting and Self-Control." Pp. 61-74 in *Out of Control: Assessing the General Theory of Crime*.

Davies, G., M. Luciano, K. McGhee, L. Lopez, A. J. Gow, *et al.* 2011. "Genome-Wide Association Studies Establish That Human Intelligence Is Highly Heritable and Polygenic." *Molecular Psychiatry* 16(10):996-1005.

Ellis, Bruce J., W. Thomas Boyce, Jay Belsky, Marian J. Bakermans-Kranenburg, and Marinus H. van Ijzendoorn. 2011. "Differential Susceptibility to the Environment: An Evolutionary–Neurodevelopmental Theory." *Development and Psychopathology* 23(1):7-28.

Foley, Debra L., Lindon J. Eaves, Brandon Wormley, Judy L. Silberg, Hermine H. Maes, *et al.* 2004. "Childhood Adversity, Monoamine Oxidase a Genotype, and Risk for Conduct Disorder." *Archives of General Psychiatry* 61(7):738-744.

Fowler, James H., Jaime E. Settle, and Nicholas A. Christakis. 2011. "Correlated Genotypes in Friendship Networks." *Proceedings of the National Academy of Sciences of the United States of America* 108(5):1993-1997.

Freese, Jeremy. 2008. "Genetics and the Social Science Explanation of Individual Outcomes." *The American Journal of Sociology* 114(S1):S1-S35.

Gibson, Chris L, Christopher J Sullivan, Shayne Jones, and Alex R Piquero. 2010. "“Does It Take a Village?” Assessing Neighborhood Influences on Children’s Self-Control." *Journal of Research in Crime and Delinquency* 47(1):31-62.

Glazier, Anne M., Joseph H. Nadeau, and Timothy J. Aitman. 2002. "Finding Genes That Underlie Complex Traits." *Science* 298(5602):2345-2349.

Gottfredson, Denise C. 2001. *Schools and Delinquency*: Cambridge University Press.

Gottfredson, Michael R and Travis Hirschi. 1990. *A General Theory of Crime*: Stanford University Press.

Guo, Guang, Michael E. Roettger, and Tianji Cai. 2008. "The Integration of Genetic Propensities into Social-Control Models of Delinquency and Violence among Male Youths." *American Sociological Review* 73(4):543-568.

Guo, Guang, Michael E. Roettger, and Jean C. Shih. 2007. "Contributions of the Dat1 and Drd2 Genes to Serious and Violent Delinquency among Adolescents and Young Adults." *Human Genetics* 121(1):125-136.

Hagan, John and Holly Foster. 2003. "S/He's a Rebel: Toward a Sequential Stress Theory of Delinquency and Gendered Pathways to Disadvantage in Emerging Adulthood." *Social Forces* 82(1):53-86.

Hannon, Lance. 2003. "Poverty, Delinquency, and Educational Attainment: Cumulative Disadvantage or Disadvantage Saturation?" *Sociological Inquiry* 73(4):575-594.

Hirschhorn, Joel N. and Mark J. Daly. 2005. "Genome-Wide Association Studies for Common Diseases and Complex Traits." *Nature Reviews Genetics* 6(2):95-108.

Hirschi, Travis. 1969. *Causes of Delinquency*. Los Angeles, CA: University of California.

Jaffee, S. R. and T. S. Price. 2007. "Gene-Environment Correlations: A Review of the Evidence and Implications for Prevention of Mental Illness." *Molecular Psychiatry* 12(5):432-442.

Kim-Cohen, J., A. Caspi, A. Taylor, B. Williams, R. Newcombe, *et al.* 2006. "MAOA, Maltreatment, and Gene-Environment Interaction Predicting Children's Mental Health: New Evidence and a Meta-Analysis." *Molecular Psychiatry* 11(10):903-913.

Lander, E. S. and D. Botstein. 1986. "Mapping Complex Genetic Traits in Humans: New Methods Using a Complete Rflp Linkage Map." *Cold Spring Harbor Symposia on Quantitative Biology* 51(1):49-62.

Lander, E. S. and N. J. Schork. 1994. "Genetic Dissection of Complex Traits." *Science* 265(5181):2037-2048.

Lee, S. Hong, Teresa R. DeCandia, Stephan Ripke, Jian Yang, Patrick F. Sullivan, *et al.* 2012. "Estimating the Proportion of Variation in Susceptibility to Schizophrenia Captured by Common Snps." *Nature Genetics* 44(3):247-U35.

Lee, Sang Hong, Naomi R Wray, Michael E Goddard, and Peter M Visscher. 2011. "Estimating Missing Heritability for Disease from Genome-Wide Association Studies." *The American Journal of Human Genetics* 88(3):294-305.

Lutgendorf, Susan K., Anil K. Sood, Steve W. Cole, Koen DeGeest, Caroline Y. Sung, *et al.* 2009. "Depression, Social Support, and Beta-Adrenergic Transcription Control in Human Ovarian Cancer." *Brain Behavior and Immunity* 23(2):176-183.

Meldrum, Ryan Charles. 2008. "Beyond Parenting: An Examination of the Etiology of Self-Control." *Journal of Criminal Justice* 36(3):244-251.

Morgan, Drake, Michael A. Nader, Kathleen A. Grant, H. Donald Gage, Robert H. Mach, *et al.* 2002. "Social Dominance in Monkeys: Dopamine D2 Receptors and Cocaine Self-Administration." *Nature neuroscience* 5(2):169-174.

Newman, Timothy K., Yana V. Syagailo, Christina S. Barr, Jens R. Wendland, Maribeth Champoux, *et al.* 2005. "Monoamine Oxidase a Gene Promoter Variation and Rearing Experience Influences Aggressive Behavior in Rhesus Monkeys." *Biological Psychiatry* 57(2):167-172.

Norman, Greg J., Louise C. Hawkley, Steve W. Cole, Gary G. Berntson, and John T. Cacioppo. 2012. "Social Neuroscience: The Social Brain, Oxytocin, and Health." *Social Neuroscience* 7(1):18-29.

Pescosolido, Bernice A., Brea L. Perry, J. Scott Long, Jack K. Martin, Jr John I. Nurnberger, *et al.* 2008. "Under the Influence of Genetics: How Transdisciplinarity Leads Us to Rethink Social Pathways to Illness." *American Journal of Sociology* 114(S1):S171-S201.

Plomin, R., J. C. DeFries, G. E. McClearn, and P. McGuffin. 2001. *Behavioral Genetics*, Edited by R. Plomin. New York: Worth Pubs.

Pratt, Travis C, Michael G Turner, and Alex R Piquero. 2004. "Parental Socialization and Community Context: A Longitudinal Analysis of the Structural Sources of Low Self-Control." *Journal of Research in Crime and Delinquency* 41(3):219-243.

Price, Alkes L., Noah A. Zaitlen, David Reich, and Nick Patterson. 2010. "New Approaches to Population Stratification in Genome-Wide Association Studies." *Nature Reviews Genetics* 11(7):459-463.

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

Rietveld, Cornelius A., Arpana Agrawal, Johan G. Eriksson, Eva Albrecht, Behrooz Z. Alizadeh, *et al.* 2013. "Gwas of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment." *Science* 340(6139):1467-1471.

Rutter, Michael, Terrie E. Moffitt, and Avshalom Caspi. 2006. "Gene-Environment Interplay and Psychopathology: Multiple Varieties but Real Effects." *Journal of Child Psychology and Psychiatry, and Allied Disciplines* 47(3-4):226-261.

Sampson, Robert J and John H Laub. 1993. *Crime in the Making: Pathways and Turning Points through Life*: Harvard University Press.

Shanahan, Michael J. and Jason D. Boardman. 2009. "Genetics and Behavior in the Life Course: A Promising Frontier." Pp. 215–35 in *The Craft of Life Course Research*, edited by G. H. E. a. J. Z. Giele. New York: Guilford.

Shanahan, Michael J. and Scott M. Hofer. 2005. "Social Context in Gene-Environment Interactions: Retrospect and Prospect." *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences* 60 Spec No 1(Special Issue 1):65-76.

Shanahan, Michael J., Stephen Vaisey, Lance D. Erickson, and Andrew Smolen. 2008. "Environmental Contingencies and Genetic Propensities: Social Capital, Educational Continuation, and Dopamine Receptor Gene Drd2." *American Journal of Sociology* 114 Suppl(1):S260-S286.

Simons, Ronald L., Man Kit Lei, Steven R. H. Beach, Gene H. Brody, Robert A. Philibert, *et al.* 2011. "Social Environment, Genes, and Aggression: Evidence Supporting the Differential Susceptibility Perspective." *American Sociological Review* 76(6):883-912.

Tung, Jenny, Luis B. Barreiro, Zachary P. Johnson, Kasper D. Hansen, Vasiliki Michopoulos, *et al.* 2012. "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* 109(17):6490-6495.

Turner, Michael G, Alex R Piquero, and Travis C Pratt. 2005. "The School Context as a Source of Self-Control." *Journal of Criminal Justice* 33(4):327-339.

Vanyukov, Michael M., Brion S. Maher, Bernie Devlin, Galina P. Kirillova, Levent Kirisci, *et al.* 2007. "The Maoa Promoter Polymorphism, Disruptive Behavior Disorders, and Early Onset Substance Use Disorder: Gene-Environment Interaction." *Psychiatric genetics* 17(6):323-332.

Vinkhuyzen, A. A. E., M. Luciano, A. Payton, M. Horan, W. Ollier, *et al.* 2012. "Common Snps Explain Some of the Variation in the Personality Dimensions of Neuroticism and Extraversion." *Translational psychiatry* 2(4):e102.

Visscher, Peter M., Matthew A. Brown, Mark I. McCarthy, and Jian Yang. 2012. "Five Years of Gwas Discovery." *American Journal of Human Genetics* 90(1):7-24.

Wagner, Brandon, Jiang Li, Hexuan Liu, and Guang Guo. 2013. "Gene-Environment Correlation: Difficulties and a Natural Experiment-Based Strategy." *American Journal of Public Health* 103 Suppl 1(Journal Article):S167-73.

Wikström, Per-Olof H and Robert J Sampson. 2003. "Social Mechanisms of Community Influences on Crime and Pathways in Criminality." Pp. 118-148 in *The Causes of*

Conduct Disorder and Serious Juvenile Delinquency, edited by B. Lahey, T. Moffitt, and A. Caspi. New York, NY: Guilford.

Yang, J., B Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, *et al.* 2010. "Common Snps Explain a Large Proportion of the Heritability for Human Height." *Nature Genetics* 42(7):565-569.

Yang, Jian, M. Geoffrey Hayes, William G. Hill, Maria Teresa Landi, Alvaro Alonso, *et al.* 2011a. "Genome Partitioning of Genetic Variation for Complex Traits Using Common Snps." *Nature Genetics* 43(6):519-525.

Yang, Jian, S. Hong Lee, Michael E. Goddard, and Peter M. Visscher. 2011b. "Gcta: A Tool for Genome-Wide Complex Trait Analysis." *The American Journal of Human Genetics* 88(1):76-82.

CHAPTER 3 GENOMIC STRUCTURE OF SOCIOECONOMIC ACHIEVEMENT OVER THE LIFE COURSE

3.1 Introduction

There has long been an interest in decomposing genetic and environmental contributions to traits related to socioeconomic achievement (Eckland 1967; Eckland 1979; Guo and Stearns 2002; Jencks 1980; Nielsen 2006; Nielsen and Roos 2015; Scarr and Weinberg 1978; Taubman 1976). Yet previous studies have mostly concentrated on one socioeconomic measure at a time (e.g., educational attainment). How the roles of genes and the environment change over time is still a mystery. This chapter aims to examine the life-course dynamics of genetic and environmental contributions to socioeconomic achievement combining recently available genome-wide data and longitudinal socioeconomic measures.

While sociological research has traditionally focused on socioenvironmental influences that shape an individual's socioeconomic achievement, there is increasing evidence that one's socioeconomic achievement can also be affected by individual characteristics, such as intelligence (Nielsen 2006; Nielsen 2008) and personality (Shanahan *et al.* 2014). All these individual characteristics are heritable to some degree (Benyamin *et al.* 2014; Deary *et al.* 2009; Johnson *et al.* 2008; Okbay *et al.* 2016a). There is also evidence that genetic factors largely contribute to stability in personal characteristics, and socioenvironmental factors are responsible for changes in these characteristics (Deary *et al.* 2012; Lyons *et al.* 2009).

However, very few studies have examined the life-course dynamics of genetic and socioenvironmental influences on socioeconomic achievement. Socioeconomic achievement is multidimensional, and the roles of different dimensions vary over the life course. Specifically, one's socioeconomic status (SES) in childhood primarily depends on his/her parents' education and occupation. Parental SES is transferred to one's own occupational status through education. In late adulthood, the role of wealth becomes more and more important. It is unclear, however, how genetic and environmental factors contribute to socioeconomic achievement at different life stages. Specifically, what are the magnitudes of genetic and environmental influences on different socioeconomic achievement measures such as educational attainment, occupational status, and wealth? Do the same or distinct genetic influences on socioeconomic achievement operate at different life stages? What are the genetic and environmental contributions to stability and mobility in socioeconomic achievement over time?

To address these issues, this study integrates genome-wide genotype data with longitudinal measures of socioeconomic achievement in the Health and Retirement Study (HRS). First, I estimate and compare social and genetic contributions to three socioeconomic achievement measures, namely educational attainment, occupational status, and wealth in late adulthood. Second, I investigate the extent to which the three measures are influenced by the same genetic variants. Finally, I examine genetic and environmental contributions to changes in socioeconomic achievement over the life course.

3.2 Social and Genetic Contributions to Socioeconomic Achievement

Conventional social stratification and social mobility theories typically view one's socioeconomic experience as a linear process over the life course. The process begins with family background typically measured as parental education and occupation, through education,

and ends with occupational status indicated as occupational prestige and income. This process, as numerous studies have demonstrated, is shaped by the social context in which one lives (Blau and Duncan 1967; Boudon 1974; Bourdieu and Passeron 1977; Di Maggio 1982; Hout 1988; Mare 1980; Sewell *et al.* 1969).

Additionally, as noticed by Blau and Duncan (1967), unmeasured individual characteristics also play a pivotal role in influencing socioeconomic achievement. In a meta-analysis of twin/family studies, Branigan *et al.* (2013) show that the mean heritability of educational attainment is around 40%, and the shared environmentality accounts for about 36% variance in educational attainment. Using genome-wide DNA data, Rietveld *et al.* (2013) identify three genetic variants as significant predictors of educational attainment (p-value < 5×10^{-8}). In a more recent genome-wide association study (GWAS), Okbay *et al.* (2016b) identify 74 genome-wide significant variants associated with the number of years of schooling completed.

The composition of genetic and socioenvironmental influences on socioeconomic achievement, however, has been found to vary across social context. A classic example is the study of Heath *et al.* (1985). Using samples of three cohorts of Norwegian twins, the authors find that the heritability of educational attainment is substantially greater, while the contribution of shared environmentality is smaller for younger cohorts than older ones. Also, they find that this change merely existed among males. They attribute this change to liberal reforms that produced more equal opportunities for younger cohorts, yet the equality of opportunity for females did not improve to the same extent as for males. Branigan *et al.* (2013) summarize gene-environment interaction findings on educational attainment. They show that heritability of educational attainment varies by gender, nation, and birth cohort.

The present study extends this line of research to examine social and genomic structure of socioeconomic achievement over the life course. It addresses three critical questions: (1) what is the impact of genetic and environmental influences on socioeconomic achievement at different life stages? (2) Are the same or different genetic influences in operation at various life stages? (3) What are the genetic and environmental contributions to stability and mobility in socioeconomic achievement over time?

3.3 Research Questions

3.3.1 What is the impact of genetic and environmental influences on socioeconomic achievement at different life stages?

There are at least two different perspectives that contribute to our understanding of life-course variations in weights of genetic and environmental components. A natural perspective is that as people age, they are increasingly subject to environmental influences, "the slings and arrows of outrageous fortune" in life, and therefore genetic influences decrease. The study of Deary *et al.* (2012), for example, show that additive genetic effects explain 48 percent of the variation in intelligence at age 11, but only 28 percent at ages 65-79. A contrasting "niche picking" perspective predicts that genetic influences grow with age. Accordingly, when people are young, they are constrained in a rearing environment provided by their parents. As they grow up, they shift away from parental influences, and are increasingly able to select their own experiences (Scarr and McCartney 1983). Evidence from twin and family studies shows that the genetic contribution to cognitive abilities rises with age, whereas the contribution of shared environmental factors reduces with age, at least until middle adulthood (McCartney *et al.* 1990; McGue *et al.* 1993; Plomin and Spinath 2004).

It is unknown that how overall genetic and environment contributions to socioeconomic achievement vary over the life span. Also, most studies on life-course patterns in genetic and environmental components rely on cross-sectional data, in which age effects may be confounded by cohort effects. *Aim 1 of this study is to assess overall genetic and environmental contributions to socioeconomic achievement at different life stages using longitudinal data*. From the "slings and arrows" perspective, the genetic contribution to wealth in later life is expected to be smaller than to occupational status than to educational attainment, while the order is reversed for environmental contribution. From the "niche picking" perspective, the genetic contribution to wealth in later life is expected to be greater than to occupational status than to educational attainment.

3.3.2 Are the same or different genetic influences in operation influencing socioeconomic achievement at various life stages?

Even if heritability is constant over the life course, specific genetic variants at work are not necessarily the same at different ages (Plomin *et al.* 1993). This is because that heritability only estimates how much of the phenotypic variation can be attributed to genetic differences among individuals, not which genes or genetic variants matter. Since an individual's genotype is fixed at conception, it is tempting to view genetic influences as static. However, as Francis Galton noted more than one century ago, "it must be borne in mind that the divergence of development, when it occurs, need not be ascribed to the effect of different nurtures, but it is quite possible that it may be due to the appearance of qualities inherited at birth, though dormant,…,in early life" (Galton 1876:402). Although an individual's genotype does not change over the life course, different genes may operate at different developmental periods (Vogler 2006). In other words, it is likely that new genetic variants are "tuned on" in a later course of a

trait (genetic innovation), while the ones operate at earlier stages become less influential over time (genetic attenuation). Previous studies have shown new genetic influences emerging during childhood and adolescence (Cardon *et al.* 1992; DeFries *et al.* 1987; Eaves 1982; Lyons *et al.* 2009; Petrill *et al.* 2004; Wilson and Matheny 1983). There is also evidence that the same genetic influences operate from early adulthood to late middle age (Lyons *et al.* 2009)

There has been limited investigation of the extent to which the same or different genetic influences are in operation affecting socioeconomic achievement at various life stages. *Aim 2 of this study is to examine this issue using life-course socioeconomic measures together with genome-wide genotype data.*

3.3.3 What are the genetic and environmental contributions to socioeocnomic stability and mobility?

Previous research has assessed genetic and environmental contributions to stability and mobility in individual traits related to socioeconomic achievement. Based on a longitudinal study of individuals in their second half of life, Plomin *et al.* (1994) show that the genetic factors contribute to 90% of cognitive stability. Lyons *et al.* (2009) examine genetic and environmental contributions to stability and change in cognitive ability during 35 years of adulthood. They observe a genetic correlation of 1.0 between cognitive ability measured in young adulthood and that measured in late middle age. Using an accelerated longitudinal design of twin samples, McGue and Christensen (2002) find a heritability of 6% for the linear change in cognitive abilities. Reynolds *et al.* (2005) examine linear and quadratic change at age 65. For linear change, they observe a heritability of 1% and non-shared environmentality 99%; for quadratic change, they find a heritability of 43% and non-shared environmentality 57%. Their findings

suggest that genetic factors were primarily responsible for stability, and non-shared environmental factors were primarily responsible for changes in cognitive ability.

It is unclear that what roles genetic and environmental factors play in determining stability and mobility in socioeconomic achievement over the life course. *Aim 3 of this study is to disentangle genetic and environmental contributions to stability and mobility in* socioeconomic *achievement over the life span.*

3.4 Data

Data for this study come from the Health and Retirement Study (HRS). HRS is a longitudinal study of Americans over age 50 conducted every two years from 1992 to 2012; it collects information on economic, health, social, and other factors relevant to aging and retirement. HRS includes six birth cohorts with different entry years: the Study of Assets and Health Dynamics Among the Oldest Old (AHEAD) cohort (born before 1924) surveyed in 1993, 1995, and 1998-2012; Children of Depression (CODA) cohort (born 1924-1930) surveyed in 1998-2012; HRS cohort (born 1931-1941) surveyed from 1992-2012; War Baby (WB) cohort (born 1942-1947) surveyed in 1998-2012; Early Baby Boomers (EBB) cohort (born 1948-1953) surveyed in 2004-2012; and Mid Baby Boomers (MBB) cohort (born 1954-60) surveyed in 2010 and 2012. When the present study is conducted, genetic data are available for AHEAD, CODA, HRS, WB, and EBB.

Genetic Samples. DNA samples were collected in 2006 and 2008. Of the collected samples, 13,129 were put into genotyping production using the Illumina Human Omni-2.5 Quad beadchip, with coverage of approximately 2.5 million single nucleotide polymorphisms (SNPs), and 12,507 passed the University of Washington Genetics Coordinating Center's (GCC)

standardized quality control processes. To minimize confounding effects of population stratification, this study focuses on non-Hispanic whites.

Socioeconomic Achievement Measures. Three life-course socioeconomic achievement measures are used in this study: educational attainment, occupational status, and late-adulthood wealth. Educational attainment is measured using years of education ("What is the highest grade of school or year of college you completed?"). Occupational status is based on the occupational prestige score of respondents' current job or last job for retirees. Late-adulthood wealth is based on household wealth (sum of all types of assets, pensions, etc.). Details of the variables are provided in Table 3.1.

These socioeconomic achievement measures may have different meanings for different birth cohorts. For example, a high school degree might indicate high socioeconomic achievement for earlier cohorts, but medium/low socioeconomic achievement for later cohorts. This issue was addressed by within-cohort standardization. Specifically, the three life-course socioeconomic achievement measures were recoded into relative indicators based on a baseline sample. The baseline sample includes onset measures for all respondents (either provided DNA or did not). These measures were taken in 1992 for HRS, 1993 for AHEAD, 1998 for CODA and WB, and 2004 for EBB. Respondents were divided into 10 categories on the basis of the 9 deciles of each of the three socioeconomic achievement measures within each birth cohort in the baseline sample. In addition to cohort effects, education and occupation measures might also be subject to gender differences. These two variables therefore were also standardized by gender within each cohort.

3.5 Methods

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

The genomic-relatedness-matrix restricted maximum likelihood (GREML) method is developed by Yang et al.(2011). In contrast to single-variant association analysis where each SNP is tested against an adjusted p-value, GREML treats all SNP effects as random effects. The basic GREML model can be described by the following equation:

$$Y=X\beta + W\mu + \varepsilon, \qquad \text{(Equation 3.1)}$$

where Y is the outcome variable; β is a vector of the coefficients of fixed covariates such as age, sex and other controls; μ is a vector of genetic effects with $\mu_i \sim N(0, \sigma^2_\mu)$, where i = 1,…, I, with I being the number of SNPs; ε is a vector of residual effects with $\varepsilon_j \sim N(0, \sigma^2_\varepsilon)$, where j = 1,…, J, with J being the number of individuals; W is a standardized genotype matrix. Yang et al. (2010) innovatively applied a previous result that has been known in animal genetics (Goddard et al. 2009). The result defines $g = W\mu$, $\mathbf{A} = WW^T/I$ and $\sigma^2_g = I\sigma^2_\mu$. Then Equation 2 is mathematically equivalent to Equation 1:

$$Y=X\beta + g + \varepsilon, \text{ with } Var(Y) = \mathbf{A}\sigma^2_g + I_\varepsilon\sigma^2_\varepsilon, \qquad \text{(Equation 3.2)}$$

where g is a J*1 vector of the total genetic effects of the individuals with $g \sim N(0, \mathbf{A}\sigma^2_g)$, **A** is the genomic relatedness matrix (GRM) and σ^2_g is the total genetic variance explained by the SNPs. Hence σ^2_g can be estimated by the restricted maximum likelihood (REML) approach, depending on the GRM estimated from all SNPs. The genetic contribution to the outcome can be assessed using the proportion of total variance in the outcome explained by all SNPs, which can be expressed as $\sigma^2_g/(\sigma^2_g + \sigma^2_\varepsilon)$, and the environmental contribution can be expressed as $\sigma^2_\varepsilon/(\sigma^2_g + \sigma^2_\varepsilon)$.

The GREML approach has been extended to estimate the genetic correlation (rG) between different traits (Lee *et al.* 2012). The bivariate GREML model can be described by the following equation:

$$Y_t = X_t b_t + g_t + e_t,\ g_t \sim N(0, A\sigma^2_{gt}),\ e_t \sim N(0, \sigma^2_{et}), \qquad \text{(Equation 3.3)}$$

where Y_t is a vector of observations for trait t (t = 1 or 2). For example, Y_1 may represent respondents' educational attainment and Y_2 represents their occupational status or wealth. Equation (4) demonstrates the covariance matrix between two traits:

$$\text{Cov}(Y_1, Y_2) = \begin{pmatrix} A\sigma^2_{g1} + I\sigma^2_{e1} & A\sigma_{g1g2} + I\sigma_{e1e2} \\ A\sigma_{g1g2} + I\sigma_{e2e1} & A\sigma^2_{g2} + I\sigma^2_{e2} \end{pmatrix}. \qquad \text{(Equation 3.4)}$$

The rG is defined as $\frac{\sigma_{g1g2}}{\sigma_{g1}\sigma_{g2}}$. A high rG between two traits indicates that SNP effects on two traits are relatively similar. Substantively speaking, an rG of 1 implies that the two traits are affected by the same genetic variants, so that any variation between the two traits is environmental. At the other extreme, an rG of 0 suggests genetically independent traits; in other words, the two traits are affected by completely different genetic variants.

The bivariate GREML model has been employed to assess genetic correlations between phenotypes such as intelligence, education, and health outcomes. For example, Deary *et al.* (2012) estimate a highly significant rG of .62 between intelligence in adolescence (age 11) and that in late adulthood (age 65-78). In a more recent study, Boardman *et al.* (2015) examine the genetic correlations between education and health measures such as BMI, depression, and self-rated health. They find that the phenotypic correlation between depression and education (rG = -.75) and that between self-rated health and education (rG = -.91) were largely explained by common genetic factors, while there was no evidence that the correlation between BMI and education is influenced by common genetic factors (rG = -.03).

Polygenic Score

Polygenic scores (PS) are typically constructed based on existing GWAS results. They can be calculated using the following formula:

$$PS_i = \sum_{j=1}^{J} \beta_j * x_{ij}, \qquad \text{(Equation 3.5)}$$

where PS_i is the PS of individual i, β_j is the coefficient for variant j estimated using GWAS data, x_{ij} is the number of risk alleles on variant j that individual i possesses.

The first PS analysis is 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).

This approach has been employed to advance sociological research. Liu and Guo (2015), for example, provide evidence that cumulative socioeconomic advantage significantly decreased an individual's BMI in middle to late adulthood only for those with a higher genetic propensity for obesity [measured by a polygenic score constructed using the 32 obesity-related SNPs in Speliotes *et al.* (2010)], but not for those with lower genetic propensities. In another recent study, Conley *et al.* (2015) decompose the intergenerational association in educational attainment into genetic and environmental components by including a polygenic score [constructed based on results in the study of Rietveld *et al.* (2013)] as a predictor in the traditional intergenerational model. They find that genetic factors account for approximately a sixth and social inheritance accounts for five-sixths of the intergenerational association in educational attainment.

3.6 Analytical Strategy

Three analyses were conducted in this chapter. I first performed univariate GREML analyses to estimate the proportion of phenotypic variance explained by genome-wide SNPs for each of the three socioeconomic achievement outcomes. Also, to assess genetic and environmental contributions to socioeconomic mobility, I estimated the proportion of variance in

occupational status explained by genome-wide SNPs after adjusting for educational attainment, and the proportion of variance in wealth explained by genome-wide SNPs after adjusting for educational attainment and occupational status.

Secondly, I conducted bivariate GREML analyses to test genetic correlations for three pairs of the socioeconomic achievement measures (i.e., education and occupation, education and wealth, and occupation and wealth). For each genetic correlation, likelihood ratio test (LRT) was performed to compare (1) the fitted model and a null model assuming no genetic correlation (i.e., rG = 0, substantively meaning that two traits are affected by completely different SNPs); (2) and the fitted model and a null model assuming perfect genetic correlation (i.e., rG = 1, substantively meaning that two traits are affected by the same SNPs).

Finally, I conducted a polygenic score constructed based on SNPs found to be associated with intelligence. Before the polygenic score calculation, I matched SNPs in HRS with the most recent GWAS results reported by Benyamin *et al.* (2014) and used the 501, 484 matched SNPs to 'score' each of respondent's genetic predisposition to intelligence. I then calculated the polygenic scores according to the methods described by Dudbridge (2013) using the PRsice software (Euesden *et al.* 2015). Polygenic scores range from -34.30-53.48 and are normally distributed in HRS (M=12.80, SD=12.23). I standardized the scores to have mean=0, sd=1 for analysis. Greater scores are associated with higher levels of intelligence. I estimated and compared the associations between polygenic scores and the three socioeconomic achievement outcomes.

To account for potential population stratification, I adjusted all the analyses for the first ten principal components computed from the genome-wide SNP data using the EIGENSOFT software (Price *et al.* 2010; Price *et al.* 2006). In addition, all three analyses were also conducted

for males and females separately to detect gender differences in genetic and environmental contributions to life-course socioeconomic achievement.

3.7 Results

3.7.1 Bivariate correlations among key variables

Table 3.2 shows the Pearson's correlations between the standardized life-course socioeconomic achievement measures. All the correlations are statistically significant (P<.001). Pearson's correlation is about .50 between educational attainment and occupational status, .30 between educational attainment and wealth, and .22 between occupational status and wealth. While the correlation between occupational status and wealth is significantly greater for male than for females, there are no significant gender differences in the correlations between educational attainment and occupational status and between educational attainment and wealth.

3.7.2 What is the impact of genetic and environmental influences on socioeconomic achievement at different life stages?

Table 3.3 displays univariate GREML results for estimating genetic and environmental contributions to SES at different life stages. As can be seen, the proportion of variance explained by genome-wide SNPs is 42% for educational attainment. This result is similar to heritability estimates from many twin and sibling studies (Branigan *et al.* 2013). As indicated by the significant p value ($p < .001$) in the bottom, dropping the genetic component causes a significant loss of information in the model. The SNP heritability is 35% for occupational status ($p < .001$), and 33% for wealth ($p < .001$). Consistent with the "slings and arrows" hypothesis, these results suggest that the genetic contribution to socioeconomic achievement decreases as people age. Yet differences in the genetic contribution estimates are not large among three socioeconomic achievement measures.

The differences in the estimated genetic contribution to three socioeconomic measures may be due to different number of observations used in estimating three models (e.g., there are more missing values in occupational status than in educational attainment and wealth). To test this possibility, I conducted a robustness test in which all models were estimated using observations without missing values in all three outcomes. The results remain, suggesting that differences in estimated genetic contribution to three socioeconomic measures are not driven by missing values.

Moreover, as previous research has shown gender differences in genetic influences on socioeconomic achievement (Branigan *et al.* 2013; Heath *et al.* 1985), I also estimated the univariate GREML models for males and females separately. As results in Table 3.3 show, the overall pattern (i.e., genetic contribution declines over the life course) holds for males but not for females. Among females, the genetic contribution to occupational status (17%) is smaller than the genetic contribution to wealth (33%). Again, I tested whether this is because of varying missing patterns in the three socioeconomic achievement measures by fitting the models to the same number of observations for females. As a result, the estimated genetic contribution to occupational status is still around 17%, but the estimated genetic contribution to wealth drops to 15%. This suggests that deviance from the linear pattern for females is due to variation in sample size used to estimate the GREML models.

3.7.3 Are the same or different genetic influences in operation influencing socioeconomic achievement at various life stages?

Table 3.4 demonstrates results of bivariate genomic association among educational attainment, occupational status, and wealth in late adulthood. Each column of Table 3.4 includes bivariate GREML results between two socioeconomic achievement measures. For example,

column 7 demonstrates results of estimating genetic and environmental contributions to the correlation between educational attainment and occupational status for both genders. It shows 42% and 35% variances in education and occupation are explained by genome-wide genotypes. These estimates are very similar to estimates in Table 3.3 produced by univariate GREML models. The genetic correlation is around .70 between educational attainment and occupational status ($p < .001$ for testing H_0: rG = 0 and $p < .001$ for testing H_0: rG = 1), .83 between educational attainment and wealth ($p < .001$ for testing H_0: rG = 0 and $p = .071$ for testing H_0: rG = 1), and .43 ($p = .003$ for testing H_0: rG = 0 and $p < .001$ for testing H_0: rG = 1) between occupational status and wealth. These results show that the three life-course socioeconomic achievement outcomes are significantly genetically correlated with each other, yet the genetic variants in operation are not exactly the same at different life stages.

The bivariate GREML models are also estimated by gender. For males, the genetic correlation is .50 between educational attainment and occupational status ($p = .09$ for testing H_0: rG = 0 and $p < .02$ for testing H_0: rG = 1), 1 between educational attainment and wealth ($p < .01$ for testing H_0: rG = 0 and $p = .5$ for testing H_0: rG = 1), and .05 ($p = .46$ for testing H_0: rG = 0 and $p = .01$ for testing H_0: rG = 1) between occupational status and wealth. For females, the genetic correlation is .81 between educational attainment and occupational status ($p = .01$ for testing H_0: rG = 0 and $p = .29$ for testing H_0: rG = 1), .90 between educational attainment and wealth ($p < .001$ for testing H_0: rG = 0 and $p = .31$ for testing H_0: rG = 1), and .52 ($p = .08$ for testing H_0: rG = 0 and $p = .19$ for testing H_0: rG = 1) between occupational status and wealth. There are more variations in the genetic correlation among three socioeconomic measures in males than in female. This is probably due to smaller male sample sizes.

3.7.4 What are the genetic and environmental contributions to socioeocnomic stability and mobility?

Based on the bivariate GREML results in Table 3.4, over 50% of the covariance between each two of the three socioeconomic achievement measures can be explained by genome-wide SNPs. This suggests that stability in socioeconomic achievement is largely attributable to genetic factors. To estimate genetic and environmental contributions to mobility, I estimated a univariate GREML model predicting respondents' occupation after adjusting for their education, and a model predicting respondent's wealth after adjusting for their education and occupation. The results show that for individuals with the same levels of educational attainment, 74% of variation in their occupational status can be explained by non-genetic environmental factors; for those with the same levels of educational attainment and occupational status, 77% of variation in their wealth can be explained by non-genetic environmental factors. Social mobility, therefore, is mainly a consequence of the environmental changes.

The mobility models are also estimated by gender. For males with the same levels of educational attainment, 69% of variation in their occupational status can be explained by environmental factors; for males with the same levels of educational attainment and occupational status, 83% of variation in wealth can be explained by environmental factors. For females with the same levels of educational attainment, 91% of variation in their occupational status can be explained by environmental factors; for females with the same levels of educational attainment and occupational status, 78% of variation in wealth can be explained by environmental factors.

3.8 An Example of Genetic Stability and Mobility

Above results have shown that while commons genetic variants are associated with three socioeconomic achievement outcomes, the genetic influences tend to decrease over the life

course. Some genes influencing education are likely to be "turned off" at some stages of life and therefore they have no impact on occupation and wealth. Here I demonstrate genetic stability and mobility using polygenic scores constructed based on intelligence-related SNPs. First, the intelligence polygenic score is significantly associated with all three socioeconomic achievement outcomes. This suggests that intelligence-related genetic variants continuously influence one's socioeconomic achievement over the life course. However, it is noticeable that magnitudes of the genetic associations with occupational status and wealth are about half of that of the genetic association with educational attainment (see Table 3.6). This provides evidence for genetic attenuation, namely that the importance of intelligence-related genetic variants for socioeconomic achievement declines when one ages.

3.8 Discussion and Conclusions

To date, few studies have investigated genetic and environmental contributions to stability and mobility of socioeconomic achievement over the life span. Recently available genome-wide data in together with longitudinal socioeconomic achievement measures in HRS provide a unique opportunity for this investigation. Results in this study suggest that both genetic and environment factors make important contributions to stability and mobility of socioeconomic achievement. Accordingly, additive genetic effects account for about 42% variation in education, 35% in occupation, and 33% in wealth. This provides suggestive evidence for a decreasing genetic influence and an increasing environmental influence on SES over the life span. As an individual ages, his/her socioeconomic achievement becomes more likely to be a consequence of contextual opportunities and constraints, rather than a consequence of his/her intrinsic characteristics determined by genes.

Moreover, results show highly significant rG among educational attainment, occupational status, and wealth. These results provide evidence for stable genetic influences. This means that the three socioeconomic achievement measures are, to a large extent, influenced by the same genetic variants. Yet the rGs among three socioeconomic achievement measures are not perfectly 1. It suggests that although the DNA sequence is stable across the life-course, the effects of some genes may only be apparent at certain life stages. There are more than one possible explanations for genetic innovation and attenuation. Genetic innovation might depict new rGs that occur as the individual is exposed to new experiences, and genetic attenuation might represent the disappearance of rGs as the individual's exposure to certain experiences reduce. Alternatively, or probably in addition to rGs, genetic innovation might represent certain developmental steps in the brain which were not previously important but become pertinent at later life, and genetic attenuation might indicate the steps previously important but become minor at later life (Pickles *et al.* 1998). I demonstrate genetic attenuation using intelligence polygenic score as an example. I find that the association between the score and occupation and wealth is only half of the genetic association with educational attainment. Mechanisms of genetic innovation and attenuation, however, are still unclear. Future research may extend this study to examine the specific mechanisms of genetic innovation and attenuation over the life course when data become available.

The findings in this chapter is summarized in Figure 3.1. As it shows, life-course socioeconomic achievement is subject to both common and uncommon genetic influences. Some genes may continuously play an important role in influencing socioeconomic achievement over time, while others may only operate at certain life stages. This finding provides important insights for sociological studies concerning complex relationships among educational attainment,

occupational status, income, and wealth. These relationships are often a mixture of socioenvironmental and genetic effects.

Social scientists have noticed that socioenvironmental effects are likely to be confounded by genetic factors. A current approach to address such confounding is to include genetic influences as covariates in regression models (Conley et al. 2015). While these models can rule out genetic confounding factors that are time-constant, they do not take into account time-varying heterogeneity. In particular, although DNA typically does not change over time, genetic influences could be time-varying due to genetic innovation and attenuation. As a consequence, the models can be biased if all genetic influences are treated as time-invariant.

It is possible that the use of different measures at different life stages might result in variability in genetic contribution estimates. Three different indicators are used in assessing socioeconomic achievement at different life stages. A more refined lifetime SES indictor might be a composite measure with varying weights on different dimensions of SES (e.g., educational attainment has more weight in measuring young adulthood SES and household wealth has more weight in measuring late adulthood socioeconomic achievement). In HRS, however, income and wealth information is only available after age 50, and educational attainment is the only indicator of young adulthood SES. Additionally, sample size is too small for GREML to identify gender differences in genetic and environmental contributions to socioeconomic achievement. Although genetic information is available for some minorities (e.g., blacks and Asians), their sample sizes are insufficient to achieve adequate statistical power for separate analyses. Future research can use more refined measures and extend the analysis in this study to other population groups when data become available.

Despite these limitations, this study demonstrates the complicated roles played by genes and social environment in influencing socioeconomic achievement over the life course. The theoretical framework and methods in this article could be expanded to study other complex traits of interest to social scientists.

REFERENCES

Benyamin, B., B. St Pourcain, O. S. Davis, G. Davies, N. K. Hansell, *et al.* 2014. "Childhood Intelligence Is Heritable, Highly Polygenic and Associated with Fnbp1l." *Molecular Psychiatry* 19(2):253-258.

Blau, P. M. and O. D. Duncan. 1967. *The American Occupational Structure*. New York: Wiley.

Boardman, Jason D., Benjamin W. Domingue, and Jonathan Daw. 2015. "What Can Genes Tell Us About the Relationship between Education and Health?" *Social Science and Medicine* 127:171-80.

Boudon, Raymond. 1974. *Education, Opportunity, and Social Inequality.* New York: John Wiley.

Bourdieu, Pierre and Jean-Claude Passeron. 1977. "Reproduction in Education, Society, and Culture." Beverly Hills, CA: Sage.

Branigan, Amelia R., Kenneth J. McCallum, and Jeremy Freese. 2013. "Variation in the Heritability of Educational Attainment: An International Meta-Analysis." *Social Forces* 92(1):109-140.

Cardon, Lon R., David W. Fulker, J. C. DeFries, and Robert Plomin. 1992. "Continuity and Change in General Cognitive Ability from 1 to 7 Years of Age." *Developmental Psychology* 28(1):64-73.

Conley, Dalton, Benjamin W. Domingue, David Cesarini, Christopher Dawes, Cornelius A. Rietveld , *et al.* 2015. "Is the Effect of Parental Education on Offspring Biased or Moderated by Genotype?" *Sociological Science* 2:82-105.

Deary, Ian J., Janie Corley, Paul Redmond, Helen C. Fox, Suzanne J. Rowe, *et al.* 2012. "Genetic Contributions to Stability and Change in Intelligence from Childhood to Old Age." *Nature* 482(7384):212-215.

Deary, Ian J., W. Johnson, and L. M. Houlihan. 2009. "Genetic Foundations of Human Intelligence." *Human Genetics* 126(1):215-232.

DeFries, J. C., Robert Plomin, and Michele C. LaBuda. 1987. "Genetic Stability of Cognitive Development from Childhood to Adulthood." *Developmental Psychology* 23(1):4-12.

Di Maggio, Paul. 1982. "Cultural Capital and School Success: The Impact of Status Culture Participation on the Grades of Us High School Students." *American Sociological Review* 47(2):189-201.

Dudbridge, Frank. 2013. "Power and Predictive Accuracy of Polygenic Risk Scores." *PLoS genetics* 9(3):e1003348.

Eaves, L.J. 1982. "The Utility of Twins." Pp. 249-276 in *Genetic Basis of Epilepsies*, edited by V. E. Anderson, W. A. Hauser, J. K. Penry, and C. F. Sing. New York: Raven.

Eckland, B. K. 1967. "Genetics and Sociology: A Reconsideration." *American Sociological Review* 32(2):173-194.

Eckland, Bruce K. 1979. "Genetic Variance in the Ses-Iq Correlation." *Sociology of Education* 52(3):191-196.

Euesden, J., C. M. Lewis, and P. F. O'Reilly. 2015. "Prsice: Polygenic Risk Score Software." *Bioinformatics* 31(9):1466-1468.

Galton, Francis. 1876. "The History of Twins as a Criterion of the Relative Powers of Nature and Nurture." *The Journal of the Anthropological Institute of Great Britain and Ireland* 5:391–406.

Guo, Guang and Elizabeth Stearns. 2002. "The Social Influences on the Realization of Genetic Potential for Intellectual Development." *Social Forces* 80(3):881-910.

Heath, A. C., K. Berg, L. J. Eaves, M. H. Solaas, L. A. Corey, *et al.* 1985. "Education Policy and the Heritability of Educational Attainment." *Nature* 314(6013):734-736.

Hout, Michael. 1988. "More Universalism, Less Structural Mobility: The American Occupational Structure in the 1980s." *American Journal of Sociology* 93(May 88):1358-1400.

Jencks, Christopher. 1980. "Heredity, Environment, and Public Policy Reconsidered." *American Sociological Review* 45(5):723-736.

Johnson, A.M., P.A. Vernon, and A.R. Feiler. 2008. "Behavioral Genetic Studies of Personality: An

Introduction and Review of the Results of 50+ Years of Research." Pp. 145-173 in *Handbook of Personality Theory and Assessment*, edited by G. Boyle, G. Matthews, and D. Saklofske. London: Sage.

Lee, S. H., J. Yang, M. E. Goddard, P. M. Visscher, and N. R. Wray. 2012. "Estimation of Pleiotropy between Complex Diseases Using Single-Nucleotide Polymorphism-Derived Genomic Relationships and Restricted Maximum Likelihood." *Bioinformatics* 28(19):2540-2542.

Liu, Hexuan and Guang Guo. 2015. "Lifetime Socioeconomic Status, Historical Context, and Genetic Inheritance in Shaping Body Mass in Middle and Late Adulthood." *American Sociological Review* 80(4):705-737.

Locke, Adam E., Bratati Kahali, Sonja I. Berndt, Anne E. Justice, Tune H. Pers, *et al.* 2015. "Genetic Studies of Body Mass Index Yield New Insights for Obesity Biology." *Nature* 518(7538):197-U401.

Lyons, Michael J., Hong Xian, Rosemary Toomey, Seth A. Eisen, William S. Kremen, *et al.* 2009. "Genes Determine Stability and the Environment Determines Change in Cognitive Ability During 35 Years of Adulthood." *Psychological Science* 20(9):1146-1152.

Mare, Robert D. 1980. "Social Background and School Continuation Decisions." *Journal of the American Statistical Association* 75(370):295-305.

McCartney, Kathleen, Monica J. Harris, and Frank Bernieri. 1990. "Growing up and Growing Apart: A Developmental Meta-Analysis of Twin Studies." *Psychological Bulletin* 107(2):226-237.

McGue, Matt and Kaare Christensen. 2002. "The Heritability of Level and Rate-of-Change in Cognitive Functioning in Danish Twins Aged 70 Years and Older." *Experimental Aging Research* 28(4):435-451.

McGue, Matt., Thomas J. Bouchard Jr., William G. Iacono, and David T. Lykken. 1993. "Behavior Genetics of Cognitive Abilities: A Life-Span Perspective." Pp. 59-76 in *Nature, Nurture, and Psychology*, edited by R. Plomin and G. E. McClearn. Washington, DC: American Psychological Association.

Nielsen, François. 2006. "Achievement and Ascription in Educational Attainment: Genetic and Environmental Influences on Adolescent Schooling." *Social Forces* 85(1):193-216.

Nielsen, François and J. Micah Roos. 2015. "Genetics of Educational Attainment and the Persistence of Privilege at the Turn of the 21st Century." *Social Forces* 94(2):535-561.

Nielsen, François. 2008. "The Nature of Social Reproduction: Two Paradigms of Social Mobility." Pp. 1-35 in *Sociologica*.

Okbay, Aysu, Bart M. L. Baselmans, Rico Rueedi, Rauli Svento, Ivana Kolcic, *et al.* 2016a. "Genetic Variants Associated with Subjective Well-Being, Depressive Symptoms, and Neuroticism Identified through Genome-Wide Analyses." *Nature genetics* 48(6):624-624.

Okbay, Aysu, Jonathan P. Beauchamp, Mark Alan Fontana, James J. Lee, Tune H. Pers, *et al.* 2016b. "Genome-Wide Association Study Identifies 74 Loci Associated with Educational Attainment." *Nature* advance online publication.

Petrill, Stephen A., Paul A. Lipton, John K. Hewitt, Robert Plomin, Stacey S. Cherny, *et al.* 2004. "Genetic and Environmental Contributions to General Cognitive Ability through the First 16 Years of Life." *Developmental Psychology* 40(5):805-812.

Pickles, Andrew, Kevin Pickering, Emily Simonoff, Judy Silberg, Joanne Meyer, *et al.* 1998. "Genetic “Clocks” and “Soft” Events: A Twin Model for Pubertal Development and Other Recalled Sequences of Developmental Milestones, Transitions, or Ages at Onset." *Behavior Genetics* 28(4):243-253.

Plomin, R., N. L. Pedersen, P. Lichtenstein, and G. E. McClearn. 1994. "Variability and Stability in Cognitive Abilities Are Largely Genetic Later in Life." *Behavior Genetics* 24(3):207-216.

Plomin, Robert, Jerome Kagan, Robert N. Emde, J. Steven Reznick, Julia M. Braungart, *et al.* 1993. "Genetic Change and Continuity from Fourteen to Twenty Months: The Macarthur Longitudinal Twin Study." *Child Development* 64(5):1354-1376.

Plomin, Robert and Frank M. Spinath. 2004. "Intelligence: Genetics, Genes, and Genomics." *Journal of Personality and Social Psychology* 86(1):112-129.

Price, Alkes L., Noah A. Zaitlen, David Reich, and Nick Patterson. 2010. "New Approaches to Population Stratification in Genome-Wide Association Studies." *Nature Reviews Genetics* 11(7):459-463.

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

Purcell, Shaun M., Naomi R. Wray, Jennifer L. Stone, Peter M. Visscher, Michael C. O'Donovan, *et al.* 2009. "Common Polygenic Variation Contributes to Risk of Schizophrenia and Bipolar Disorder." *Nature* 460(7256):748-752.

Reynolds, Chandra A., Deborah Finkel, John J. McArdle, Margaret Gatz, Stig Berg, *et al.* 2005. "Quantitative Genetic Analysis of Latent Growth Curve Models of Cognitive Abilities in Adulthood." *Developmental Psychology* 41(1):3-16.

Rietveld, Cornelius A., Arpana Agrawal, Johan G. Eriksson, Eva Albrecht, Behrooz Z. Alizadeh, *et al.* 2013. "Gwas of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment." *Science* 340(6139):1467-1471.

Scarr, Sandra and Kathleen McCartney. 1983. "How People Make Their Own Environments: A Theory of Genotype → Environment Effects." *Child Development* 54(2):424-435.

Scarr, Sandra and Richard A. Weinberg. 1978. "The Influence of "Family Background" on Intellectual Attainment." *American Sociological Review* 43(5):674-692.

Sewell, William H., Archibald O. Haller, and Alejandro Portes. 1969. "The Educational and Early Occupational Attainment Process." *American Sociological Review* 34(1):83-92.

Shanahan, Michael J., Shawn Bauldry, Brent W. Roberts, Ross Macmillan, and Rosemary Russo. 2014. "Personality and the Reproduction of Social Class." *Social Forces* 93(1):209-240.

Simonson, Matthew A., Amanda G. Wills, Matthew C. Keller, and Matthew B. McQueen. 2011. "Recent Methods for Polygenic Analysis of Genome-Wide Data Implicate an Important Effect of Common Variants on Cardiovascular Disease Risk." *BMC Medical Genetics* 12(1):146-146.

Speliotes, Elizabeth K., Reedik Mägi, Thomas A. Buchanan, Joshua C. Randall, Sailaja Vedantam, *et al.* 2010. "Association Analyses of 249,796 Individuals Reveal 18 New Loci Associated with Body Mass Index." *Nature Genetics* 42(11):937-48.

Taubman, Paul. 1976. "The Determinants of Earnings: Genetics, Family, and Other Environments: A Study of White Male Twins." *American Economic Review* 66(5):858-870.

Thorleifsson, G., C. N. Palmer, T. Johnson, J. Luan, J. H. Park, *et al.* 2010. "Hundreds of Variants Clustered in Genomic Loci and Biological Pathways Affect Human Height." *Nature* 467(7317):832-838.

Vogler, G.P. 2006. "Behavior Genetics and Aging." in *Handbook of the Psychology of Aging*, edited by J. E. Birren and K. W. Schaie. Amsterdam: Academic Press.

Wilson, Ronald S. and Adam P. Matheny. 1983. "Assessment of Temperament in Infant Twins." *Developmental Psychology* 19(2):172-183.

Wood, A. R., T. Esko, J. Yang, S. Vedantam, T. H. Pers, *et al.* 2014. "Defining the Role of Common Variation in the Genomic and Biological Architecture of Adult Human Height." *Nature Genetics* 46(11):1173-1186.

CHAPTER 4 SOCIAL AND GENOMIC CONTRIBUTIONS TO MULTIGENERATIONAL STABILITY AND MOBILITY OF EDUCATIONAL ATTAINMENT

4.1 Introduction

Intergenerational transmission of education is a central theme in the social stratification and mobility literature. Yet the mechanisms through which educational attainment is transmitted from one generation to another have remained unclear. In this chapter, I take advantage of the recently available molecular genomic data to assess the roles of genes and environment in intergenerational stability and mobility of educational attainment.

Education is a central concept in the social stratification and mobility literature as it is both a key process of social reproduction and a source of social mobility. While sociologists have traditionally focused on economic and cultural explanations of intergenerational transmission of education, there is an increasing awareness of the role of genetic inheritance (Conley *et al.* 2015; Guo and Stearns 2002; Nielsen 2006; Nielsen and Roos 2015; Nielsen 2008). Importantly, parents and children share 50% of their DNA. Because of that, parental influences on children are often ambiguously social and genetic. To obtain an accurate estimate of the social effects, it is important to identify and adjust for the genetic influences (Conley *et al.* 2015; Liu and Guo 2016).

The heritability of educational attainment is estimated to be around 40% on average (Branigan *et al.* 2013). As Jencks (1980) pointed out more than three decades ago, however,

heritability estimates only set an upper limit on educational attainment traceable to genetic variation, they cannot separate exogenous environmental variance (i.e., variation in educationalattainment due to environmental factors independent of genotypes) and endogenous environmental variance (i.e., variation in educational attainment due to environmental factors that vary in response to genotypes). Therefore, heritability estimates can provide limited use for public policy as they cannot be used to assess effects of creating new environments.

Advances in genomic science and technology have provided sociologists opportunities to reinvestigate genetic and environmental contributions to intergenerational transmission of educational attainment. Molecular genetic data are now available from thousands of genetically independent individuals. These data allow us to examine influences of the whole human genome on phenotypes based on observed genetic measures without making assumptions in heritability studies (e.g., equal environments for identical and fraternal twins, absence of assortative mating, and generalizability of the twin/family sample).

Taking advantage of genome-wide genotype data in the Health and Retirement Study, this chapter examines the role of genes and the environment in multigenerational processes of educational attainment. I first conceptualize a model of genetic and environmental influences on multigenerational processes of educational attainment. I then test this model using genome-wide data in conjunction with educational measures from three generations in HRS.

4.2 Multigenerational Transmission of Educational Attainment

4.2.1 Intergenerational influences

There is a long tradition in sociology to assess intergenerational influences based on statistical associations between parents' and children's social status (Blau and Duncan 1967; Featherman and Hauser 1976; Hout 1984; Hout 1988; Mare 1993). Large intergenerational

associations mean that the offspring's socioeconomic status is largely dependent upon their family background, or in other words, children are mostly likely to obtain the same level of socioeconomic status as their parents and less likely to fall far behind or surpass their parents. Smaller intergenerational associations indicate more independence of offspring's status from their family background, and therefore more opportunities for social mobility.

Sociologists have typically focused on economic and cultural explanations of intergenerational transmission of education. The economic perspective emphasizes family resources' direct (e.g., tuition) and indirect (e.g., forgone earnings) influences on education (Bailey and Dynarski 2011; Blossfeld and Shavit 1993; Boudon 1974; Reardon 2011). Accordingly, higher levels of family economic resources should be associated with higher levels of educational attainment. The cultural perspective highlights the importance of cultural capitals such as norms, values, attitudes, expectations (Bourdieu and Passeron 1977; De Graaf and Ganzeboom 1993; Di Maggio 1982; Sakamoto *et al.* 2009). Transmitted within the family, these cultural capitals are necessary for educational success. Compared to those from culturally disadvantaged groups, children from culturally advantaged groups are more likely to continue their education and obtain higher levels of educational attainment.

In addition to the socioeconomic and cultural explanations, there is a growing realization in the role of biological mechanisms in intergenerational transmission of educational outcomes (Conley *et al.* 2015; Guo and Stearns 2002; Mare 2011; Nielsen 2006; Nielsen and Roos 2015; Nielsen 2008; Turkheimer *et al.* 2003). There is no such an "education gene" that determines one's educational attainment. Yet genes can influence education indirectly through various pathways. Genetic influences on education can be mediated by psychological characteristics including intelligence, self-control, and interpersonal skills. Also, some genes may affect

education through behaviors such as geographic mobility and mate choice (Belsky et al. 2016). Importantly, genetic effects and socioenvironmental effects are correlated and entangled (Jaffee and Price 2007; Wagner *et al.* 2013). If the genetic effects are ignored, estimates of pure socioenvironmental effects are likely to be biased (Conley *et al.* 2015; Liu and Guo 2016).

It has been a challenge to estimate genetic effects as gene effects were not directly observable. Recent advances in molecular genetics provide social scientists an opportunity to address this issue (Okbay *et al.* 2016; Rietveld *et al.* 2013a). This study takes the opportunity to disentangle social and biological mechanisms in intergenerational transmission of educational attainment. Moreover, this study extends the analysis of intergenerational influences on educational attainment to three generations.

4.2.2 Multigenerational processes

Research of intergenerational transmission of socioeconomic status (SES) has been extended to investigate the effects of grandparents and other family members. There are two different models on grandparents' influence on grandchildren: Markovian model and non-Markovian model. The Markovian model assumes that grandparents' influences on grandchildren are completely medicated by parents' SES. Several empirical studies have provided evidence for the Markovian model. For example, in their landmark book "The new American grandparent: A place in the family, a life apart," Cherlin and Furstenberg (1992) conclude that grandparents do not play a pivotal role directly influencing their grandchildren. Using data from the Wisconsin Longitudinal Studies, two studies show limited grandparental effects on grandchildren's educational success (Jæger 2012; Warren and Hauser 1997). Yet findings from other studies suggest that grandparents can influence their grandchildren independent of parents. Research has shown multiple pathways through which grandparents can

directly affect their grandchildren's education, including providing financial assistance (Aldous 1995), offering advice (Cherlin and Furstenberg 1992), and monitoring grandchildren's school activities (Deleire and Kalil 2002).

Other than the socioeconomic pathways, biological mechanisms may also play an important role in the mutigenerational processes. Since grandparent and grandchildren share 25% of their DNA, associations in traits between grandparents and grandchildren might be partially ascribed to genetic inheritance. A unique grandparents' effect is through genetic transmission in the female line. As a woman's eggs are made shortly after conception, the eggs that contain her children's DNA spend about nine months in their grandmother's uterus. This allows for indirect environmental influences of grandmothers on their grandchildren's genetic and epigenetic processes, and therefore health and development in the long term (Gluckman *et al.* 2008). Although the biological pathway of multigenerational influence is well-established, it has not been incorporated in studies of social stratification and mobility.

4.2.3 A model of multigenerational transmission of educational attainment

In this section, I conceptualize a model of multigenerational transmission of educational attainment. This model focuses on genetic and environmental contributions to multigenerational processes of transmission of education. To simplify, this model is based on the Markovian assumption.

A parent's genes can contribute to his/her child's education in at least two different ways. The parent' DNA is inherited by the child and the inherited genes may influence the child's education through pathways as stated above. Moreover, the parent's genes may influence the parent's educational status which, in turn, interact with exogenous socioeconomic experiences over time shaping the child's education (Dickens and Flynn 2001; Jencks 1980; Scarr and

McCartney 1983).While the social pathway is part of sociologists' interest, the existence of the biological pathways may bias the estimates of social influences in intergenerational models.

Effects of grandparents on children's education are even more complicated. First, a quarter of a grandparent' DNA is directly transmitted to his/her grandchildren and these genes may influence their education attainment (Bio + Bio). Second, grandparents' genes inherited by parents may play a role in creating an endogenous environment for grandchildren (Bio + Soc). Third, grandparents' genes influence their own education, and therefore provide a context for parents, which will, in turn, influence grandchildren later (Soc + Soc).

Figure 4.1 demonstrates multigenerational transmission of educational status. Suppose $g_{i,j}$ represents the genetic association between generation i and j; $p_{i,j}$ represents the environmental association in education between generation i and j; a_i represents the genetic association with education in generation i. Assume $r_{i,j}$ is the overall educational correlation between generation i and j. Then, the educational correlation between generation 1 (G1) and generation 2 (G 2) can be expressed as:

$$r_{1,2} = p_{1,2} + a_1 * g_{1,2} * a_2$$

The educational correlation between generation 1 (G1) and generation 3 (G3) is:

$$r_{1,3} = p_{1,2} * p_{2,3} + a_1 * g_{1,2} * a_2 * p_{2,3} + a_1 * g_{1,2} * g_{2,3} * a_3$$

As an extension, the educational correlation between generation 1 (G1) and generation n (Gn) is:

$$r_{1,n} = p_{1,2} * p_{2,3} * \ldots * p_{n-1,n} + a_1 * g_{1,2} * a_2 * p_{2,3} * \ldots * p_{n-1,n} + \ldots + a_1 * g_{1,2} * \ldots * g_{n-2,n-1} *$$

$$a_{n-1} * p_{n-1,n} + a_1 * g_{1,2} * \ldots * g_{n-1,n} * a_n$$

Assume random mating (i.e., $g_{1,2} = g_{2,3} = \ldots = g_{n-1,n} = g$), the same genetic association with education across generations (i.e., $a_1 = a_2 = \ldots = a_n = a$), and the same environmental association in education between two consecutive generations (i.e., $p_{1,2} = p_{2,3} = \ldots = p_{n-1,n} = p$), then

$$r_{1,n} = p^{n-1} + a^2 * g * p^{n-2} + \ldots + a^2 * g^{n-1} = p^{n-1} + a^2 * \sum_{i=1}^{n-1} g^i * p^{(n-1)-i}$$

The proportion of the genetic contribution (through both biological and social pathways) to educational correlation between G1 and Gn is:

$$\frac{a^2 * \sum_{i=1}^{n-1} g^i * p^{(n-1)-i}}{p^{n-1} + a^2 * \sum_{i=1}^{n-1} g^i * p^{(n-1)-i}}$$

The ratio of the genetic contribution (through both biological and social pathways) to the environmental contribution is:

$$\frac{a^2 * \sum_{i=1}^{n-1} g^i * p^{(n-1)-i}}{p^{n-1}} > \frac{a^2 * g^{(n-1)}}{p^{n-1}} = a^2 * \left(\frac{g}{p}\right)^{(n-1)}$$

Due to assortative mating g is often greater than 0.5 and p is smaller than 0.5, the ratio will be larger as n increases.

4.3 Increasingly Available Molecular Genetic Data and Analytical Methods

The role of genes in human traits has been conventionally 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 outcomes (e.g., Boardman *et al.* 2010; Boardman *et al.* 2012; 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 directly observed, and genetic and environmental contributions are estimated as latent variables based on relatedness among genetic relatives. Also, twin/family studies rely on critical assumptions such as equal environments for identical and fraternal twins and absence of assortative mating. These assumptions have been 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.* 2008; 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 particular, 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). 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. Genome-wide genotype data of a similar scale have been collected in the National Longitudinal Study of Adolescent to Adult Health (Add Health). 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.

Researchers have attempted to identify specific genetic variants associated with educational attainment. In a pioneer study, Rietveld *et al.* (2013a) conduct a GWAS using 126,559 individuals. As a result, three SNPs are identified as significant predictors of educational attainment (p-value$<5\times10^{-8}$). In a more recent GWAS meta-analysis, Okbay *et al.* (2016) identify 74 genome-wide significant loci associated with the number of years of schooling completed using 293,732 individuals.

The genomic-relatedness-matrix restricted maximum likelihood estimation (GREML) method has been developed to assess contribution of the whole genome to phenotypes based on molecular genetic data. GREML allows the use of genetically unrelated samples to examine genetic, environmental effects, as well as gene-environmental interactions. This method has also been deployed for a variety of phenotypes, including height (Yang *et al.* 2010), BMI (Yang *et al.* 2011), 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).

The GREML approach has also been used to estimate the genetic correlation (rG) between different traits (Lee *et al.* 2012a). Deary *et al.* (2012) estimate a highly significant rG of .62 between intelligence in adolescence (age 11) and in late adulthood (age 65-78). In a more recent study, Boardman *et al.* (2015) examine the genetic correlation between educational attainment and health measures such as BMI, depression, and self-rated health. They find that the correlation between depression and education (rG = -.75) and between self-rated health and

education (rG = -.91) were largely explained by common genetic factors, but there was no evidence that BMI and education were influenced by common genetic factors (rG = -.03).

While bivariate GREML is useful in estimating genetic correlations, it does not directly assess intergenerational genetic mediation. Polygenic score analysis can be used to achieve this end. The first polygenic score analysis was conducted in a study of schizophrenia (Purcell *et al.* 2009), which find 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). This approach has been employed to advance sociological research. Liu and Guo (2015), for example, provide evidence that cumulative socioeconomic advantage significantly decreased an individual's BMI in middle to late adulthood only for those with a higher genetic propensity for obesity [measured by a polygenic score constructed using the 32 obesity-related SNPs in Speliotes *et al.* (2010)], but not for those with lower genetic propensities. In another recent study, Conley *et al.* (2015) decompose the intergenerational association in educational attainment into genetic and environmental components by including a polygenic score [constructed based on results in the study of Rietveld *et al.* (2013a)] as a predictor in the traditional intergenerational model. They find that genetic factors account for approximately a sixth and social inheritance accounts for five-sixths of the intergenerational association in educational attainment. The polygenic influence on educational attainment has been replicated using AddHealth (Domingue *et al.* 2015).

4.4 Aims of the Study

In this study, I assess the multigenerational transmission of education model using bivariate GREML and polygenic scores based on genome-wide genotype data in the Health and Retirement Study (HRS). There are four specific aims of this study.

Aim 1: Examine genetic and environmental contributions to the parent-child association in educational attainment. I estimate the exogenous environmental effects on children's education.

Aim 2: Extend Aim 1 to three generations by assessing genetic and environmental contributions to the grandparent-grandchild association in educational attainment.

Aim 3: Investigate the extent to which parents' genes influence children's educational attainment through biological and social pathways.

Aim 4: Examine genetic and environmental contributions to intergenerational mobility in educational attainment.

4.5 Data

Data for this study come from the Health and Retirement Study (HRS). HRS is a longitudinal study of Americans over age 50 conducted every two years from 1992 to 2012; it collects information on economic, health, social, and other factors relevant to aging and retirement. DNA samples were collected in 2006 and 2008. Of the collected samples, 13,129 were put into genotyping production using the Illumina Human Omni-2.5 Quad beadchip, with coverage of approximately 2.5 million single nucleotide polymorphisms (SNPs), and 12,507 passed the University of Washington Genetics Coordinating Center's (GCC) standardized quality control processes. To minimize confounding effects of population stratification, this study focuses on non-Hispanic whites.

HRS respondents provided information about their parents' and children's educational attainment. There are two issues in the data that might affect the analysis. First, some children

were not old enough to complete their highest grade when their parents were interviewed. Second, GREML analysis requires genetically unrelated individuals to obtain unbiased results, yet a number of HRS respondents had more than one child. To address these issues, I selected the oldest child in each household who was most likely to complete the highest grade of school, and excluded those who had not reached 30 years of age when the last wave of data was collected. Years of education is time-sensitive. Completion of twelve years of education is more common in more recent cohorts than in earlier ones. Also, women typically received less education than men in older generations. To address these issues, years of education is standardized by cohort and gender.

4.6 Methods

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

The genomic-relatedness-matrix restricted maximum likelihood (GREML) method was developed by Yang et al. (2011). In contrast to single-variant association analysis where each SNP is tested against an adjusted p-value, GREML treats all SNP effects as random effects. The basic GREML model can be described by the following equation:

$$Y=X\beta + W\mu + \varepsilon, \qquad \text{(Equation 4.1)}$$

where Y is the outcome variable; β is a vector of the coefficients of fixed covariates such as age, sex and other controls; μ is a vector of genetic effects with $\mu_i \sim N(0, \sigma^2_\mu)$, where i = 1,…, I, with I being the number of SNPs; ε is a vector of residual effects with $\varepsilon_j \sim N(0, \sigma^2_\varepsilon)$, where j = 1,…, J, with J being the number of individuals; W is a standardized genotype matrix. Yang et al. (2010) innovatively applied a previous result that has been known in animal genetics (Goddard et al. 2009). The result defines g = Wμ, $\mathbf{A} = WW^T/I$ and $\sigma^2_g = I\sigma^2_\mu$. Then Equation 2 is mathematically equivalent to Equation 1:

$$Y=X\beta + g + \varepsilon, \text{ with } Var(Y) = \mathbf{A}\sigma_g^2 + I_\varepsilon\sigma_\varepsilon^2, \qquad \text{(Equation 4.2)}$$

where g is a J*1 vector of the total genetic effects of the individuals with g ~ N (0, $\mathbf{A}\sigma_g^2$), **A** is the genomic relatedness matrix (GRM) and σ_g^2 is the total genetic variance explained by the SNPs. Hence σ_g^2 can be estimated by the restricted maximum likelihood (REML) approach, depending on the GRM estimated from all SNPs. The genetic contribution to the outcome can be assessed using the proportion of total variance in the outcome explained by all SNPs, which can be expressed as $\sigma_g^2/(\sigma_g^2 + \sigma_\varepsilon^2)$, and the environmental contribution can be expressed as $\sigma_\varepsilon^2/(\sigma_g^2 + \sigma_\varepsilon^2)$.

The GREML approach has been extended to estimate the genetic correlation (rG) between different traits (Lee et al. 2012a). The bivariate GREML model can be described by the following equation:

$$Y_t = X_t b_t + g_t + e_t,\ g_t \sim N(0, A\sigma_{gt}^2),\ e_t \sim N(0, \sigma_{et}^2), \qquad \text{(Equation 4.3)}$$

where Y_t is a vector of observations for trait t (t = 1 or 2). For example, Y_1 may represent respondents' educational attainment and Y_2 represents their children's educational attainment. Equation (4.4) demonstrates the covariance matrix between two traits:

$$Cov(Y_1, Y_2) = \begin{pmatrix} A\sigma_{g1}^2 + I\sigma_{e1}^2 & A\sigma_{g1g2} + I\sigma_{e1e2} \\ A\sigma_{g1g2} + I\sigma_{e2e1} & A\sigma_{g2}^2 + I\sigma_{e2}^2 \end{pmatrix}. \qquad \text{(Equation 4.4)}$$

The rG is defined as $\frac{\sigma_{g1g2}}{\sigma_{g1}\sigma_{g2}}$. A high rG between two traits indicates that SNP effects on two traits are relatively similar. Substantively speaking, an rG of 1 implies that the two traits are affected by the same genetic variants, so that any variation between the two traits is environmental. At the other extreme, an rG of 0 suggests genetically independent traits; in other words, the two traits are affected by completely different genetic variants.

Polygenic Score (PS)

Polygenic scores are typically constructed based on existing GWAS results. They can be calculated using the following formula:

$$PS_i = \sum_{j=1}^{J} \beta_j x_{ij}, \qquad \text{(Equation 4.5)}$$

where PS_i is the PS of individual i, β_j is the coefficient for variant j estimated using GWAS data, x_{ij} is the number of risk alleles on variant j that individual i possesses.

Polygenic scores in this study were constructed using the P-values and β-weights from the recent GWAS based on years of education (Okbay *et al.* 2016) according to the methods described by Dudbridge (2013) using the PRsice software (Euesden *et al.* 2015). Polygenic scores ranged from 2.13 to 113.60 and were normally distributed in HRS (mean = 58.54, sd =14.88). I standardized the scores to have mean = 0, sd = 1 for analysis. Greater polygenic scores are associated with higher levels of educational attainment.

4.7 Analytic Strategy

First, I performed a bivariate GREML analysis to assess the overall genetic correlations in education among three generations. I estimated the proportion of covariance in education between each pair of generations (i.e., G1 and G2, G2 and G3, and G1 and G3). This estimate can be interpreted as the genetic contribution to educational attainment through both biological and social pathways.

Second, I performed a polygenic score analysis to examine intergenerational genetic mediation of educational attainment. Specifically, I predicted G2's educational attainment using G1's educational attainment adjusting for the G2's polygenic score. This is a replication of analysis in the study of Conley et al. 2015 but is based on better-powered polygenic scores. This analysis focuses on the association between two generations' educational attainment for children with the same level of genetic propensity to education. It eliminates the biological pathway and

estimates the exogenous intergenerational association in educational attainment. This analysis was then stratified for father-son, father-daughter, mother-son, and mother-daughter pairs.

Third, I extended the polygenic score analysis to three generations. I predicted G3's educational attainment using G1's educational attainment adjusting for the G2's polygenic score. Since HRS only has genotype data from G2, I used G2's polygenic score as a proxy of the polygenic score of G3. This analysis was then stratified based on the gender combination of the grandparent (G1), the parent (G2), and the child (G3) (i.e., father's father, father, and son; father's mother, father, and son; mother's father, mother, and son; mother's mother, mother, and son; father's father, father, and daughter; father's mother, father, and daughter; mother's father, mother, and daughter; mother's mother, mother, and daughter).

Finally, I predicted G3's educational attainment using G2's educational attainment adjusting for G2's polygenic score. This analysis examined the association between two generations' educational attainment for parents with the same level of genetic propensity to educational attainment. By doing this, I assessed the extent to which parents' genes influence children's educational attainment through biological and social pathways. This analysis was also stratified for father-son, father-daughter, mother-son, and mother-daughter pairs.

4.8 Results

4.8.1 Bivariate correlations

Table 4.2 shows the Pearson's correlations in educational attainment among three generations. All the correlations are statistically significant (P<.001). The correlations in educational attainment between G1 and G2 and between G2 and G3 are around .40. This is consistent with the findings of Blau and Duncan (1967). The correlation between G1 and G3 (.17) is slightly lower than half the correlation between G1 and G2.

The polygenetic scores constructed based on G2's genotypes are significantly associated with educational attainment of all three generations. Assuming the level of gene expression is the same across generations, since one shares ½ DNA with his/her parent and child, the association between one's genes and his/her parent or child's phenotype is supposed to be half of the genetic association with his/her own phenotype if intergenerational genetic influences work only through biological pathways. Yet results in Table 4.2 shows that while the association between G2's polygenic score and G1' educational attainment (.14 for father and .12 for mother) is approximately half of the association between G2's polygenic scores and G2's educational attainment (.26), the association between G2's polygenic scores and G3' educational attainment (.19) is significantly greater than half of the association between G2's polygenic scores and G2's educational attainment. One possible explanation is that in addition to genetic transmission, parents' genes can influence their children's phenotype through non-biological pathways. Alternatively, this might be due to increasing educational assortative mating or genetic innovation in younger generations (i.e., some education-related genes are more likely to be expressed in younger generations than in older generations).

4.8.2 Genomic contribution to multigenerational association in educational attainment

Bivariate GREML models were estimated for father-child and mother-child pairs separately. As shown by Table 4.3, rGs in educational attainment among three generations are all significant at the .05 level. Consistent with results in Table 4.2, genetic correlations between G1 and G3 (rG=.464 between [G1] grandfather and [G3] grandchildren; rG= .393 between [G1] grandmother and [G3] grandchildren) are about, if not smaller than, half of the genetic correlations between G2 and G3 (rG=.940 between [G2] father and [G3] children; rG= .984 between [G2] mother and [G3] children). Genetic correlations between G2 and G3 are greater

than that between G1 and G2. Again, this might be because genotypes of G2 is linked to G1's phenotypes only through biological pathways, while genotypes of G2 can affect G3's phenotypes through both biological and non-biological pathways.

4.8.3 Genetic mediation of the association in educational attainment between parents and children

Table 4.4 shows results of assessing genetic mediation of the educational association between G1 and G2. As can be seen in Panel 1, the association in educational attainment between (G1) fathers and (G2) children reduces 10% after controlling for (G2) children's PGS, the association between (G1) mothers and (G2) children reduces 7% after controlling for (G2) children's PGS. Panels (2) and (3) display the results by G2's gender. Whereas associations between (G1) fathers and mothers and (G2) sons respectively reduce 6% and 6% after controlling for (G2) sons' PGS, associations between (G1) fathers and mothers and (G2) daughters respectively reduces 9% and 8% after controlling for (G2) daughters' PGS.

4.8.4 Genetic mediation of the association in educational attainment between grandparents and grandchildren

Table 4.5 demonstrates results of testing genetic mediation of the educational association between G1 and G3. The association in educational attainment between (G1) father's father and (G3) children reduces 14%, and the association between (G1) father's mother and (G3) children reduces 11% adjusting for father (G2)'s PGS. The association between (G1) mother's father and (G3) children reduces 13%, and the association between (G1) mother's mother and (G3) children reduces 12% adjusting for (G2) mother's PGS. The genetic contribution to educational association between G1 and G3 is greater than that between G1 and G2. This is consistent with the prediction of the multigenerational transmission of education model.

4.8.5 Social and biological pathways of the influence of parents' genes on children's educational attainment

The association between G2's PGS and G3's educational attainment reduces about 50% after controlling for G2's educational attainment. As results in Table 4.6 demonstrate, around half of parents' genetic influence on children's educational attainment is through biological pathways and the other half through non-biological pathways. This analysis is conducted for each possible parent-child gender combination: father and son (bio: 58%; soc: 42%), mother and son (bio: 52%; soc: 48%), father and daughter (bio: 52%; soc: 48%), mother and daughter (bio: 41%; soc: 59%). This suggests that father's genes influence son's education more through biological pathways, mother's genes influence daughter's education more through social pathways.

I replicate the analysis using univariate GREML models. The first model predicts G3' education using G2's whole-genome SNPs (see columns 3 and 4 in Table 4.7). As a result, 34% of variance in G3's educational attainment can be predicted by G2's genomic information. The second univariate GREML models predicts G3'education using G2's whole-genome SNPs adjusting for G2's education. The results show the SNP heritability reduces to 15%. Similarly, this suggests that half of parental genetic influence on children's education is through biological pathways, and the other half through socioenvironmental pathways.

4.8.6 Genetic and environmental contributions to intergenerational mobility in educational attainment

Self PGS still significantly predicts educational attainment adjusting for both parents' education (see Table 4.4). This suggests that genes also play an important role in educational mobility. I replicated the analysis using univariate GREML models (see columns 1 and 2 in

Table 4.7). The first model predicts G2' education using G2's whole-genome SNPs. As a result, 40% of variance in G2's educational attainment can be predicted by the SNPs. The second univariate GREML model is adjusted for G1's education (both father and mother). The results show the SNP heritability reduces to 23%, about half of the estimate in the first model. This says, for those with the same level of parental education, genetic factors account for 23% of the remaining variation in education.

The differences in the estimated genetic contribution to education may be due to different number of observations used in estimating the two models (e.g., missing values in G1's education). To test this possibility, a robustness test is conducted using observations without missing values in both G1 and G2's education. The results remain. Therefore the difference in estimated genetic contribution to education between the two models is not driven by missing values.

4.9 Discussions and Conclusion

This study makes important contributions to the social stratification and mobility literature. First, one of the biggest challenges in social stratification and mobility research is to address previously unmeasured genetic/biological confounding variables in the estimation of the intergenerational influence. In this study, I conceptualize a model of multigenerational influence and test the model using educational measures from three generations in conjunction with genome-wide genotype data in HRS. I find significant genetic correlations in educational attainment across three generations. This suggests genetic factors play an important role in stabilizing intergenerational educational attainment in the U.S. Importantly, the parent-child association in education is reduced around 6-10% adjusting for (either parents' or children's) polygenic scores.

Secondly, I extend the two-generation model to three or more generations. Bivariate GREML results show that genetic correlation in education between grandparents and grandchildren is about half of the genetic correlation between parents and children. I also find that the grandparent-grandchild association in education reduces 10-15% adjusting for parents' polygenic scores. This finding is consistent with the prediction of the multigenerational transmission model, namely that genes play a larger role in educational association between more distal generations.

Thirdly, I examine the biological and social pathways through which parents' genes influence children's educational attainment. Results from both polygenic score and GREML models provide evidence that about half of parent's genetic influence can be ascribed to genetic transmission and the other half is medicated by parents' education. Moreover, I stratify the analysis by gender. Around 60% of the association between fathers' polygenic scores and sons' education is mediated by fathers' education, while only 40% of the association between mothers' polygenic scores and daughters' educational is medicated by mothers' education. One interesting speculation is that fathers' genes influence sons mainly through biological pathways, whereas mothers' genes influence daughters mainly through socioenvironmental pathways. This finding needs to be replicated using other independent samples.

Finally, this study also provides evidence for genetic contribution to intergenerational mobility in educational attainment. GREML results show genome-wide SNPs of G2 can still predict 23% variation in G2's education adjusted for G1's education. It is possible that some genes that were silent in the previous generation are expressed in the subsequent generation. Genes do not determine educational destiny, instead, they provide opportunities for change.

Genetic influences are largely intertwined with the environment. Genes may affect people by modifying the environment in various ways.

Some limitations need to be acknowledged. First, the polygenic score approach often suffers from lacking of power in assessing genetic variation. According to GREML results, the genome-wide SNPs explains around 40% of variation in education, while the polygenic scores only result in a Rsq of 6.6%. This discrepancy may be because the true GWAS coefficients differ across cohorts, perhaps due to heterogeneity in phenotype measurement or gene-environment interactions, or the failure of some of the assumptions underlying the calculation of Rsq (Okbay *et al.* 2016). Second, genomic data are only available for G2. Because of that, it is impossible to estimate the genetic association with education for G1 and G3. Third, minority samples are not included in the analysis. Although genetic information is available for some minorities (e.g., blacks and Asians), their sample size is insufficient to achieve adequate statistical power for separate analysis. Future research can extend the analysis in this study to other racial populations when data become available.

Despite these limitations, this study demonstrates how incorporating genetic information into social science research can help enrich our understanding of critical social issues. Molecular genetic data are increasingly available in large-scale datasets (e.g., the Fragile Families Study, the Framingham Heart Study, the National Longitudinal Study of Adolescent to Adult Health, and the Wisconsin Longitudinal Study), providing researchers unprecedented opportunities to study interactive influences of socioenvironmental factors and genetic factors on sociological outcomes.

REFERENCES

Aldous, Joan. 1995. "New Views of Grandparents in Intergenerational Context." *Journal of Family Issues* 16(1):104-122.

Bailey, Martha J. and Susan M. Dynarski. 2011. "Inequality in Postsecondary Education." Pp. 117–131 in *In Whither Opportunity?*, edited by G. J. Duncan and R. J. Murnane. New York: Russell Sage Foundation.

Benjamin, Daniel. J. , Magnus Johannesson, Peter Visscher, M., David Cesarini, Matthijs. J. H. M. van der Loos, *et al.* 2012. "The Genetic Architecture of Economic and Political Preferences." *Proceedings of the National Academy of Sciences* 109(21):8026-8031.

Blau, P. M. and O. D. Duncan. 1967. *The American Occupational Structure*. New York: Wiley.

Blossfeld, Hans-Peter and Yossi Shavit. 1993. "Persisting Barriers: Changes in Educational Opportunities in Thirteen Countries." Pp. 1-23 in *Persistent Inequality: Changing Educational Attainment in Thirteen Countries.*, edited by Y. Shavit and H.-P. Blossfeld. Boulder, CO: Westview Press.

Boardman, Jason D., Casey L. Blalock, and Fred C. Pampel. 2010. "Trends in the Genetic Influences on Smoking." *Journal of Health and Social Behavior* 51(1):108-123.

Boardman, Jason D., Benjamin W. Domingue, and Jonathan Daw. 2015. "What Can Genes Tell Us About the Relationship between Education and Health?" *Social Science and Medicine* 127:171-80.

Boardman, Jason D., Michael E. Roettger, Benjamin W. Domingue, Matthew B. McQueen, Brett C. Haberstick, *et al.* 2012. "Gene–Environment Interactions Related to Body Mass: School Policies and Social Context as Environmental Moderators." *Journal of Theoretical Politics* 24(3):370-388.

Boudon, Raymond. 1974. *Education, Opportunity, and Social Inequality.* New York: John Wiley.

Bourdieu, Pierre and Jean-Claude Passeron. 1977. *Reproduction in Education, Society, and Culture.* Beverly Hills, CA: Sage.

Branigan, Amelia R., Kenneth J. McCallum, and Jeremy Freese. 2013. "Variation in the Heritability of Educational Attainment: An International Meta-Analysis." *Social Forces* 92(1):109-140.

Caspi, Avshalom, Antony Braithwaite, Richie Poulton, Karen Sugden, Terrie E. Moffitt, *et al.* 2003. "Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-Htt Gene." *Science (New York, N.Y.)* 301(5631):386-389.

Caspi, Avshalom, Ahmad R. Hariri, Andrew Holmes, Rudolf Uher, and Terrie E. Moffitt. 2010. "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* 167(5):509-527.

Caspi, Avshalom, Joseph McClay, Terrie E. Moffitt, Jonathan Mill, Judy Martin, *et al.* 2002. "Role of Genotype in the Cycle of Violence in Maltreated Children." *Science* 297(5582):851-854.

Chabris, Christopher F., Benjamin M. Hebert, Daniel J. Benjamin, Jonathan P. Beauchamp, David Cesarini, *et al.* 2012. "Most Reported Genetic Associations with General Intelligence Are Probably False Positives." *Psychological Science* 23(11):1314-1323.

Charney, Evan and William English. 2012. "Candidate Genes and Political Behavior." *American Political Science Review* 106(1):1-34.

Cherlin, A. J. and F. F. Furstenberg. 1992. *The New American Grandparent: A Place in the Family, a Life Apart.* Cambridge, MA: Harvard University Press.

Conley, Dalton, Benjamin W. Domingue, David Cesarini, Christopher Dawes, Cornelius A. Rietveld , *et al.* 2015. "Is the Effect of Parental Education on Offspring Biased or Moderated by Genotype?" *Sociological Science* 2:82-105.

Davies, G., M. Luciano, K. McGhee, L. Lopez, A. J. Gow, *et al.* 2011. "Genome-Wide Association Studies Establish That Human Intelligence Is Highly Heritable and Polygenic." *Molecular Psychiatry* 16(10):996-1005.

Daw, Jonathan, Michael Shanahan, Kathleen Mullan Harris, Andrew Smolen, Brett Haberstick, *et al.* 2013. "Genetic Sensitivity to Peer Behaviors: 5httlpr, Smoking, and Alcohol Consumption." *Journal of Health and Social Behavior* 54(1):92-108.

De Graaf, Paul M. and Harry B. G. Ganzeboom. 1993. "Family Background and Educational Attainment in the Netherlands for the 1891–1960 Birth Cohorts." in *Persistent Inequality: Changing Educational Attainment in Thirteen Countries*, edited by Y. Shavit and H.-P. Blossfeld. Boulder, CO: Westview Press.

Deary, Ian J., Janie Corley, Paul Redmond, Helen C. Fox, Suzanne J. Rowe, *et al.* 2012. "Genetic Contributions to Stability and Change in Intelligence from Childhood to Old Age." *Nature* 482(7384):212-215.

Deleire, Thomas and Ariel Kalil. 2002. "Good Things Come in Threes: Single-Parent Multigenerational Family Structure and Adolescent Adjustment." *Demography* 39(2):393-413.

Di Maggio, Paul. 1982. "Cultural Capital and School Success: The Impact of Status Culture Participation on the Grades of Us High School Students." *American Sociological Review* 47(2):189-201.

Dickens, William T. and James R. Flynn. 2001. "Heritability Estimates Versus Large Environmental Effects: The Iq Paradox Resolved." *Psychological Review* 108(2):346-369.

Domingue, Benjamin W, Daniel W. Belsky, Dalton Conley, Kathleen Harris, and Jason D Boardman. 2015. "Polygenic Influence on Educational Attainment: New Evidence from the National Longitudinal Study of Adolescent to Adult Health." *AERA Open* 1(3):1-13.

Dudbridge, Frank. 2013. "Power and Predictive Accuracy of Polygenic Risk Scores." *PLoS genetics* 9(3):e1003348.

Duncan, Laramie E. and Matthew C. Keller. 2011. "A Critical Review of the First 10 Years of Candidate Gene-by-Environment Interaction Research in Psychiatry." *American Journal of Psychiatry* 168(10):1041-1049.

Euesden, J., C. M. Lewis, and P. F. O'Reilly. 2015. "Prsice: Polygenic Risk Score Software." *Bioinformatics* 31(9):1466-1468.

Featherman, David L. and Robert M. Hauser. 1976. "Changes in the Socioeconomic Stratification of the Races, 1962-73." *American journal of sociology* 82(3):621-651.

Gluckman, Peter D., Mark A. Hanson, Cyrus Cooper, and Kent L. Thornburg. 2008. "Effect of in Utero and Early-Life Conditions on Adult Health and Disease." *The New England journal of medicine* 359(1):61-73.

Goldberger, Arthur S. 1979. "Heritability." *Economica* 46(184):327-347.

Guo, Guang, Michael E. Roettger, and Tianji Cai. 2008. "The Integration of Genetic Propensities into Social-Control Models of Delinquency and Violence among Male Youths." *American Sociological Review* 73(4):543-568.

Guo, Guang, Michael E. Roettger, and Jean C. Shih. 2007. "Contributions of the Dat1 and Drd2 Genes to Serious and Violent Delinquency among Adolescents and Young Adults." *Human Genetics* 121(1):125-136.

Guo, Guang and Elizabeth Stearns. 2002. "The Social Influences on the Realization of Genetic Potential for Intellectual Development." *Social Forces* 80(3):881-910.

Hindorff, L. A. , J (European Bioinformatics Institute) MacArthur, J (European Bioinformatics Institute) Morales, H. A. Junkins, P. N. Hall, *et al.* "A Catalog of Published Genome-Wide Association Studies.", vol. 2016. www.genome.gov/gwastudies.

Hirschhorn, Joel N. and Mark J. Daly. 2005. "Genome-Wide Association Studies for Common Diseases and Complex Traits." *Nature Reviews Genetics* 6(2):95-108.

Hout, Michael. 1984. "Status, Autonomy, and Training in Occupational Mobility." *American Journal of Sociology* 89(6):1379-1409.

Hout, Michael. 1988. "More Universalism, Less Structural Mobility: The American Occupational Structure in the 1980s." *American Journal of Sociology* 93(May 88):1358-1400.

Jæger, Mads Meier. 2012. "The Extended Family and Children's Educational Success." *American Sociological Review* 77(6):903-922.

Jaffee, S. R. and T. S. Price. 2007. "Gene-Environment Correlations: A Review of the Evidence and Implications for Prevention of Mental Illness." *Molecular Psychiatry* 12(5):432-442.

Jencks, Christopher. 1980. "Heredity, Environment, and Public Policy Reconsidered." *American Sociological Review* 45(5):723-736.

Lee, S. H., J. Yang, M. E. Goddard, P. M. Visscher, and N. R. Wray. 2012a. "Estimation of Pleiotropy between Complex Diseases Using Single-Nucleotide Polymorphism-Derived Genomic Relationships and Restricted Maximum Likelihood." *Bioinformatics* 28(19):2540-2542.

Lee, S. Hong, Teresa R. DeCandia, Stephan Ripke, Jian Yang, Patrick F. Sullivan, *et al.* 2012b. "Estimating the Proportion of Variation in Susceptibility to Schizophrenia Captured by Common Snps." *Nature Genetics* 44(3):247-U35.

Liu, Hexuan and Guang Guo. 2015. "Lifetime Socioeconomic Status, Historical Context, and Genetic Inheritance in Shaping Body Mass in Middle and Late Adulthood." *American Sociological Review* 80(4):705-737.

—. 2016. "Opportunities and Challenges of Big Data for the Social Sciences: The Case of Genomic Data." *Social Science Research* 59(2016):13-22.

Locke, Adam E., Bratati Kahali, Sonja I. Berndt, Anne E. Justice, Tune H. Pers, *et al.* 2015. "Genetic Studies of Body Mass Index Yield New Insights for Obesity Biology." *Nature* 518(7538):197-U401.

Mare, Robert D. 1993. "Educational Stratification on Observed and Unobserved Components of Family Background." in *Persistent Inequality: Changing Educational Attainment in Thirteen Countries*, edited by Y.Shavit and H. P. Blossfeld. Boulder, CO: Westview.

—. 2011. "A Multigenerational View of Inequality." *Demography* 48(1):1-23.

Mitchell, C., J. Brooks-Gunn, I. Garfinkel, S. McLanahan, D. Notterman, *et al.* 2015. "Family Structure Instability, Genetic Sensitivity, and Child Well-Being." *American Journal of Sociology* 120(4):1195-1225.

Mitchell, Colter, Daniel Notterman, Jeanne Brooks-Gunn, John Hobcraft, Irwin Garfinkel, *et al.* 2011. "Role of Mother's Genes and Environment in Postpartum Depression." *Proceedings of the National Academy of Sciences of the United States of America* 108(20):8189-8193.

Nielsen, François. 2006. "Achievement and Ascription in Educational Attainment: Genetic and Environmental Influences on Adolescent Schooling." *Social Forces* 85(1):193-216.

Nielsen, François and J. Micah Roos. 2015. "Genetics of Educational Attainment and the Persistence of Privilege at the Turn of the 21st Century." *Social Forces* 94(2):535-561.

Nielsen, François. 2008. "The Nature of Social Reproduction: Two Paradigms of Social Mobility." Pp. 1-35 in *Sociologica*.

Okbay, Aysu, Jonathan P. Beauchamp, Mark Alan Fontana, James J. Lee, Tune H. Pers, *et al.* 2016. "Genome-Wide Association Study Identifies 74 Loci Associated with Educational Attainment." *Nature* advance online publication.

Purcell, Shaun M., Naomi R. Wray, Jennifer L. Stone, Peter M. Visscher, Michael C. O'Donovan, *et al.* 2009. "Common Polygenic Variation Contributes to Risk of Schizophrenia and Bipolar Disorder." *Nature* 460(7256):748-752.

Reardon, Sean F. 2011. "The Widening Achievement Gap between the Rich and the Poor: New Evidence and Possible Explanations." Pp. 91-115 in *In Whither Opportunity?*, edited by G. J. Duncan and R. J. Murnane. New York: Russell Sage Foundation.

Rietveld, Cornelius A., Arpana Agrawal, Johan G. Eriksson, Eva Albrecht, Behrooz Z. Alizadeh, *et al.* 2013a. "Gwas of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment." *Science* 340(6139):1467-1471.

Rietveld, Cornelius A., Robert F. Krueger, Meike Bartels, David A. Collier, Daniel J. Benjamin, *et al.* 2013b. "Molecular Genetics and Subjective Well-Being." *Proceedings of the National Academy of Sciences* 110(24):9692-9697.

Risch, Neil, Richard Herrell, Thomas Lehner, Kung-Yee Liang, Lindon Eaves, *et al.* 2009. "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* 301(23):2462-2471.

Sakamoto, Arthur, Kimberly Goyette, and Chang Hwan Kim. 2009. "Socioeconomic Attainments of Asian Americans." *Annual Review of Sociology* 35:255-176.

Scarr, Sandra and Kathleen McCartney. 1983. "How People Make Their Own Environments: A Theory of Genotype → Environment Effects." *Child Development* 54(2):424-435.

Simons, Ronald L., Man Kit Lei, Steven R. H. Beach, Gene H. Brody, Robert A. Philibert, *et al.* 2011. "Social Environment, Genes, and Aggression: Evidence Supporting the Differential Susceptibility Perspective." *American Sociological Review* 76(6):883-912.

Simonson, Matthew A., Amanda G. Wills, Matthew C. Keller, and Matthew B. McQueen. 2011. "Recent Methods for Polygenic Analysis of Genome-Wide Data Implicate an Important Effect of Common Variants on Cardiovascular Disease Risk." *BMC Medical Genetics* 12(1):146-146.

Speliotes, Elizabeth K., Reedik Mägi, Thomas A. Buchanan, Joshua C. Randall, Sailaja Vedantam, *et al.* 2010. "Association Analyses of 249,796 Individuals Reveal 18 New Loci Associated with Body Mass Index." *Nature Genetics* 42(11):937-48.

Thorleifsson, G., C. N. Palmer, T. Johnson, J. Luan, J. H. Park, *et al.* 2010. "Hundreds of Variants Clustered in Genomic Loci and Biological Pathways Affect Human Height." *Nature* 467(7317):832-838.

Turkheimer, Eric, Andreana Haley, Mary Waldron, Brian D'Onofrio, and Irving I. Gottesman. 2003. "Socioeconomic Status Modifies Heritability of Iq in Young Children." *Psychological Science* 14(6):623-628.

Vinkhuyzen, A. A. E., M. Luciano, A. Payton, M. Horan, W. Ollier, *et al.* 2012. "Common Snps Explain Some of the Variation in the Personality Dimensions of Neuroticism and Extraversion." *Translational psychiatry* 2(4):e102.

Wagner, Brandon, Jiang Li, Hexuan Liu, and Guang Guo. 2013. "Gene-Environment Correlation: Difficulties and a Natural Experiment-Based Strategy." *American Journal of Public Health* 103 Suppl 1(Journal Article):S167-73.

Warren, John Robert and Robert M. Hauser. 1997. "Social Stratification across Three Generations: New Evidence from the Wisconsin Longitudinal Study." *American Sociological Review* 62(4):561-572.

Wood, A. R., T. Esko, J. Yang, S. Vedantam, T. H. Pers, *et al.* 2014. "Defining the Role of Common Variation in the Genomic and Biological Architecture of Adult Human Height." *Nature Genetics* 46(11):1173-1186.

Yang, J., B Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, *et al.* 2010. "Common Snps Explain a Large Proportion of the Heritability for Human Height." *Nature Genetics* 42(7):565-569.

Yang, Jian, M. Geoffrey Hayes, William G. Hill, Maria Teresa Landi, Alvaro Alonso, *et al.* 2011. "Genome Partitioning of Genetic Variation for Complex Traits Using Common Snps." *Nature Genetics* 43(6):519-525.

CHAPTER 5: CONCLUDING REMARKS

This dissertation demonstrates how social theories and genomics can be integrated to improve our understanding of sociological issues such as delinquency and violence, and social stratification and mobility. Chapter 1 overviews the opportunities provided by the recently available genomic data to social science researchers. It highlights 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.

Chapters 2, 3, and 4 investigate three substantive sociological issues combining genomic data and conventional social science measures. Chapter 2 focuses on the interaction of social environment and genetic factors on delinquency and violence. Specifically, it examines how genetic influences on delinquent and violent behaviors differ between individuals who live under adverse social environments (e.g., low attachment to family and school, disadvantaged neighborhoods, etc.) and those who experience favorable social environments (e.g., high attachment to family and school, advantaged neighborhoods, etc.) using data from Add Health. Chapter 3 assesses genetic and environmental contributions to socioeconomic stability and mobility over the life course. It examines genetic and environmental influences on socioeconomic measures at different life stages (e.g., educational attainment, occupational status, wealth, etc.). Chapter 4 investigates the extent to which genetic and environmental factors contribute to stability and changes in educational attainment across generations. These two chapters take advantage of the longitudinal design of HRS and genome-wide data in HRS.

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 delinquency and educational attainment, 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. 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. There is a recent awareness of considering group-level social and cultural processes in the investigation of gene-environment interactions. 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.

Table 2.1 Variable Description

Variable Name	Description	Mean or Proportion	SD
Delinquency and Violence			
Wave I	Serious Delinquency Score, Wave I	.155	.288
Wave III	Serious Delinquency Score, Wave III	.066	.167
Wave I	Violence Score, Wave I	.155	.311
Wave III	Violence Score, Wave III	.058	.169
Demographics			
Age	Respondent's age at the time of Wave I	16.152	1.738
Female	Respondent's gender	.505	
Parenting Factors			
High parental attachment	High emotional attachment to resident parents at Wave I	.524	
Low parental attachment	Low emotional attachment to resident parents at Wave I	.476	
Strict parental supervision	Strict parental supervision at Wave I	.501	
Weak parental supervision	Weak parental supervision at Wave I	.499	
School Factors			
High school attachment	High emotional attachment to school at Wave I	.586	
Low school attachment	Low emotional attachment to school at Wave I	.414	
Strict school discipline	Strict school discipline at Wave I	.518	
Low school discipline	Weak school discipline at Wave I	.482	
Neighborhood			
High education	Respondent lives in higher education blocks at Wave I	.500	
Low education	Respondent lives in lower education blocks at Wave I	.500	
High income	Respondent lives in higher income blocks at Wave I	.500	
Low income	Respondent lives in lower income blocks at Wave I	.500	
Low unemployment rate	Respondent lives in blocks with lower unemployment rate at Wave I	.500	
High unemployment Rate	Respondent lives in blocks with higher unemployment rate at Wave I	.500	
Low single/no parent household rate	Respondent lives in blocks with lower single/no-parent household rate at Wave I	.500	
High single/no parent household rate	Respondent lives in blocks with higher single/no-parent household rate at Wave I	.500	

Table 2.2 Genomic Contribution to Serious Delinquency and Violence and Standard Errors

	Serious Delinquency (Wave III)	Violence (Wave III)
Genomic contribution (Proportion of total variance explained by SNPs)	.046(.094)	.000(.094)
Intercept	.704(.195)***	.412(.190)***
Female	-.067(.005)***	-.065(.005)***
Age	-.069(.024)**	-.031(.024)
Age^2	.002(.001)*	.001(.001)
Prior delinquency or violence (Wave I)	.148(.009)***	.128(.009)***
N	4075	4088

Note: The genomic contribution is estimated by mixed linear models. Models are fit using the genome-wide complex trait analysis (GCTA) software package developed by Yang et al. (2010). The models also include the first 10 principle components as covariates to account for population stratification.
*p≤ .05; ** p≤ .01; *** p≤ .001 (two-tailed tests)

Table 2.3 Genomic Contribution to Serious Delinquency and Violence under High-Social-Control and Low-Social-Control Conditions

	Serious Delinquency (Wave III)				Violence (Wave III)			
	High-Social-Control Conditions		Low-Social-Control Conditions		High-Social-Control Conditions		Low-Social-Control Conditions	
	Collective Genetic Contribution (1)	Number of Persons (2)	Collective Genetic Contribution (3)	Number of Persons (4)	Collective Genetic Contribution (5)	Number of Persons (6)	Collective Genetic Contribution (7)	Number of Persons (8)
Parenting Factors								
Parental attachment	.000	1955	.217	2054	.061	1962	.112	2060
Parental supervision	.000	1920	.177	2088	.000	1926	.029	2095
School Factors								
School attachment	.000	1325	.258	1703	.195	1327	.336	1710
School discipline	.000	1785	.474	1217	.086	1791	.668	1220
Neighborhood								
High education	.000	2131	.036	1900	.077	2138	.000	1906
High income	.000	2253	.049	1773	.000	2260	.000	1779
Low unemployment	.000	2496	.340	1495	.000	2503	.110	1501
Low single pare. rate	.000	2696	.229	1276	.000	2705	.058	1280

Note: The genomic contribution is estimated by mixed linear models controlling for gender, age, age^2, Wave I delinquency or violence, and first 10 PCs.

Table 2.4 Gene-Environment Correlation

	Serious Delinquency (Wave III)	Violence (Wave III)
Parenting Factors		
Parental attachment	-.374(.365)	-.240(.497)
Parental supervision	.451(.424)*	1.000(1.430)**
School Factors		
School attachment	-.110(.109)	-.377(.432)**
School discipline	-.179(.157)	-.022(.242)
Neighborhood		
High education	.211(.114)**	.089(.177)
High income	.044(.092)	-.172(.304)
Low unemployment	.122(.115)	-.105(.253)
Low single pare. rate	-.095(.097)	-.361(.566)*

*p≤ .05; ** p≤ .01; *** p≤ .001 (Likelihood Ratio Test)

Table 3.1 Summary Statistics of Key Variables in the Genetic Sample

Variable	Mean (SD)	Median
Education (years of school)	13.16 (2.57)	12.00
Occupation (occupational prestige score)	45.36 (12.86)	37.00
Household Wealth (thousand dollars)	542.69 (1872.47)	252.00

Table 3.2 Bivariate Correlation between Standardized Key Variables (Standard Error)

	Male	Female	Overall
Education and Occupation	.474(.013)	.488(.013)	.483(.008)
Education and Wealth	.305(.013)	.295(.013)	.301(.010)
Occupation and Wealth	.259(.015)	.190(.015)	.224(.013)

Note: Three life-course socioeconomic achievement measures are recoded into relative indicators based on the basis of the 9 deciles of each measure.

Table 3.3 Genetic and Environmental Contributions to Life-course Socioeconomic Achievement Outcomes

	Males			Females			Males and Females		
	Education	Occupation	Wealth	Education	Occupation	Wealth	Education	Occupation	Wealth
Genetic Variance	2.732(1.305)	1.832(1.046)	1.458(1.154)	3.694(1.013)	1.043(.941)	2.608(.887)	3.763(.598)	2.221(.535)	2.553(.525)
Residual Variance	5.851(.911)	4.241(.732)	5.695(.812)	5.406(.701)	5.099(.663)	5.312(.618)	5.267(.411)	4.147(.372)	5.181(.366)
Phenotypic variance	8.582(.452)	6.073(.358)	7.153(.095)	9.100(.367)	6.141(.321)	7.921(.319)	9.030(.237)	6.368(.201)	7.734(.205)
SNP Heritability	.318(.137)	.302(.156)	.204(.151)	.406(.097)	.170(.145)	.329(.100)	.417(.057)	.349(.061)	.330(.061)
logL	-5317.078	-4047.185	-5082.129	-7576.101	-5204.096	-7335.415	-12934.148	-9299.131	-12464.753
logL0	-5319.376	-4048.845	-5082.870	-7583.060	-5204.684	-7340.098	-12956.196	-9308.098	-12477.769
LRT	4.598	3.320	1.481	13.918	1.177	9.366	44.097	17.935	26.033
df	1	1	1	1	1	1	1	1	1
p-value	0.01601	0.03422	.1118	9.549e-05	.139	.001	1.563e-11	1.143e-05	1.678e-07
N	3519	3022	3531	4957	3800	4981	8476	6822	8512

Note: All models control for the largest 10 principal components for adjusting population stratification.

Table 3.4 Genetic Correlations among Life-course Socioeconomic Achievement Measures

	Males			Females			Males and Females		
	T1: Education T2: Occupation (1)	T1: Education T2: Wealth (2)	T1:Occupation T2: Wealth (3)	T1: Education T2: Occupation (4)	T1: Education T2: Wealth (5)	T1:Occupation T2: Wealth (6)	T1: Education T2: Occupation (4)	T1: Education T2: Wealth (5)	T1:Occupation T2: Wealth (6)
Genetic variance									
T1	2.717(1.305)	2.910(1.295)	1.677(1.033)	3.723(1.013)	3.708(1.011)	1.192(.937)	3.764(.598)	3.811(.596)	2.235(.530)
T2	2.005(1.013)	1.590(1.145)	1.454(1.154)	1.198(.889)	2.675(.882)	2.601(.887)	2.211(.509)	2.581(.520)	2.578(.525)
Cov(T1, T2)	1.190(.903)	2.151(.903)	.070(.798)	1.717(.741)	2.829(.700)	.920(.659)	1.918(.429)	2.599(.413)	1.029(.382)
Residual variance									
T1	5.861(.911)	5.716(.903)	4.355(.724)	5.387(.700)	5.396(.699)	5.001(.659)	5.266(.411)	5.234(.410)	4.144(.368)
T2	4.111(.708)	5.614(.804)	5.697(.812)	4.965(.626)	5.267(.614)	5.317(.618)	4.137(.354)	5.161(.362)	5.164(.365)
Cov(T1, T2)	2.230(.631)	.580(.630)	1.490(.562)	2.109(.517)	.158(.483)	.619(.461)	1.859(.296)	.317(.284)	.669(.266)
Phenotypic variance									
T1	8.579(.451)	8.627(.450)	6.032(.353)	9.109(.367)	9.104(.367)	6.194(.321)	9.031(.237)	9.045(.237)	6.380(.200)
T2	6.117(.349)	7.204(.391)	7.152(.392)	6.163(.307)	7.942(.318)	7.919(.319)	6.348(.194)	7.742(.203)	7.741(.205)
SNP Heritability									
T1	.316(.137)	.337(.134)	.278(.156)	.409(.097)	.407(.097)	.192(.142)	.417(.057)	.421(.057)	.350(.074)
T2	.328(.149)	.221(.148)	.203(.151)	.194(.136)	.337(.099)	.329(.100)	.348(.071)	.333(.060)	.333(.060)
rG Test									
rG	.510(.267)	1.000(.389)	.045(.501)	.813(.265)	.898(.197)	.522(.370)	.665(.100)	.829(.112)	.429(.144)
df	1	1	1	1	1	1	1	1	1
p-value (rG=0)	0.09334	0.008778	.4661	0.01073	2.076e-05	0.08372	2.406e-06	5.057e-11	.003469
p-value (rG=1)	0.01997	0.5	.01032	0.2856	0.3086	.1867	0.001	0.071	.0003
N	3271	3525	3277	4379	4969	4391	7649	8494	7667

Table 3.5 Genetic Contribute to Socioeconomic Mobility

	Males		Females		Males and Females	
	Residuals (Occupation~ Education)	Residuals (Wealth~ Education + Occupation)	Residuals (Occupation~ Education)	Residuals (Wealth~ Education + Occupation)	Residuals (Occupation~ Education)	Residuals (Wealth~ Education + Occupation)
Genetic Variance	1.479(.789)	1.076(1.135)	.400(.703)	.144(.984)	1.254(.404)	1.503(.561)
Residual Variance	3.270(.552)	5.152(.800)	4.153(.498)	6.289(.701)	3.509(.284)	5.136(.396)
Phenotypic variance	4.749(.272)	6.228(.381)	4.554(.238)	6.433(.333)	4.763(.150)	6.639(.208)
SNP Heritability	.311(.150)	.173(.173)	.088(.150)	.224(.152)	.263(.078)	.226(.078)
logL	-3667.223	-4146.509	-4683.321	-5378.302	-8399.241	-9574.165
logL0	-3669.312	-4146.947	-4683.478	-5378.313	-8404.333	-9578.048
LRT	4.791	0.875	.313	0.021	10.184	7.766
df	1	1	1	1	1	1
p-value	0.02046	0.1748	0.288	0.4418	0.0007084	0.002662
N	3015	3015	3794	3794	6809	6809

Table 3.6 Associations between Intelligence-related Polygenic Scores and Socioeconomic Achievement Outcomes

	Education	Occupation	Wealth
Males	.171(.045)***	.093(.041)*	.123(.042)**
Females	.179(.039)***	.106(.038)**	.067(.037)
Males and Females	.176(.030)***	.095(.028)***	.089(.028)**

Table 4.1 Summary Statistics of Key Variables in the Genetic Sample

Variable	Mean(SD)	Median
Generation 1's (Grandfather's) Education (years of school)	10.00(4.03)	11.52
Generation 1's (Grandmother's) Education (years of school)	10.38(3.71)	11.52
Generation 2's (Self) Education (years of school)	13.16 (2.57)	12.00
Generation 3's (Children's) Education (years of school)	14.06(2.19)	14.00
Generation 2's Polygenic Scores on Education	58.54 (14.88)	58.87

Note: Education is standardized by birth cohort and gender.

Table 4.2 Bivariate Correlation between Standardized Key Variables (95% Confidence Interval)

	(1)	(2)	(3)	(4)	(5)
Generation 1 Grandfather' Education (1)	1				
Generation 1 Grandmother' Education (2)	.646(.633, .658)	1			
Generation 2 Self Education(3)	.373(.354, .391)	.357(.339, .376)	1		
Generation 3 Children' Education(4)	.164(.140, .188)	.165(.141, .188)	.390(.370, .410)	1	
Generation 2's Polygenic Scores (5)	.142(.121, .163)	.120(.099, .140)	.256(.237,.275)	.192(.169,.214)	1

Table 4.3 Bivariate GREML Results for Educational Attainment

	T1:Generation 1 Father T2:Generation 2 Self	T1:Generation 1 Mother T2:Generation 2 Self	T1:Generation 1 Grandfather T2:Generation 3 Grandchildren	T1:Generation 1 Grandmother T2:Generation 3 Grandchildren	T1:Generation 2 Self (Father) T2:Generation 3 Children	T1:Generation 2 Self (Mother) T2:Generation 3 Children
Genetic variance						
T1	.388(.061)	.314(.049)	.373(.062)	.304(.049)	.300(.155)	.277(.083)
T2	.258(.043)	.260(.043)	.261(.068)	.272(.068)	.218(.187)	.230(.102)
Cov(T1, T2)	.259(.039)	.184(.035)	.145(.047)	.113(.042)	.240(.130)	.248(.071)
Residual variance						
T1	.493(.042)	.429(.035)	.497(.042)	.432(.034)	.404(.107)	.337(.057)
T2	.389(.030)	.388(.030)	.524(.048)	.516(.048)	.608(.131)	.516(.071)
Cov(T1, T2)	.074(.037)	.092(.024)	.018(.032)	.030(.029)	.084(.090)	.057(.049)
Phenotypic variance						
T1	.880(.024)	.743(.020)	.870(.024)	.736(.020)	.704(.052)	.614(.029)
T2	.648(.017)	.648(.017)	.785(.026)	.788(.026)	.825(.062)	.745(.036)
SNP Heritability						
T1	.440(.059)	.423(.056)	.429(.061)	.413(.058)	.426(.190)	.451(.116)
T2	.399(.057)	.402(.057)	.332(.078)	.345(.077)	.264(.208)	.308(.124)
rG Test						
rG	.817(.129)	.645(.089)	.464(.123)	.393(.134)	.940(.389)	.984(.217)
df	1	1	1	1	1	1
p-value	6.138e-13	5.255e-09	.0006	0.0022	.032	0.0001798
N	7958	8093	6895	7031	2268	3993

Note: All models control for the largest 10 principal components for adjusting population stratification.

Table 4.4 Genetic Mediation of the Association in Educational Attainment between (G1) Parents and (G2) Selves

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Males and Females							
G2 PS	.221(.009)***		.173(.008)***		.182(.008)***		.167(.008)***
G1 Father Edu		.324(.009)***	.290(.008)***			.201(.011)***	.182(.011)***
G1 Mother Edu				.326(.009)***	.302(.009)***	.182(.012)***	.174(.012)***
Rsq	.066	.139	.180	.128	.173	.150	.191
N	9301	8304	8304	8603	8603	8094	8094
Males							
		Father and Son		Mother and Son		Father, Mother, and Son	
G2 PS	.211(.014)***		.173(.013)***		.184(.013)***		.170(.013)***
G1 Father Edu		.309(.013)***	.289(.013)***			.212(.017)***	.197(.017)***
G1 Mother Edu				.296(.015)***	.279(.014)***	.147(.019)***	.141(.018)***
Rsq	.059	.131	.174	.100	.173	.138	.191
N	3909	3509	3509	3583	3583	3404	3404
Females							
		Father and Daughter		Mother and Daughter		Father, Mother and Daughter	
G2 PS	.228(.011)***		.172(.011)***		.178(.011)***		.163(.011)***
G1 Father Edu		.318(.011)***	.290(.011)***			.191(.015)***	.170(.015)***
G1 Mother Edu				.344(.012)***	.317(.011)***	.207(.016)***	.197(.015)***
Rsq	.071	.144	.184	.148	.191	.174	.210
N	5392	4795	4795	5020	5020	4690	4690

Table 4.5 Genetic Mediation of the Association in Educational Attainment between (G1) Grandparents and (G3) Children

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
		Father's Father and Child		Father's Mother and Child		Father's Parents and Child	
G2 PS	.195(.019)***		.178(.020)***		.177(.020)***		.175(.020)***
G1 Grandfather Edu		.136(.020)***	.117(.020)***			.088(.027)**	.074(.027)**
G1 Grandmother Edu				.137(.022)***	.122(.021)***	.075(.029)**	.069(.028)*
Rsq	.039	.018	.051	.016	.049	.020	.051
N	2563	2334	2334	2370	2370	2275	2275
		Mother's Father and Child		Mother's Mother and Child		Mother's Parents and Child	
G2 PS	.176(.014)***		.141(.015)***		.152(.015)***		.136(.015)***
G1 Grandfather Edu		.169(.015)***	.147(.015)***			.109(.020)***	.093(.020)***
G1 Grandmother Edu				.183(.015)***	.161(.015)***	.102(.021)***	.095(.021)***
Rsq	.035	.033	.053	.034	.059	.040	.060
N	4386	3888	3888	4078	4078	3809	3809

Table 4.6 Social and Biological Pathways of the Influence of (G2) Parents' Genes on (G3) Children's Educational Attainment

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Males and Females							
G2 PS	.184(.011)***	.195(.019)***		.109(.018)***	.176(.014)***		.084(.014)***
G2 Father Edu			.422(.020)***	.392(.021)***			
G2 Mother Edu						.439(.015)***	.413(.016)***
Rsq	.037	.039	.147	.180	.035	.156	.163
N	6949	2563	2563	2563	4386	4386	4386
Males							
			Father and Son			Mother and Son	
G2 PS	.194(.016)***	.207(.026)***		.121(.013)***	.187(.020)***		.098(.018)***
G2 Father Edu			.433(.028)***	.399(.013)***			
G2 Mother Edu						.451(.021)***	.424(.022)***
Rsq	.041	.045	.131	.166	.038	.168	.177
N	3572	1314	1314	1314	2258	2258	2258
Females							
			Father and Daughter			Mother and Daughter	
G2 PS	.171(.016)***	.183(.028)***		.096(.027)***	.164(.020)***		.067(.019)***
G2 Father Edu			.413(.029)***	.387(.029)***			
G2 Mother Edu						.426(.023)***	.402(.024)***
Rsq	.032	.033	.142	.151	.031	.143	.148
N	3377	1249	1249	1249	2128	2128	2128

Table 4.7 Genetic and Environmental Contributions to Intergenerational Mobility in Educational Attainment

	Generation 2	Generation 2 Adjusting for G1's Education (both father and mother)	Generation 3	Generation 3 Adjusting for G2's Education
Genetic Variance	.259(.043)	.110(.037)	.264(.069)	.097(.057)
Residual Variance	.389(.030)	.372(.026)	.521(.048)	.536(.040)
Phenotypic variance	.648(.017)	.482(.014)	.785(.026)	.633(.021)
SNP Heritability	.400(.058)	.228(.071)	.336(.078)	.153(.086)
logL	-1792.371	-712.734	-2021.843	-1538.212
logL0	-1812.427	-717.751	-2030.317	-1539.835
LRT	40.112	10.034	156.946	3.246
df	1	1	1	1
p-value	1.199e-10	0.0007686	1.923e-05	0.0358
N	8393	7380	6311	6310

Note: All models control for the largest 10 principal components for adjusting population stratification.

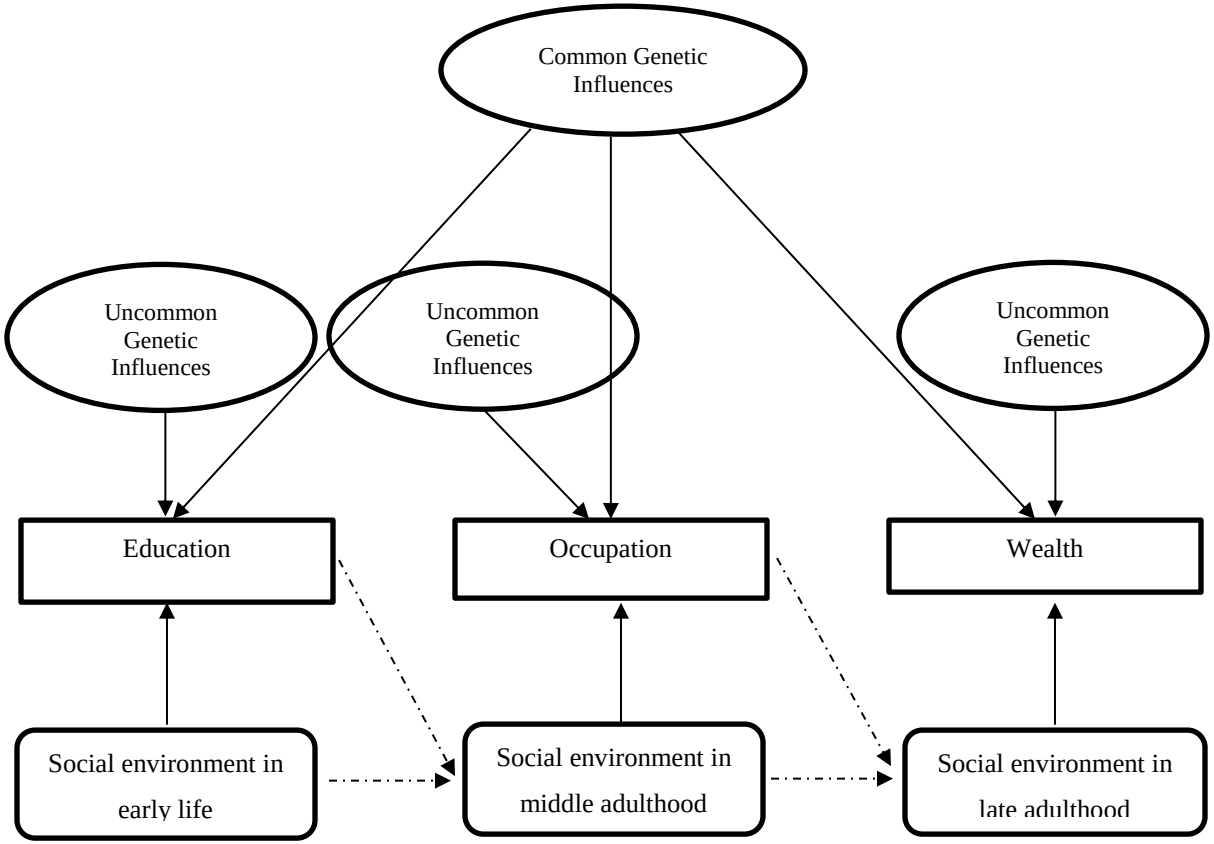

Figure 3.1 Social and Genetic Contributions to Lifetime Socioeconomic Achievement

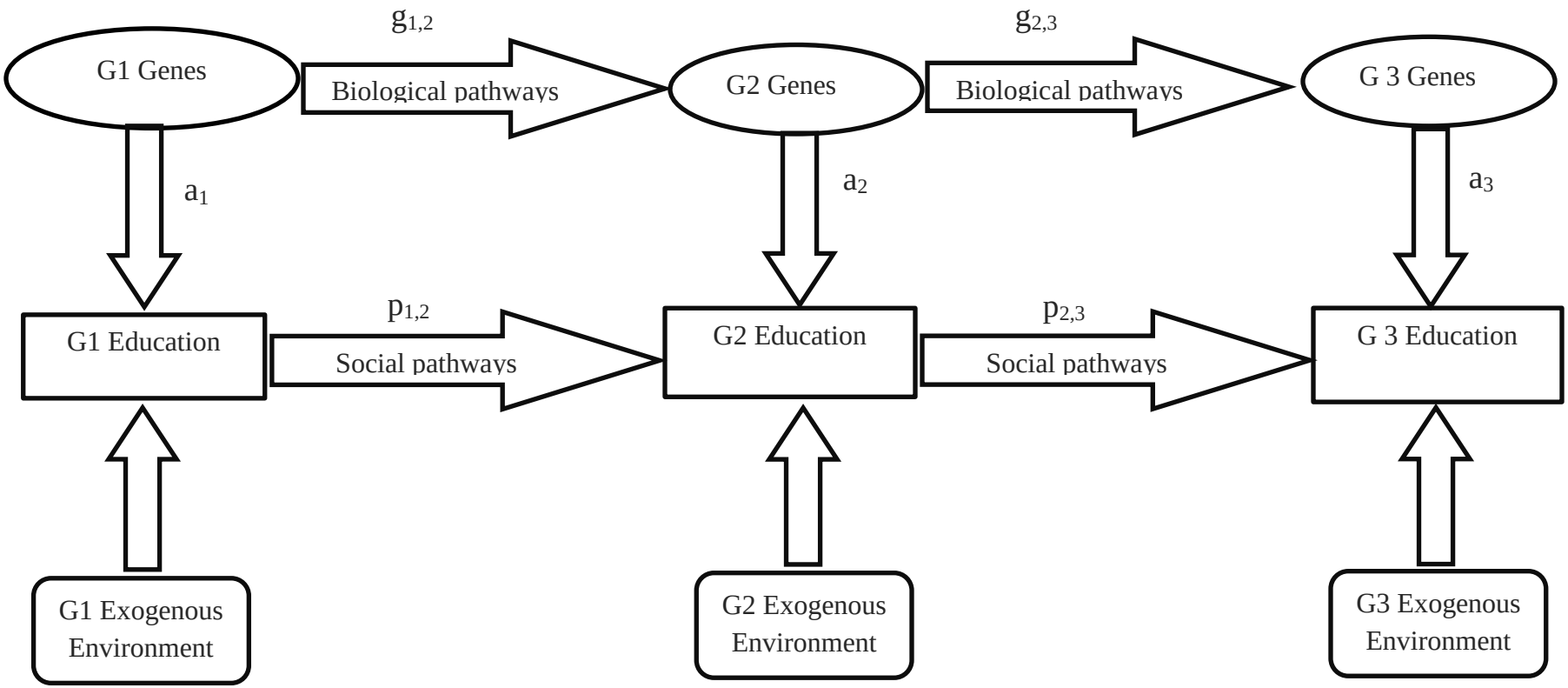

Figure 4.1 Multigenerational Transmission of Educational Attainment