ANNUAL REVIEWS

Annual Review of Sociology

Sociology, Genetics, and the Coming of Age of Sociogenomics

Melinda C. Mills[1] and Felix C. Tropf[2]

[1]Leverhulme Centre for Demographic Science, Department of Sociology, University of Oxford and Nuffield College, Oxford OX1 1JD, United Kingdom; email: melinda.mills@sociology.ox.ac.uk

[2]École Nationale de la Statistique et de L'administration Économique, Center for Research in Economics and Statistics, 91764 Palaiseu, France; email: felix.tropf@ensae.fr

Annu. Rev. Sociol. 2020. 46:553–81

First published as a Review in Advance on May 11, 2020

The *Annual Review of Sociology* is online at soc.annualreviews.org

https://doi.org/10.1146/annurev-soc-121919-054756

Keywords

sociogenomics, social science genetics, molecular genetics, behavior genetics, biosociology, gene-environment interaction, diversity

Abstract

Recent years have seen the birth of sociogenomics via the infusion of molecular genetic data. We chronicle the history of genetics, focusing particularly on post-2005 genome-wide association studies, the post-2015 big data era, and the emergence of polygenic scores. We argue that understanding polygenic scores, including their genetic correlations with each other, causation, and underlying biological architecture, is vital. We show how genetics can be introduced to understand a myriad of topics such as fertility, educational attainment, intergenerational social mobility, well-being, addiction, risky behavior, and longevity. Although models of gene-environment interaction and correlation mirror agency and structure models in sociology, genetics is yet to be fully discovered by this discipline. We conclude with a critical reflection on the lack of diversity, nonrepresentative samples, precision policy applications, ethics, and genetic determinism. We argue that sociogenomics can speak to long-standing sociological questions and that sociologists can offer innovative theoretical, measurement, and methodological innovations to genetic research.

INTRODUCTION

Although sociogenomics—the combination of genetics and sociology—has been deliberated for decades (Conley 2016, Eckland 1967, Freese 2008, Guo 2006, Udry 1995), dramatic advances have taken place in the last five years.[1] Radical drops in the cost of genome sequencing and growth in computational power precipitated an unprecedented explosion of data, novel methods, applications, and results (Conley & Fletcher 2017). As of 2020, markers for genetic loci[2] were isolated in over 4,500 studies (Mills & Rahal 2019, 2020). Traits (termed phenotypes in this literature) ranged from height (Wood et al. 2014) and body mass index (BMI) (Locke et al. 2015) to behaviors often studied by sociologists such as educational attainment (EDU) (Lee et al. 2018), fertility (Barban et al. 2016, Mathieson et al. 2020, Mills et al. 2020b), well-being (Baselmans et al. 2019), risky behavior (Karlsson Linnér et al. 2019), same-sex sexual behavior (Ganna et al. 2019), and substance use (Liu et al. 2019). For the first time in history, sociologists are able to ask and test fundamentally new types of questions and also falsify previous results (Mills 2019).

Sociology has had a fractious relationship with biology and genetics (Duster 2006, Freese & Shostak 2009). Udry (1995) attributed this tension to sociology's Durkheimian boundaries, where "the determining cause of a social fact should be sought among the social facts preceding it" [Durkheim 1938 (1895), p. 110)]—or in other words, self-imposing deliberate blinders to the nonsocial. While sociogenomic findings are increasingly embraced by top sociology journals (Gaydosh et al. 2018, Liu 2018, Wedow et al. 2018), some claim that it is a dangerous enterprise (Bliss 2018, Duster 2006).

Since the publication of previous reviews on genetics and sociology (Conley 2016, Freese & Shostak 2009), the field has experienced radical shifts in the availability of data, new methods, and topics under study. We do not explore previously covered topics such as geneticization, homophily, and the genetics of geography and race, yet there remains some inevitable overlap. Although this review is aimed at sociologists, it occupies an interdisciplinary space that demands scientific omnivores; we thus expand our review beyond sociology journals, albeit narrowing in on sociological topics. Readers desiring a primer in human genetics, evolution and migration, gene-environment interaction (G×E), and instructions on how to carry out applied statistical applications can refer to introductory texts specifically aimed at social scientists (Mills et al. 2020a).

We first chronicle the history of this research and key milestones and then focus on the more recent molecular genetic era. We then unpack the big data revolution, multiple genetic discoveries, and the rise of polygenic scores (PGSs) and their potential application in sociology. Key results are reviewed, with an emphasis on core traits of interest to sociologists. G×E and gene-environment correlation (rGE) theories are then introduced with exemplar applications in addition to the enduring challenges. We then balance our review with a critical reflection on the lack of diversity, nonrepresentative samples, ethics, genetic determinism, potential for policy applications, public understanding, and risks of sociogenomics. We conclude with anticipating the future of this research.

A BRIEF HISTORY OF BEHAVIOR GENETICS

Early Genetic Research

A brief history of human genetic research is illustrated in **Figure 1**. Early research from the mid- to late-1800s (Galton 1869) and twentieth century (Jensen 1968) focused largely on diseases but also

[1]The term sociogenomics was first used in 2005, but in relation to honeybees (Robinson et al. 2005).

[2]Genetic loci refer to locations on the genome, which are often single-nucleotide polymorphisms (SNPs). SNPs are the way we examine single-base differences or markers in DNA that allow us to examine variation in

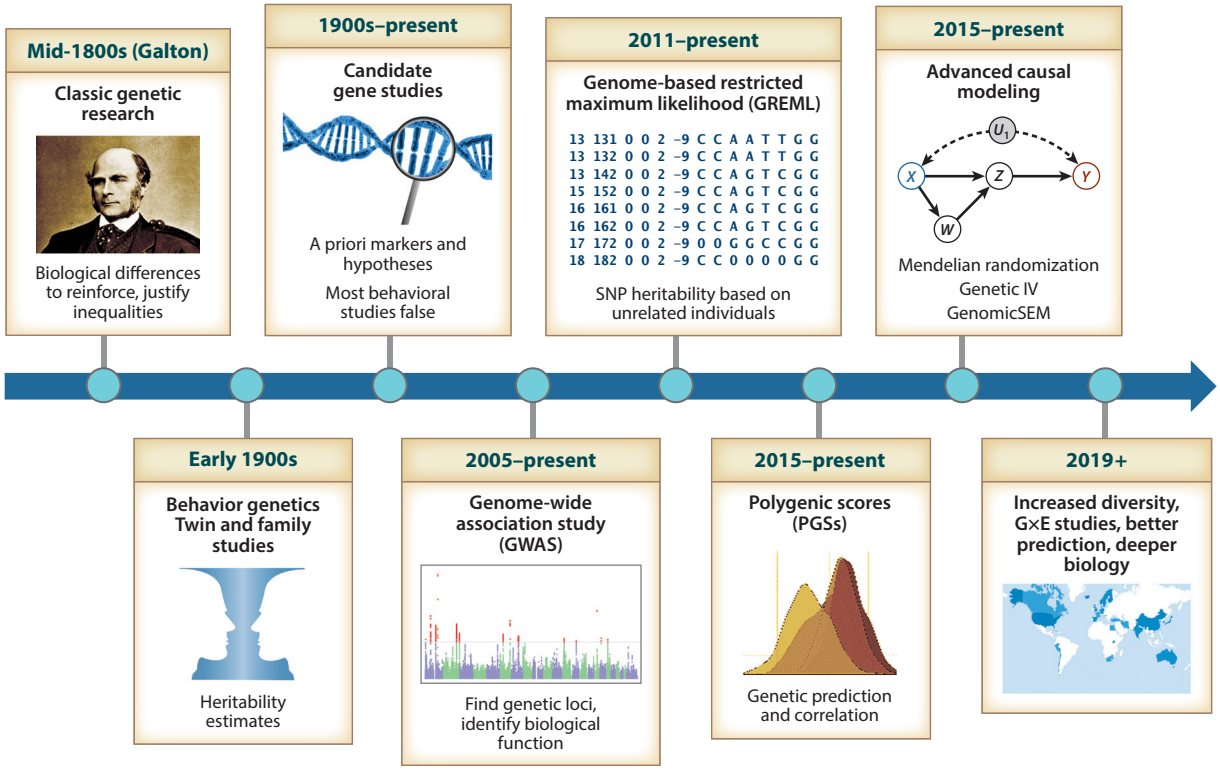

Figure 1

A brief history of genetics and sociogenomics from Galton, twin and family studies, candidate genes, genome-wide association studies, polygenic scores, causal modeling, and G×E (gene-environment interaction) to increased diversity and better prediction.

had a preoccupation with differences in cognitive ability and intelligence (Eckland 1967), often between racial and socioeconomic groups. This research generally stemmed from intelligence tests and attempted to use biological difference to justify punitive policies, resist desegregation and immigration, and reinforce racial and socioeconomic inequalities (Gillborn 2016, Martschenko et al. 2019).

Heritability and Twin Studies

Heritability forms the basis of much of our understanding of the genetic and socioenvironmental influences on outcomes and is defined as the proportion of variation of a trait in a population that is attributable to genetic differences. From the early 1900s, twin studies were used to derive twin heritability. Twin heritability is estimated by comparing a trait's correlation between identical [monozygotic (MZ)] twins, who share virtually 100% of their genetic material, with that of fraternal [dizygotic (DZ)] twins, who share ~50% of segregating genetic material. If a trait is heritable, holding environmental influences constant, individuals who are more genetically related should be more similar. An online atlas of 17,894 twin heritability estimates showed higher heritability for physical traits such as height (60%) and lower heritability for what are

a population. For a gentle introduction to terminology in genetics for social scientists, see Mills et al. (2020a, chapter 1).

often referred to as complex behavioral traits (e.g., 30% for societal values) (Polderman et al. 2015). Twin models make various assumptions, including no assortative mating, no G×E or rGEs, the equal environment assumption, and have other issues beyond the scope of this review. While some of these assumptions have been heavily criticized, sociological studies barely find evidence, for example, of a violation of the equal environment assumption (i.e., that MZ twins share environmental influences to the same extent as DZ twins) (Conley et al. 2013). We discuss alternative estimates of heritability based on molecular genetic data and relaxing key assumptions shortly.

Candidate Gene Studies

Early candidate gene studies focused on predefined loci (often fewer than 10) based on the loci's supposed biological functions. Such studies were conducted largely in psychology on traits such as externalizing and antisocial behavior (Caspi 2002), with an almost exclusive focus on neurotransmission (around 90% of studies) (Duncan et al. 2014), and also in sociology (Guo et al. 2008). A spate of early candidate gene studies were stimulated by the now infamous Caspi (2002) publication on the *5-HTTLPR* region and sensitivity to stressful life events. Most of these studies have been debunked (Duncan et al. 2014). Failure to replicate such studies was attributed to most hypotheses being wrong, increased risk of false positives due to statistical significance norms, seriously underpowered studies, and a strong publication bias toward positive findings. This situation even led the Editor of *Behavior Genetics* to write, "many of the published findings in the last decade are wrong or misleading" (Hewitt 2012, p. 1).

Candidate gene studies, however, can still be useful for certain traits or diseases such as Alzheimer's (*APOE*), obesity (*FTO*), and ability to metabolize alcohol (*ADH1B/ALDH2*), where effect sizes are large and associations are replicated. The *APOE* gene, for example, explains 13% of the heritability in Alzheimer's disease, whereas the presence of each risk allele translates into a fivefold increase in the risk of developing Alzheimer's (Adams et al. 2016). In contrast, in more complex traits associated with multiple genes (i.e., polygenic traits), such as educational attainment or reproductive behavior, clear candidate genes are uncommon. For example, the first three discovered single-nucleotide polymorphisms (SNPs, pronounced "snips") for one of the first EDU genome-wide association studies (GWASs) explained only 0.2% of the overall variance in that trait and an average increase in educational level by only one month per allele (Rietveld et al. 2013).

SOCIOGENOMICS AND THE DYNAMIC ERA OF MOLECULAR GENETICS

The birth of the GWAS (pronounced "gee-woz") in 2005 allowed researchers to identify associations between outcomes (phenotypes) and SNPs. SNPs are measures of single-base differences, or markers, and are sometimes referred to as genetic loci in DNA that allow us to examine genetic variation in a population. A GWAS is a method that adopts an unbiased, hypothesis-free approach to discover SNPs that are associated with a trait—a procedure sociologists often term as data-mining. To be considered a hit or discovery, the loci need to meet a significance threshold of $p < 5 \times 10^{-8}$ (i.e., $p < 0.00000005$), which is Bonferroni corrected for multiple testing.[3] A meta-analysis of all data is performed, often combining summary statistics from multiple data sets to obtain the largest sample possible. In 2018, the MTAG method (Multi-Trait Analysis of GWAS) (Turley et al. 2018) and genomic structural equation modeling (SEM) (Grotzinger et al. 2019)

[3]It is $p < 5 \times 10^{-8}$ since only a small fraction of loci are associated with the outcome, so it is akin to testing 700,000–800,000 variants due to strong correlations between genotypes at nearby genetic variants due to LD (see footnote 4).

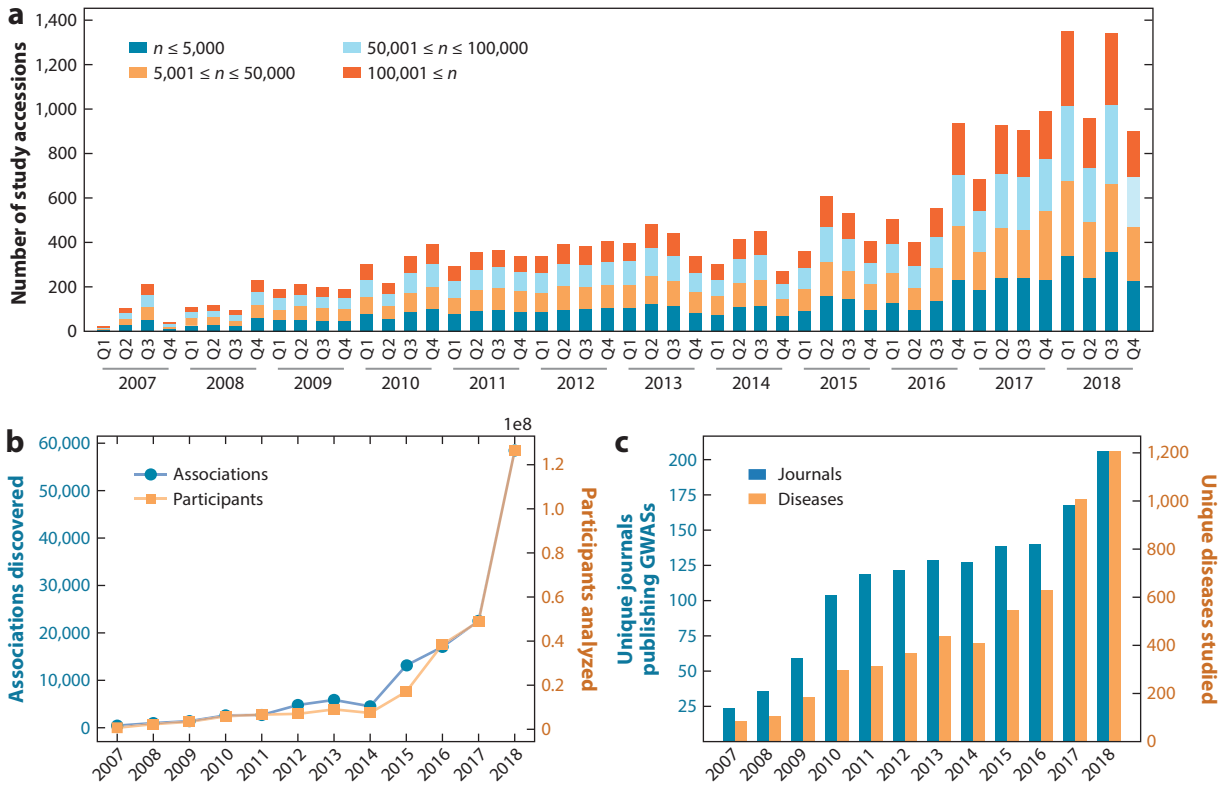

Figure 2

The growth of genome-wide association studies (GWASs) from 2007 to 2018. (*a*) Number of study accessions published per quarter over time, colored according to sample size to show the growth of larger ($100{,}001 \leq n$) GWASs. (*b*) A strong positive correlation between the number of associations found and the number of participants used in GWASs over time. (*c*) Growth in the number of unique traits examined as well as in the number of unique journals publishing GWASs over time. Figure adapted from Mills & Rahal (2019) under the Creative Commons Attribution (CC BY 4.0) International license, derived from National Human Genome Research Institute–European Bioinformatics Institute (NHGRI-EBI) GWAS Catalog, available at **https://www.ebi.ac.uk/gwas/**. Reuse of their data is subject to the terms of use for European Molecular Biology Laboratory–European Bioinformatics Institute (EMBL-EBI) services. For a daily updated version, see **https://www.gwasdiversitymonitor.com**, as described by Mills & Rahal (2020).

further advanced this approach by allowing joint analyses of summary statistics on related traits to increase the number of discovered SNPs and understand genetic correlations of related traits.

The Genome-Wide Association Study Big Data Revolution

Since 2005, momentous changes have occurred in the field of statistical genetics in the availability of the type and size of data, infusion of new phenotypes, more powerful genotyping and computational ability, and novel statistical methods. Rapid computational and technological progress meant that it was now possible to collect genetic data cheaply. Based on a previous review of nearly all 4,000 GWASs as of late 2018 (Mills & Rahal 2019), **Figure 2** shows the growth in GWASs over time by sample size, particularly since 2016, as well as associations and participants and the number of traits and journals where this research is published. Readers can also refer to an online, real-time dashboard to explore all GWASs to date by various characteristics over time (**https://www.gwasdiversitymonitor.com**, as described by Mills & Rahal 2020). Very large data

sources emerged, particularly as of 2017, namely the 500,000-sample UK Biobank (Bycroft et al. 2018), as well as participation in research by direct-to-consumer (DTC) genetic companies with large samples such as 23andMe. Since any particular SNP accounts for less than 0.03% of variation of a trait within a population for behavioral traits (Chabris et al. 2013), it was also recognized that large and statistically well-powered samples were required to detect more genetic variants and increase precision, particularly for complex polygenic traits (Dudbridge 2013).

The Rise of Polygenic Scores

Derived from GWAS summary statistics, a PGS is a single quantitative variable that summarizes an individual's genetic predisposition to a trait and is the main tool or variable that sociologists can use to integrate genetics into their research. Although there are monogenic diseases (i.e., one genetic variant causing the disease) such as Huntington's, the genetic architecture of complex behavioral traits that sociologists are often more interested in (e.g., education, fertility) is extremely polygenic. Polygenicity refers to multiple genetic variants (SNPs) being associated with the outcome, with each having a tiny association (Boyle et al. 2017). A PGS is a linear combination of multiple SNP effects on a trait across the whole genome, weighted by the GWAS effect sizes. Formally, a PGS is the weighted sum of a person's genotype at M loci. A PGS for individual i can be calculated as the sum of the reference or risk allele counts a_{ij} (0, 1, or 2) for each SNP $j = 1, \ldots M$, multiplied by a weight w_j,

$$PGS_i = \sum_{j=1}^{M} a_{ij} w_j,$$

where the weights w_j are transformations of GWAS coefficients. A PGS is not only linear but usually additive, because we weight each risk allele for each SNP included in the score equally, and given the large number of SNPs, the score follows a normal distribution. An introduction to PGSs and how to construct and use them can be found elsewhere (Dudbridge 2013, Euesden et al. 2014, Mills et al. 2020a, Ware et al. 2018).

Missing and Hidden Heritability

In the early 2010s, GREML (genome-based restricted maximum likelihood) analyses leveraged the investigation of the genetic similarity of unrelated individuals, transcending the need to use twin data for this purpose (Yang et al. 2011). Moving beyond the twin heritability estimates described previously, this allowed the calculation of SNP heritability, which is the proportion of trait variance of the total narrow-sense or additive heritability contribution of a trait based on all measured SNPs. In addition to this, there is also GWAS heritability, which is the proportion of variance accounted for by genetic variants derived from a GWAS discovery and combined into a PGS.

A general surprise was that the heritability estimates derived from GWASs were considerably lower than heritability estimates from twin and SNP heritability, prompting debates about missing or hidden heritability (Lee et al. 2011, Manolio et al. 2009, Tropf et al. 2017). Studies of same-sex sexual behavior, for instance, had a twin heritability of 40% and a SNP heritability of 8–25%, with the GWAS PGS explaining less than 1% (Ganna et al. 2019, Mills 2019). SNP heritability measures the potential ceiling of genetic explanatory power if the GWAS was statistically well powered enough to capture all polygenic effects of the measured SNPs, among other issues (Wray & Maier 2014). Sociologists have also shown that a key reason for the drop in heritability estimates

is due to the potential watering down of GWAS heritability, since GWAS meta-analyses are based on combining data from multiple historical periods, birth cohorts, and countries. This, in turn, introduces heterogeneity in genetic effects, most likely due to G×E (Tropf et al. 2017). Assuming homogeneity in the GWAS meta-analysis across often divergent time periods and institutional contexts highlights effects shared across them, while twin and SNP heritability estimates are typically based on less heterogeneous data, i.e., from one country or even specific birth cohorts, that also capture context-specific effects.

Heritability is also often misinterpreted. Crucially, it is not an individual-level estimate; rather, it is about the population from which calculations emanate, and it is also not equivalent to inheritance (Visscher et al. 2008). Since heritability estimates do not isolate specific genetic variants, they cannot be easily used to model individual or subgroup mechanisms, which is unsatisfying for sociologists' focus on underlying mechanisms. Low heritability also does not necessarily entail the absence of genetics (Visscher et al. 2008). Although the number of eyes you have is largely explained by genetics, since it is the same in everyone and has been suppressed by natural selection, there is no genetic variance and hence zero heritability. If there is variance in the number of eyes, it is only due to environmental causes (i.e., losing an eye). Heritability also does not relate to mean differences in or between populations. The Dutch grew 20 centimeters (almost 8 inches) over the past 200 years, and although height has a strong genetic basis and there is evidence of natural selection, most of the change is attributed to environmental shifts in diet and health (Stulp et al. 2015). Since estimates are population specific, they provide relatively little information on the effects of social environments such as family, education, employment and gender systems, welfare regimes, historical period, and birth cohort variation (de Vlaming et al. 2017, Tropf et al. 2017).

Genetic Correlations

Using a method called linkage disequilibrium (LD) score regression, which exploits the LD[4] structure of the data to show how likely different traits are to share the same genetic loci, researchers started to compare the summary statistics from GWAS discoveries to reveal genetic correlations and similarities across traits (Finucane et al. 2015). **Figure 3** illustrates correlations for the most recent GWAS of age at first birth (AFB) (Mills et al. 2020b) and EDU (Lee et al. 2018) for 22 related traits, grouped by various categories.[5] If the genetic correlation between two traits is 1, all genetic variance in trait 1 and 2 has a common base. If the genetic correlation is 0, the genetically based variance between trait 1 and 2 is independent.

Focusing on AFB, we see strong positive correlations and genetic overlap particularly with the behavioral trait of EDU [0.74, standard error (SE) 0.01] and to some extent openness to experience (0.19, SE 0.07) but also reproductive traits such as age at menarche (0.17, SE 0.03) and menopause (0.19, SE 0.03). There is also a strong positive correlation with AFB and age at initiation of smoking, with both variables capturing a unique window into adolescence and early adulthood (0.73, SE 0.03). Logically, there is a strong negative correlation with number of children ever born (NEB) (−0.69, SE 0.02), since later AFB is tied with lower overall levels of fertility and with traits such as ever smoking (−0.47, SE 0.02) and BMI (−0.27, SE 0.01).

[4]LD is a measure of the nonrandom association between alleles at different genetic loci at the same chromosome in a given population. An allele refers to each of the two or more alternative forms of a SNP, found at the same place on a chromosome, that arise by mutation. SNPs are in LD when the frequency of association of their alleles is higher than expected under random assortment. LD therefore concerns patterns of correlations between SNPs.

[5]Results are based on the most recent GWAS of all traits, with references available in Mills et al. (2020b).

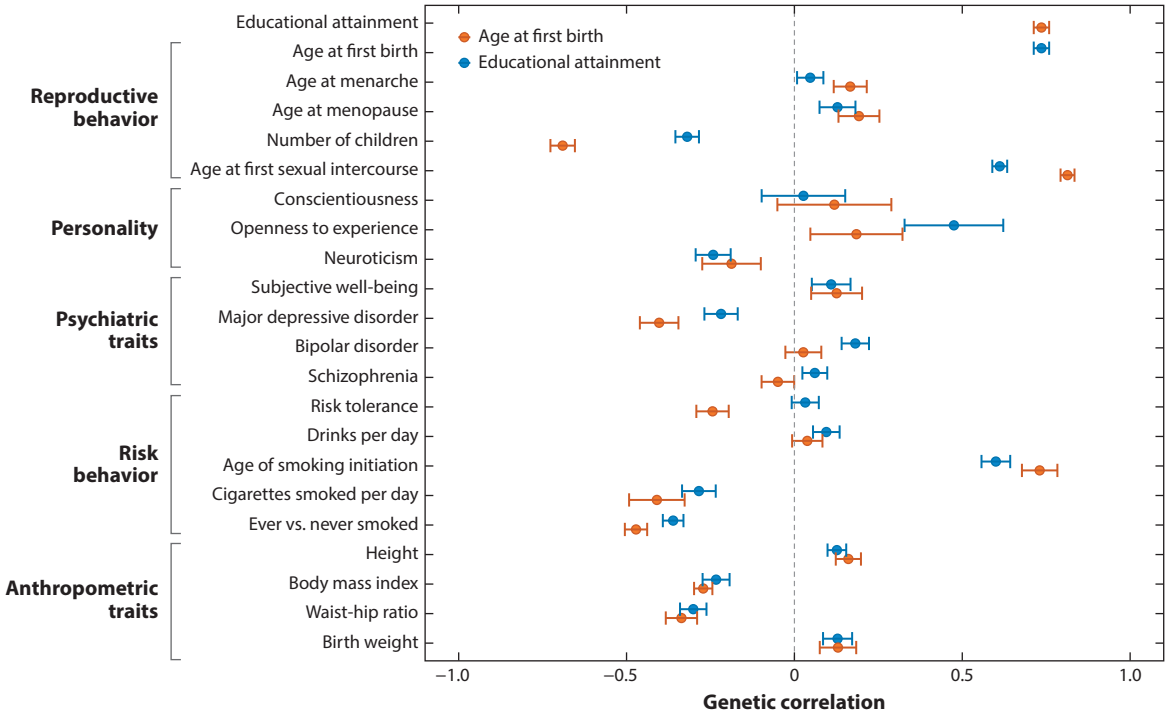

Figure 3

Genetic correlations of age at first birth (*orange*) and educational attainment (*blue*) with various traits. Horizontal bars represent 95% confidence intervals.

The genetic correlations between AFB and EDU have a striking overlap, explored in detail elsewhere (Mills et al. 2020b). We see a clustering of staying in education longer with postponing first birth; having fewer children; later sexual intercourse; being more open, less neurotic, or less depressed; higher well-being; being less likely to have ever smoked (and if they did, much later and less); taller; higher birth weight; and lower BMI and waist-hip ratio. With strong caveats about causality, population stratification, selection in environments, social mediators, and diversity of the data, an unforgiving interpretation is that we seem to have isolated middle-class genes. Crucially, a sociologist immediately sees that genetic correlations are not necessarily measuring only genetic effects. Correlations are likely picking up multiple environmentally mediated pathways and sample selection in addition to ubiquitous biological pleiotropy (see the next section). The fact that genetic results implicitly contain multiple socio-environmental factors could explain how lower socioeconomic status, for instance, is correlated with higher levels of substance use, risky and externalizing behavior, unhealthy eating, and the inability to realize one's educational potential, explored shortly in our section titled Gene-Environment Interplay.

Understanding Shared Genetic Architecture

A crucial element to understanding genetic correlations and interpreting PGS estimates is the shared genetic architecture of traits (van Rheenen et al. 2019, Young et al. 2019). Two aspects are important in this respect: genotype-phenotype (trait) associations and genetic pleiotropy across traits.

Genotype-phenotype associations. Genotype-phenotype associations can emerge for several reasons. First, there are direct effects, which are biological. For example, some SNPs isolated for late AFB and low number of children in men have been associated with lower sperm quality and mobility in downstream biological analyses (Barban et al. 2016).

Second, there might be environmental confounding due to population stratification, which refers to the population structure or patterns found in the genetic data that allow us to determine an individual's origins and that are often mirrored by geography (Conley 2016, Novembre et al. 2008). A common example used to explain this is the so-called chopstick gene (Hamer & Sirota 2000). If we ran a global GWAS on chopstick usage, we would identify variants more prevalent in Southeast Asia. This would not be due to direct genetic effects but rather rGE for cultural reasons and also differences in allele frequencies of geographically situated groups due to population stratification and historical migration patterns (Hamer & Sirota 2000). For the United States, state clustering has been observed for genetic variants associated with smoking and EDU (Domingue et al. 2016). Another study illustrates that genetic clustering in the United Kingdom reflects recent migration driven by socioeconomic status (Abdellaoui et al. 2019). In our upcoming section titled Gene-Environment Correlation, we discuss the related topic of social genetic effects and sorting.

Third, there might be indirect effects, or what has been termed genetic nurture (Kong et al. 2018). Using genetic data from parents and offspring, Kong et al. (2018) demonstrated that parents transfer not only their genetic material related to EDU. Rather, through nontransmitted genetic material, they also produce environmental conditions of nurture conducive to higher education and social mobility, discussed in more detail shortly. Finally, due to assortative mating such as by education, the effect of a causal variant may also capture these correlated inherited variants, potentially resulting in an overestimation of direct genetic effects (Robinson et al. 2017). This links with the literature on assortative mating in sociology related to population composition (Schwartz 2013), with some research already undertaken but considerably more room for sociologists to explore (Domingue et al. 2014).

Pleiotropy. Genetic correlations are related to pleiotropy in biology, which is when one gene is associated with two or more traits. Pleiotropy is now considered to be ubiquitous, leading some to argue that gene regulatory networks are so interconnected that all genes affect the functions of others, with pleiotropy so pervasive that we should think in terms of not polygenic but omnigenic models (Boyle et al. 2017). A recent study examining over 4,000 GWASs in 558 traits found that 90% of trait-associated genetic loci overlap with loci from multiple traits (Watanabe et al. 2019).

Three pleiotropy scenarios have been described: horizontal (biological), vertical (mediated), and spurious (van Rheenen et al. 2019). Horizontal pleiotropy refers to associations that could be spurious due to genes that have a common cause, or in other words, a model of when a specific gene has a direct effect on two traits. For example, certain SNPs may influence both bipolar disorder and neuroticism or, as in **Figure 3**, both AFB and years of EDU could be influenced by personality or risk aversion traits. Vertical or mediated pleiotropy is a genetic correlation that emerges if the genetic variants are, for instance, causal for EDU but also associated with fertility behavior due to a causal effect of education on fertility behavior. Spurious pleiotropy is when there is no direct link between a genetic variant and multiple phenotypes and also no direct link between the phenotypes. For example, it could be that SNPs that have a causal effect on both smoking and risk behavior are in close proximity but are not independently inherited (due to LD). This would result in a spurious association between the outcomes, and the genetic basis is divergent (Young et al. 2019).

From Correlation to Causation

Until now, the discussion has largely focused on correlation, the size and direction of the genetic association between variables, whereas there has been increased adoption and development of causal models (Conley & Zhang 2018). A variety of multivariate causal sociogenomic models have applied direct, reverse, bidirectional, mediation, and moderation models (for details, see Mills et al. 2020a, chapter 2).

Mendelian randomization (MR) is one of the most widely used statistical techniques in this area to examine causal effects and uses PGSs as an instrument, thereby exploiting the random assignment of an individuals' genotype at conception (Davey Smith 2005). MR is essentially an instrumental variable (IV) approach, statistically akin to a randomized control trial. The IV in MR assumes that it is associated with the exposure but not with any confounder of the exposure-outcome association and contends that there is no causal pathway from the IV to the outcome other than via the exposure.

MR is widely applied but often suffers from its inability to meet basic model assumptions. One of the main violations is the exclusion restriction assumption, which assumes that the only way in which the instrument should affect the outcome is via the exposure. Yet since PGSs are the sum of many genetic effects scattered across the genome (with widespread pleiotropy and biological pathways less often understood), it is likely that the PGS will violate this assumption. One recent solution has been MR-Egger, which tests for pleiotropic violation of the exclusion restriction assumption via an unconstrained intercept of a regression of the gene-outcome on the gene-exposure associations (Burgess & Thompson 2017). The intercept from MR-Egger regression is interpreted as the average pleiotropic effect, with a nonzero intercept indicating the presence of pleiotropy and subsequent violation of the exclusion restriction assumption.[6] Another recent extension is MR-TRYX (from TReasure Your eXceptions), which exploits horizontal pleiotropy to identify alternative causal pathways (Cho et al. 2020).

An extension of the MR framework is genetic instrumental variable regression (GIV) (DiPrete et al. 2018), which aims to control for biological pleiotropic effects. The method splits a GWAS sample into two independent subsamples and estimates two separate conditional PGSs. One PGS is then used to instrument for the other, deriving a new estimated PGS (GIV) free of measurement error that can be used in the second stage of analysis.

Applications using MR abound. A creative study was the interaction of the education PGS with the IV of veteran status from the US Vietnam War–era draft lottery (Schmitz & Conley 2017), which found that veterans with below-average PGSs had fewer years of education than comparable nonveterans. This suggested that the exogenous environmental influence of the draft lottery induced heterogeneous treatment effects by genetic disposition. Using the education PGS, other research found a causal negative effect of education on allostatic load (exposures that accumulate across the life course), BMI, and HbA1c (average blood glucose levels) (Ding et al. 2019). Another study by Davies et al. (2018b) used a regression discontinuity design for PGSs and a school reform to investigate the causal effects of staying in school for longer or shorter periods of time on health outcomes. Staying in school reduced the risk of diabetes and mortality, but there was evidence of violation of model assumptions, such as a causal effect of height on EDU, likely due to assortative mating.

A recent technique is genomic SEM, which interrogates the underlying causality of correlations when fitting genetic multivariate regression models (Grotzinger et al. 2019). It estimates a

[6]There are also other techniques, such as a weighted median method that permits valid genetic variants to contribute more weight to the analysis and avoid the estimate being influenced by noisy and invalid genetic instruments (Bowden et al. 2016).

genetic variance-covariance matrix from GWAS summary statistics for multiple traits using multivariate LD score regression. For sociologists, this type of model may seem more familiar, since it requires that theory and an understanding of causal mechanisms are specified in advance. Genomic SEM is not a method to infer causality per se, but rather a useful tool to estimate assumed causal relationships and compare how these theoretical models fit the data. The researcher specifies causal genetic pathways among multiple traits, producing estimates of the genetic correlation of *X* with *Y*, independent of *Z*. Recent applications include how the genetic relationship between psychiatric traits [e.g., attention deficit hyperactivity disorder (ADHD)] and externalizing behavior on age at first sex and AFB is potentially mediated by EDU (Mills et al. 2020b). There are also many additional useful functions of the software such as increasing the power of GWAS findings or identifying variants that are common or heterogeneous across various traits. A recent study used Genomic SEM, for instance, to estimate GWAS-by-subtraction for non-cognitive skills by residualizing SNP associations with educational attainment independent of cognitive ability (Demange et al. 2020).

GENOME-WIDE ASSOCIATION STUDIES OF SOCIOGENOMIC BEHAVIORAL OUTCOMES

How Sociologists Can Apply Genetic Discoveries in Sociological Research

The most relevant application of recent genetic discoveries for empirical sociologists entering this area of research is the application of PGSs derived from the summary statistics of large GWAS discoveries. In this section, we introduce the most recent results on selected traits relevant to sociologists, which are often embraced by neighboring disciplines but yet to be fully discovered by sociologists. Since there are thousands of traits discovered, mostly related to diseases, we cover only a fraction of the discovered behavioral outcomes.

Data and computing requirements. To undertake this research, an obvious requirement is a data set that contains both genetic data and the trait of interest. Several data archives exist such as the US-based dbGaP or METADAC in the United Kingdom. Data available for public release by certified researchers include the UK Biobank, the Health and Retirement Study (HRS), Add Health: National Longitudinal Study of Adolescent to Adult Health, the Wisconsin Longitudinal Study (WLS), and Understanding Society: The UK Household Longitudinal Study (UKHLS). A list of around 2,000 data sources from the largest GWAS to date can be found elsewhere (GitHub site linked in Mills & Rahal 2019).[7] Since genomic data are truly big data, computational demands are high and generally demand moving to a cluster computing environment and arranging considerable storage.[8]

Obtaining genome-wide association study summary statistics. All GWAS discoveries can be interactively explored in the NHGRI-EBI GWAS Catalog (**https://www.ebi.ac.uk/gwas/**), the Atlas of GWAS Summary Statistics (**http://atlas.ctglab.nl**; Watanabe et al. 2019), and a dashboard searchable by trait that focuses on GWAS ancestry and geographical diversity (**https://**

[7]A list of around 2,000 of the most used data sources for the largest one-third of GWASs between 2005 and 2018 linked with Mills & Rahal (2019) can be found here (**https://github.com/crahal/GWASReview/blob/master/tables/Manually_Curated_Cohorts.csv**).

[8]The imputed data of the UK Biobank, for instance, are around 2.1 terabytes, which then scale in a linear fashion with the number of SNPs and individuals. Although computing time is of course highly dependent on the type of analysis, for a simple association analysis, running a standard (BOLT-LMM) association on the UK Biobank takes around 100 MB of RAM and several days to run if it is given eight processors to use.

www.gwasdiversitymonitor.com; Mills & Rahal 2020). In some cases, expertly constructed PGSs are already provided with the data release (e.g., HRS, Add Health, WLS). If the constructed scores are not available, most summary statistics can be found at the GWAS Catalog website. Alternatively, a GWAS of virtually all traits in the UK Biobank has been made available online by Ben Neale's lab (**http://www.nealelab.is/uk-biobank/**). If only raw summary statistics are obtained, it is the task of the researcher to create PGSs, which requires advanced skills, particularly if adopting a Bayesian approach using LDPred (Euesden et al. 2014, Mills et al. 2020a, Vilhjálmsson et al. 2015). Researchers should also ensure that if the data set where the PGS will be applied was in the original GWAS discovery, it must be removed before calculating the PGS (for details, see Mills et al. 2020a, chapters 5, 10). The inclusion of genetic variables into statistical models also needs to be conducted with considerable care, with the introduction of additional control variables (e.g., principle components), and applied only within ancestry groups (Martin et al. 2017), with increasing acknowledgment that PGSs may also vary by age, sex, and socioeconomic status within ancestry groups (Mostafavi et al. 2020, Mills et a. 2020b) or across birth cohort and country (Tropf et al. 2017).

Educational Attainment, Cognitive Traits, and Household Income

Socioeconomic markers are core topics within sociology. As of 2020, three GWASs examined EDU, defined by years of education and typically inferred from country-specific international standard classification of education measures. The first GWAS appeared in 2013 (Rietveld et al. 2013), with a second study of around 300,000 individuals identifying 74 genetic loci (Okbay et al. 2016b), followed by a 2018 study that included 1.1 million individuals and identified around 1,300 loci (Lee et al. 2018). The combined PGSs for the latest study predicted around 11% of the total variance (incremental R^2), which neared R^2 effect estimates of classic predictors in sociology such as parents' education (19%), cognitive ability (14%), and income (6%). A series of related GWASs have been conducted on general cognitive ability (148 loci) (Davies et al. 2018a), intelligence (Savage et al. 2018), and noncognitive skills (Demange et al. 2020). Additional GWASs include social deprivation and household income measured in 112,151 individuals in the United Kingdom (Hill et al. 2016), extended in 2019 to find 149 loci for income in a larger sample (Hill et al. 2019a). Examples of studies applying these scores are discussed in the section titled From Correlation to Causation and are addressed in later sections.

Reproduction, Fertility, and Sexual Behavior

There is a history of integrating biology into the study of fertility and sexual behavior, particularly in demography (Kohler et al. 1999). Previous twin studies demonstrated that fertility was highly heritable (up to 40%) across multiple cohorts and countries (Mills & Tropf 2015). SNP heritability is 15% for AFB and 10% for NEB (Tropf et al. 2015). In 2016, the first GWAS on AFB and NEB was conducted on samples of around 250,000 (AFB) and 345,000 (NEB), identifying 10 independent loci for AFB and 2 for NEB (Barban et al. 2016).

This research was extended in 2020 by two considerably larger studies. One was of AFB (~540,000) and age at first sexual intercourse (AFS) (~389,000) (Mills et al. 2020b). The most recent study isolated almost 100 and 300 loci for AFB and AFS, respectively, and additional sex-specific loci, with an explained variance of the PGSs of 5–6%, once again edging toward the predictive power of classic demographic and sociological variables. A second extension examined NEB (~717,000) and childlessness (~450,000), identifying almost 50 new loci but uniquely linking the contemporary findings to ancient genome data, showing evidence of natural selection (Mathieson

et al. 2020). Related work includes age of onset at menarche (Perry et al. 2014), age at menopause (Stolk et al. 2012), and age at voice breaking for boys (Day et al. 2015).

Multiple applications exist, such as how a higher PGS propensity toward younger childbearing predicts teen pregnancy (Belsky et al. 2019) or is higher for girls raised in homes with an absent father (Gaydosh et al. 2018). Others show how a later AFB PGS is linked to longevity (Mills et al. 2020b, Mostafavi et al. 2017), to a longer healthy life span (Zenin et al. 2019), and with infertility and hormonal traits such as sex-steroid receptors (Maney 2017). An increase in one standard deviation of the PGS of NEB is associated with a decrease of around 9% in the probability of remaining childless for women (Mills et al. 2018). One-standard-deviation increases in the age at first sexual intercourse and birth PGSs are associated with a 7.3- and 6.3-month delay in sexual debut and age at birth, respectively (Mills et al. 2020b).

Well-Being and Psychological Traits

Psychological traits are a large area of research in this field, with only a fraction discussed here. A 2016 study of well-being that found 3 variants (Okbay et al. 2016a) was eclipsed by another with 2 million cases that found 304 signals (Baselmans et al. 2019). Other studies have examined self-reported health (Harris et al. 2016), lack of perseverance, and impulsivity (Sanchez-Roige et al. 2019). Considerable research has examined personality dimensions such as neuroticism (Baselmans et al. 2019), with some studies showing that PGSs related to lower anxiety, tension, and worry are associated with genetic variants of better self-rated health, higher intelligence, and longer life (Hill et al. 2019b). Others have found a genetic correlation between ADHD, externalizing behavior and substance use (e.g., early smoking) with earlier sexual intercourse and births (Mills et al. 2020b). Several large studies have also been published on depressive symptoms (Baselmans et al. 2019).

Risk Tolerance, Risky Behavior, and Addiction

Risk tolerance and risky behavior, often related to substance use, externalizing behavior, and behavioral disinhibition, particularly in adolescence, is another sizeable area of research. In 2019, the largest study on risk tolerance was published, which isolated 99 genetic loci (Karlsson Linnér et al. 2019), looking at related traits such as risk-taking tendency, adventurousness, automobile speeding, and number of sexual partners. In another study of more than 1.2 million individuals, 566 genetic variants were detected associated with multiple stages of tobacco use (initiation, cessation, level) and alcohol use, of which 150 locations were shared between traits (Liu et al. 2019). Interestingly, smoking variants increased the risk for many health conditions, while alcohol use showed a negative association. The genetics of smoking has been a strong area of research, since plausible biological pathways that control nicotinic, dopaminergic, and glutamatergic neurotransmission were found. This has resulted in multiple applications within sociology, often focusing on G×E (Boardman et al. 2010, 2011; Fletcher 2012). Other recent studies have found variants related to alcohol dependence and consumption (Kranzler et al. 2019, Liu et al. 2019) or links with risky behavior and early substance use with early sexual intercourse and births (Mills et al. 2020b).

Longevity- and Mortality-Related Traits

Ageing, mortality, morbidity, and healthy life spans are other substantial areas of research. Recent findings isolated 25 loci related to parental (mother's and father's) longevity, with overlaps of 8 loci tied to survival, Alzheimer's disease, and cardiovascular disease (Pilling et al. 2017). Another study found sex-specific loci related to longevity in Han Chinese, albeit with a very small sample

(Zeng et al. 2018), with a larger study finding six loci associated with parental life span (Wright et al. 2019), and another examined the human life span beyond 40 years (Joshi et al. 2016) and the health span (Zenin et al. 2019). Most GWAS discoveries have been in the area of health and disease. There are a multitude of studies of disease and causes of death that we do not report, with some, for instance, examining suicide, such as attempts and their relationship to depression (Mullins et al. 2019). The expectation is that there will be a continued expansion of the discovery of multiple traits relevant to sociologists.

INTERGENERATIONAL TRANSMISSION OF BEHAVIOR AND SOCIAL MOBILITY

A sustained area of sociogenomics research is in the area of the intergenerational transmission of behavior and social mobility. The transmission of EDU, life course outcomes, and impact of the parental environment have been studied for decades by stratification and mobility researchers in sociology (Breen & Jonsson 2005, Erikson & Goldthorpe 1993, Ganzeboom et al. 1991, Sorokin 1927). Yet, there is surprisingly little attention paid to sociological foundations in recent genetic research, with ample room for sociologists to explore and innovate this area of research.

A large study examined whether the education PGS was related to socioeconomic mobility across the life course in the United States, England and Wales, and New Zealand (Belsky et al. 2018). They found that the PGS was associated with social attainment and uncovered an rGE with those with higher PGSs growing up in higher-socioeconomic households. Education PGSs were linked to social mobility, regardless of children's origins; children with higher PGSs for EDU had better outcomes in terms of education, occupation, and wealth. This finding also held in comparison to siblings in the same family. Since genetic effects were attenuated when controlling for social origins, the association between education-related genetics and social attainment reflected the impact of parents' genetics on the creation of family environments.

This discussion relates to a recent study mentioned above and published in *Science* (Kong et al. 2018), which modeled the intergenerational transmission of education using unique multigenerational trio (mother, father, child) genetic data from Iceland and resulted in top geneticists discovering the importance of sociology.[9] With genetic and phenotypic data from both parents and offspring, the researchers were able to identify the 50% of the parental genome that had been passed on to a child as well as the 50% that had not. The authors demonstrated that not only do genes transmit information from parents to offspring but there must be another transmission channel: a social one. They found that both the inherited and noninherited genes for EDU have an impact on the child's education, with stronger effects from the mother. Since there was no biological pathway in which the noninherited genetic variants could impact children, they concluded that the environment parents create based on their nontransmitted genes has a causal impact on their EDU, which the authors termed genetic nurture. The impact of the inequality of opportunity and family background on subsequent EDU has, however, been long established within sociology (Ganzeboom et al. 1991).

Another study found that only 3% of the intergenerational correlation of EDU was explained by genetic effects (Conley et al. 2015). A more recent study using the same data, three generations, and the new educational PGSs found genetic effects of around 7% (Liu 2018). Rescaling results from PGSs of heritability from other studies to deal with missing heritability, some studies showed that around one-fifth of the intergenerational transmission of education is due to genetic

[9]In personal communication, a top geneticist and author confessed "I didn't at first want to learn about sociology, but now I realize that I have to. I have no choice."

inheritance (Conley et al. 2015, Tucker-Drob 2017). This suggests that although sociologists can no longer ignore the role of genetics, contrary to what some behavioral geneticists claim (Plomin 2018), family, school, neighborhood, and socioeconomic factors still play the strongest role. A recent study by Selzam et al. (2019) suggested that genetic nurture is entirely driven by parents' socioeconomic status, whereas others demonstrated that genetics may directly influence parenting (Wertz et al. 2019a) and that genetic nurturing effects might not stem only from the intergenerational transmission of education. Work examining parental investment has pointed to the continued importance of considering both genetic and socioenvironmental factors (Wertz et al. 2019b). One crucial implication of these findings is that genetic discovery likely overestimates genetic effects due to rGE. Potential solutions are to conduct within-family analyses in discovery and prediction, yet within-family prediction might also underestimate the true effects on PGSs (Trejo & Domingue 2019).

GENE-ENVIRONMENT INTERPLAY

Gene-environment interplay includes G×E and rGE. More research has focused on G×E, which studies whether the effect of the genotype varies across different environments, whereas rGE is the process by which an individual's genotype influences or is associated with exposure to the environment (i.e., how genes and the environment operate in tandem).

Gene-Environment Interaction

Perhaps one of the most compelling areas of research for sociologists has been the promise of G×E (Conley 2016, Freese & Shostak 2009). Four main theoretical models have been used to understand G×E: the diathesis–stress, bioecological (social compensation), differential susceptibility, and social control models. Recognition of how the social environment (social structure) enables and constrains agency forms the basis of most sociological thinking.

The diathesis–stress (also known as vulnerability, contextual triggering) model assumes that a predisposition for a trait lies dormant until triggered by an environmental stressor (Monroe & Simons 1991). Often inspired by the now-controversial Caspi (2002) study, many have applied this approach. In testing both diathesis–stress and the protective influence of genes, a higher PGS for well-being buffered against increased depressive symptoms following the death of a spouse (Domingue et al. 2017b). Another study validated this theory, showing how a higher PGS for major depressive disorder became relevant only after a series of stressful life events (Arnau-Soler et al. 2019).

The bioecological (or social compensation) model supposes that genetic influences are maximized in stable, adaptive, and often higher-socioeconomic environments, enabling individuals to reach their genetic potential (Bronfenbrenner & Ceci 1994). Seminal work by Turkheimer et al. (2003) showed G×E by socioeconomic status, demonstrating that the heritability of IQ is the highest in families with high socioeconomic status and lowest in families with low socioeconomic status. This work tested the classic Scarr-Rowe hypothesis (Scarr-Salapatek 1971) that families with high resources can help their children realize their genetic potential and echoes our above discussion of intergenerational social mobility. Others demonstrated that students with higher education PGSs were tracked to more advanced math courses and remained there longer, with advantaged schools tracking higher-PGS students into advanced courses and protecting those with lower PGSs from dropping out (Harden et al. 2019).

Growing research has demonstrated G×E by how cross-national and social structural differences impact heritability and PGS prediction (Belsky et al. 2018, Engzell & Tropf 2019, Rimfeld

et al. 2018, Tropf et al. 2017). A meta-analysis (Tucker-Drob & Bates 2016) showed that the variability of IQ by socioeconomic status held for the United States but did not extend to Europe or Australia, likely due to a safety net for disadvantaged households. A study in Florida, however, found no evidence that socioeconomic status moderated the genetic influence of test scores (Figlio et al. 2017). Rimfeld et al. (2018) analyzed the interaction between education PGSs and political system based on the Soviet and post-Soviet era in Estonia, finding that the independence of Estonia was related to a higher predictive power for the PGSs for both education and occupational status. The theoretical reasoning was that the transition from a communist to a capitalist society under meritocratic principles enabled one's genetic potential to unfold. Another study meta-analyzing previous research from 10 Western countries showed that populations and birth cohorts with higher social mobility shared higher heritability of EDU and lower environmental influences (Engzell & Tropf 2019).

Whereas the diathesis–stress model focuses almost exclusively on negative environmental influences, the differential susceptibility model argues that plasticity varies by individual, with some individuals more genetically susceptible (orchids) to both positive and negative environments, while others remain resilient across all situations (dandelions) (Belsky & Pluess 2009). This has been applied, for instance, to examine the orchid effect of how physical health differs with respect to marital relationship quality (South & Krueger 2013).

The social control or social push model supposes that genetic influences are filtered or buffered in particular environments such as by social norms or structural constraints (e.g., taxes, banning smoking in public places) (Shanahan & Hofer 2005). A study by Guo et al. (2015) applied social control and learning theory, utilizing the random assignment of college students as a natural experiment to estimate the social and genetic influence of drinking behavior. They found that room assignment to a peer who engaged in precollege drinking increased binge drinking for individuals with a moderate genetic predisposition for drinking. Other applications include how having the social control of a romantic partner limits alcohol misuse (Barr et al. 2019) or how the heritability of smoking was significantly reduced in US states that introduced restrictive policies on the sale of cigarettes and higher taxes (Boardman 2009).

Wedow et al. (2018) theoretically extended the typical gene-environment paradigm to a multidimensional framework, analyzing how environmental changes jointly modify genetic effects on multiple outcomes. They found that throughout the past century, the genetic correlation between EDU and smoking behavior increased. Potential explanations for this phenomenon of so-called genetic correlation–by-environment interaction (rG×E) included shared genetic influences on education and smoking behavior, for example, based on risk behavior.

Gene-Environment Correlation

A longstanding area of research has been how an individual's genotype influences or is associated with exposure to the environment. The three main rGE processes are passive, evocative, and active rGE (Plomin et al. 1977). Passive rGE is the association between the genotype a child inherits from parents and the environment where she or he is raised. Assortative mating of highly educated parents, for instance, results in genetic transmission of cognitive ability but also creates an environment of learning, monitoring, and higher educational expectations (Kong et al. 2018). In a test of the PGSs of education and parental caregiving, PGSs predicted warm, sensitive, and stimulating caregiving and positive home environments, with scores mediated by parent's cognitive abilities and self-control skills (Wertz et al. 2019a).

Evocative (or reactive) rGE is when an individual's heritable traits evoke reactions from others in the environment. If someone, for instance, is naturally prone to being introverted or shy, they

may appear unapproachable or aloof, which may in turn reinforce that trait through reactions to them. Active rGE (or niche creation) is when individuals actively select or create environments associated with their own genetic predispositions and could include risk-taking or, conversely, seeking highly regulated environments. This links with research on social genetic effects or the effects of peers' genotypes on ego's social outcomes. A provocative claim is that friends assort based on their genome, which in turn directly influences the social outcomes of an individual (Domingue & Belsky 2017, Domingue et al. 2018). This work has been criticized, however, by those claiming that the genome-wide relationship between friends is extremely small and explained by subtle population stratification, with the significant individual-friend educational PGS association expected under homophily, even without social genetic effects (Yengo et al. 2019).

Challenges and Solutions for Gene-Environment Interplay Research

The use of PGSs for the detection of G×E remains challenging because they might be noisy and underpowered, which has led to several unexpected null findings in the literature (Trejo et al. 2018). There are several persistent challenges but also potential solutions (for a summary, see Mills et al. 2020a).

An initial problem is that the theoretical models underlying G×E described previously are generally not formalized and often overlap, with an inability to articulate plausible biological pathways for G×E associations, which in turn leads to a black box explanation. A second problem is considerable noise in models due to an inconsistent measurement of the environmental exposure or trait at the stage of either the initial GWAS or subsequent replication attempts. When including variables that need to represent exogenous environmental shocks, researchers need to carefully consider their actual exogeneity (Conley 2017). There are also often a multitude of environmental influences, which is rarely acknowledged in simple one-event, exogenous-shock frameworks. Recently, structured linear mixed models were developed to identify and characterize genetic loci that are more likely to interact with one or more environments, allowing the examination of interactions across hundreds of environmental variables (Moore et al. 2019). Third, when the environment is viewed as a proximate environmental moderator (e.g., maternal smoking), it is often considerably downstream from the social environmental factors that structure the actual exposure (Boardman et al. 2013). A fourth issue is that there is often too much attention placed on the individual environment, negating environmental influences such as normative, legislative, religious, and cultural influences, or even that of the natural environment (e.g., pollution), and other factors that likewise shape behavior.

A fifth, more glaring issue is the lack of statistical power and large samples that are required to detect effects in G×E models, with power calculations rarely conducted. A sixth concern is selection and the lack of environmental diversity in many genetic and medical data sets (Mills & Rahal 2019, 2020), elaborated upon below. A seventh issue is that the interaction is sensitive to the scaling of the metric, necessitating consideration of both multiplicative and additive scales (Mills et al. 2020a). Finally, detected interactions may actually be driven by confounders (e.g., age, socioeconomic status) rather than by genetic or environmental variables when researchers incorrectly enter confounders as covariates in general linear models—a problem that is easily rectified (Keller 2014).

rGEs are exceedingly difficult to study, since the relationships between environmental exposure and the outcome are often the result of a complex and reciprocal causal process. Passive rGE, for example, may measure a spurious relationship between the environment and a trait. Innovations include quasi-experimental designs of statistical matching of sibling pairs to rule out evocative rGEs related to men's marital status and antisocial behavior (Jaffee et al. 2013). One cleverly

designed study tested a passive rGE by examining parents whose children were conceived via assisted reproductive technologies. The parents were biologically related to the child but also used external sperm, egg, or embryo donation, resulting in a split sample of some children related to both parents, some the mother only, and some the father only (Rice et al. 2013). This allowed the authors to examine depression in parents and children and evaluate how it was related to parental behavior and positivity and how this varied by children's genetic relatedness to their parents.

DIVERSITY AND NONREPRESENTATIVE SAMPLES

Lack of Diversity Exacerbates Inequality and Threatens Applicability

Although genetic research is often introduced in optimistic terms, a serious issue undermining genetic research is the lack of ancestral, geographic, and demographic diversity in the data that are currently used. The term ancestry that is used in human genetics is not interchangeable with the terms race or ethnicity used within sociology. Race is a socially and politically derived construct and not a biological category, which has been discussed extensively elsewhere (Conley 2016, Mills et al. 2020a, Nelson 2016). Ancestry in genetics mirrors geography, migration patterns, and population stratification by continents, countries, or regions (Conley 2016; Mills et al. 2020a, chapter 3; Mills & Rahal 2020). Using broad terms adopted by geneticists, Mills & Rahal (2019) found that 212 unique terms were used to classify individuals in relation to race, region, country, ethnicity, and ancestry.[10]

All GWASs (2005–2018) are overwhelmingly conducted on individuals of European ancestry, ranging from between 88% in 2017 to 95% in 2007 (Mills & Rahal 2019), and quasi-daily real-time monitoring of diversity is now in place from funders, governments, journal editors, and researchers to push for accountability (Mills & Rahal 2020). Contrary to two-point time comparisons that suggested increased diversity (Popejoy & Fullerton 2016), diversity has in fact decreased over time (Mills & Rahal 2019, 2020). Although studies of Asian ancestry groups increased, genetic discoveries using particularly African, African-American, or Afro-Caribbean and Hispanic and Latin American ancestry groups remain astoundingly scant (Mills & Rahal 2020) (**Figure 2**). At the initial discovery phase, all GWASs of cancer research, for instance, in 2019, consisted of 96.3% individuals of European ancestry compared with 0.1% of African, 0% of African-American and Afro-Caribbean, and 0.5% Hispanic or Latin American ancestries (Mills & Rahal 2020). A recent study of people from rural Uganda, for instance, showed that a genetic variant derived from European ancestry groups that is commonly used to diagnose diabetes would result in the incorrect diagnosis of diabetes in the Ugandan population (Gurdasani et al. 2019). Furthermore, 72% of all genetic discoveries emanated from using data from people from just three countries (United States, United Kingdom, and Iceland), which is striking since results for complex traits vary by country of origin (Tropf et al. 2017). Data from Iceland, which has a population of around 300,000, were responsible for around 12% of all genetic discoveries (Mills & Rahal 2019, table 2).

[10]The coding and data are openly available for other researchers to download, replicate, refine, or use (**https://github.com/crahal/GWASReview**) in addition to an extension in Mills & Rahal (2020). The approach and results from Mills & Rahal (2019) differ from a previous sociological study that mapped ancestry labels against the US Census racial classification system (Panofsky & Bliss 2017). The Panofsky & Bliss (2017) study included only selected early studies from one journal, and it was not explained or justified as to why the ancestry of all studies (including non-US studies) was mapped to the context-specific US racial classification system (but it was possibly intentional ambiguity or scientific authority).

Problems and Solutions for Nonrepresentative Samples

Similar to other disciplines such as psychology that base their research on WEIRD (western, educated, industrialized, rich, and democratic) samples (Henrich et al. 2010), our current genetic knowledge from GWASs not only misses information from around roughly 76% of the world's population but is highly nonrepresentative (Mills & Rahal 2019, 2020). Nonrepresentative samples in genetics translate into an overrepresentation of individuals in the data who are often older, healthier, and female (Mills & Rahal 2019). Sex differences have been largely unexplored, yet there are known genetically related sex differences, particularly for complex traits (Khramtsova et al. 2019). Research in the area of reproduction and sexual behavior, for instance, has found different sex-specific loci and divergent downstream biological affects by sex (Barban et al. 2016, Ganna et al. 2019, Mills et al. 2020b). Others have shown how the prediction accuracy of PGSs depends on the age or sex composition of the GWAS discovery sample and study design (Mostafavi et al. 2020).

Sociologists and demographers place considerable importance on representativeness, generalizability, external validity, and examining heterogeneity of effects across contexts and groups. Conversely, geneticists and epidemiologists are generally more invested in internal validity and the notion of universal associations and causal effect, thereby negating the need for representative samples (Elwood 2013). Some argue that representativeness should even be avoided, since, given universal causal effects, external validity emanates from the understanding of causal mechanisms, also in the absence of statistical generalizability (Rothman et al. 2013).

Many recent genetic discoveries are now drawn from the 500,000 samples of the UK Biobank, which has a 5.5% response rate and is overrepresented by healthy, low-BMI, nonsmoking, high-socioeconomic individuals (Fry et al. 2017). Many argue that, although prevalence and incidence should be judged with caution in these nonrepresentative samples, measures of association and estimates are valid since they do not require representative populations. These samples have become the cornerstones for genetic discovery and drug development.

We contend that the myth of not requiring a representative sample is highly problematic for many genetic analyses. While some argue that exposure-disease correlations are still generalizable (Fry et al. 2017), we now know that genetic associations are modifiable (de Vlaming et al. 2017, Wedow et al. 2018), particularly for behavioral and lifestyle outcomes (Tropf et al. 2017). These modifiers bias the GWAS estimates toward those that are mainly true for that overrepresented group.

Second, selection also impacts the predicted results in the target (i.e., nondiscovery) data. Sample selection clearly has the potential to influence measures of association. Whether or not an association observed in one study is similar in another target population (i.e., has external validity) is dependent upon the distribution of the exposure-outcome relationship in the discovery and target populations. Using a simulated empirical example, a recent study examined the transferability of the association with a disease that differed across an original healthy volunteer sample, which was then applied to a target population where the disease prevalence differed (Keyes & Westreich 2019). Countering the claim that representativity of the sample did not matter, the magnitude of the association was in fact highly dependent on the prevalence of other factors that interact with the exposure and outcome. They showed that as the prevalence of the trait decreased, the magnitude of the association also decreased.

We contend that this serious issue could be solved by developing weighting methods to adjust for results in the study sample to match results in the target population, such as for mortality selection in the HRS (Domingue et al. 2017a), a common procedure used with sociological and demographic data. Other solutions could involve modeling known moderators or selection

models such as a Heckman selection model (Heckman 1979). In the end, however, collection of more representative and diverse data and diversification of the reference groups and tools used in these analyses would be a longer-term solution.

GENETIC SOCIAL POLICY AND ETHICS

Can Genetics Be Used to Develop Precision Policy Measures?

Some sociologists staunchly oppose the use of genetics, particularly in relation to the study of EDU, arguing that it is the backdoor to eugenics (Bliss 2018). Although multiple reviews have pointed out large factual errors and overstated claims in the aforementioned work (Freese 2018), this and other carefully researched work (Duster 2006, Nelson 2016, Reardon 2017) still raises vital questions about the use and understanding of genetic results. Prominent behavioral geneticist Robert Plomin (Plomin 2018, Plomin & von Stumm 2018), for instance, argued that the PGSs for EDU could be used akin to precision medicine to develop precision education tools. It may be that PGSs might eventually be useful to determine early dyslexia or other traits that are often diagnosed much later.

Beyond obvious ethical concerns, however, there remain core technical barriers to applying current intelligence, cognitive ability, and EDU genetic scores for educational policy. First, given the fact that these genetic scores largely emanate from European ancestry groups in healthy and wealthy populations—and we know that PGSs derived from one ancestry group cannot be reliably applied to another group—these policies would exacerbate existing inequalities (Martin et al. 2017, 2019). Genetic policy applications thus have the potential to enhance already advantaged groups and even increase socioeconomic disparities (Selzam et al. 2019). Second, even if diversity in samples increases, there remains an implicit bias against non-European ancestry groups in the technical tools that are used such as the limited reference panels used to generate PGSs for non-European ancestry groups (M.S. Kim et al. 2018). Third, the ubiquity of pleiotropy means that the biologically causal function of genes is often a black box, making genetically-based interventions ineffective and risking unintended consequences. As noted earlier, the genetic correlation between EDU and AFB is 0.74, signaling an extremely high genetic overlap between these two traits. EDU genetically-based policies could thus also inadvertently influence the timing of fertility. Fourth, there is often poor measurement and harmonization of the phenotype or outcomes of studies, making them imprecise. It is particularly in this area of research where sociologists are masters of measurement in education, family, other lifestyle indicators, and the social structural environment where clear contributions can be made. A fifth point is that the PGSs for complex polygenic traits are simply not precision measurements. Although proponents argue that we could use, for instance, the upper percentile PGSs for precision education policy (Plomin & von Stumm 2018), if we look at a scatterplot from their own article, someone with a PGS in the 98th percentile could score anywhere between the 2nd and 98th percentile. Put differently, if someone asked you to board a plane within the 2nd to 98th percentile range of crashing, we would hope that you would not get on. Finally, in contrast to medical research, where it might be necessary to predict a condition or disease before its manifestation, no PGS will predict educational success better than measured educational success. A recent study showed no additional explanatory power of the EDU PGS on top of previous achievement (Morris et al. 2019).

Genetic Determinism, Identity, Public Understanding, and Privacy Risks

A final emerging area of research is the infiltration of genetic data into everyday lives, identity politics, public understanding, and risks of privacy and identification. With the growth of DTC

genetics companies, millions of individuals are being genotyped and interpreting results without a genetic counselor (Khan & Mittelman 2018). Nelson (2016, p. 69) recently elaborated that the rise in genetic genealogy testing "aligns with an enduring human desire: the search for roots and identity." There is not only a drive to understand oneself but also a tendency for individuals to attribute more scientific validity to genetic results, even in light of small predictions with mixed results on whether receiving results actually alters health behavior (McBride et al. 2010).

The new freedom of discovering one's genome, however, is accompanied with risks, described in detail elsewhere (see Mills et al. 2020a, chapter 14). Personal information risks include unmasking adoptions, infidelity, or nondisclosed family secrets. The commercialization of genetic tests means that DTC companies often pair horoscope-like predictions for psychological or lifestyle characteristics with more solid predictors related to physical traits such as hair color, where genetics is more straightforward, making it difficult for a nonexpert to distinguish reality.

A growing risk is privacy and identification. A 2018 study analyzing 1.28 million anonymous individuals using MyHeritage data found that 60% of Americans of European ancestry can be identified via one of their relatives (Erlich et al. 2018). Public genetic databases like GEDmatch and DTCs have been used to solve rape and murder cold cases (Mills et al. 2020a), with a recent study showing that 30% of people in forensic databases can be linked to relatives in consumer databases (J. Kim et al. 2018). Since these databases are overrepresented by individuals of European ancestry, their use for forensic purposes may in fact counter previous inequalities of incarcerated minority groups. There are, likewise, continued concerns about using genetic results for embryo selection or designer babies, with a recent study showing that using polygenic scores to screen for IQ or height would have limited utility (Karavani et al. 2019). Some also fear that the results could be used for or against so-called cures or interventions for behaviors such as same-sex behavior (Mills 2019). Another concern is genetic discrimination by an employer or insurance company. International coverage is mixed, with individuals in the United States protected by the 2008 Genetic Information Non-Discrimination Act, yet not all groups (e.g., employers with fewer than 15 employees, US military, disability) receive protection (Hudson et al. 2008).

A final worry is that right-wing groups and pseudoscientists misuse genetic data to propagate racism and misinformation. This hidden racism is often based in hereditarian writing, with groups presenting their work as neutrally scientific, full of promise for a new society, and adopting a supposedly empirical and colorblind façade (Gillborn 2016). This misuse has taken place in cases such as a flawed analysis applying educational PGSs among 53 Jewish respondents in a data set, with sociologists quick to invalidate on technical grounds (Freese et al. 2019). Rather surprisingly (as of early 2020), the Editor of the article in question decided not to retract it from the American Psychological Association journal where it was published, even after direct contact was made to clearly signal the analytical mistakes (correspondence with M.C. Mills available upon request). It remains a difficult balance of whether to ignore weak and incorrect research or give it unwarranted oxygen.

CONCLUSION

We have highlighted astonishing developments within sociogenomics, particularly in the last decade, in addition to what we hope has been a balanced and diverse review of the literature and data. The advent of molecular genetic data has offered unprecedented possibilities and the potential to expand many sociological theories and empirical and statistical applications. Instead of remaining confined in twin-based behavior genetic models, limited to often calculating the level of heritability of a trait, for the first time we can move toward understanding how genetic predispositions are related to outcomes and how they interact across various environmental circumstances. We have moved beyond the nature versus nurture debate and can for the first time

look at the interaction between nature and nurture. This will open exciting new avenues of research into age-old topics in sociology such as assortative mating, EDU, social mobility, fertility and reproduction, health, intergenerational transmission of behavior, risky behavior, substance use, externalizing behavior, and beyond.

What would happen if sociologists stood behind Durkheim and ignored sociogenomics and the integration of genetics into their research? For some topics that are biologically distant and with little or no genetic component, ignoring genetics will likely make no difference, and it is likely a waste to pursue. Yet for other topics that have a plausible biological link, such as fertility and reproductive behavior, for instance, where PGSs alone now explain 5–6% of the overall variance and have both a social and a biological component, moving forward without considering both may be a mistake. Or, can the latest PGS for EDU that predicts around 10% of the overall variance be ignored? This new field has the potential to change the type of social science data that are collected and also how sociologists conduct their research. The ability to include PGSs in our sociological toolkit is momentous. Theorists may also need to rethink many theoretical models that have a high reliance on individual agency and choice and ignore innate genetic predispositions and their interaction and correlation with the social environment.

Yet as our critical review reveals, we still need to understand both what the PGSs are measuring and the nature of genetic correlations and pleiotropy. We furthermore highlight the current lack of diversity in data collection and the absence of representativeness in the majority of contemporary data sources. Likewise, little attention is paid to the variation of within-ancestry PGSs (e.g., by sex, age, historical period, birth cohort, country, socioeconomic status) and the imprecision of measurement of some outcomes and the environment (Mostafavi et al. 2020). This means that sociologists can no longer ignore genetics and also that genetics needs sociological thinking.

Given the speed at which genetics and sociogenomics have developed, we anticipate an exciting future. For sociologists, prospects include the integration of more diverse data, particularly from African, African-American, and Hispanic and Latin American initiatives, and also data with richer socioeconomic and geographical backgrounds and that are longitudinal to examine changes across the life course. Other frontiers include projects able to measure more of the genome, deep and precision phenotypes (traits), and environment to better detect G×E. This includes initiatives to link sociogenomic data to administrative and health records and devices (e.g., accelerometer). For this area of research to truly flourish, however, students and researchers must not only acquire skills that are disciplinary-specific to sociology but also look beyond borders to develop genetic, biological, statistical, and programming skills. These are increasingly available in specialized courses, summer schools, and textbooks that actively engage social scientists (Mills et al. 2020a). Sociologists can opt to be active and equal participants in this field, codeveloping new theories, methods, data, and findings, or choose to stand by and watch as other disciplines study their core topics, often lacking the insight of decades of sociological research.

DISCLOSURE STATEMENT

The authors are not aware of any biases, affiliations, memberships, funding or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

M.C.M. is grateful for support from European Research Council grants 615603 (SOCIOGENOME, **https://www.sociogenome.org**) and 835079 (CHRONO); NCRM/ESRC SOCGEN (ES/N011856/1); The Leverhulme Trust; The Leverhulme Centre for Demographic

Science; and Nuffield College, University of Oxford. F.C.T. is grateful for support from the Laboratory of Excellence in Economics and Decision Sciences (LabEx ECODEC), funded under the French National Research Agency (ANR) Investissements d'Avenir (grant ANR-11-LABX-0047). We gratefully acknowledge the Editor and reviewers for their comments; Kayla Schulte (for polishing **Figure 1**); Charles Rahal (for updating **Figure 2**); David Brazel (for providing code for **Figure 3** and comments); Xuejie Ding, Riley Taiji, Evelina Akimova, and Domante Grendaite for their comments; and trailblazing researchers in this area, whose work we were honored to read and review.

LITERATURE CITED

Abdellaoui A, Hugh-Jones D, Yengo L, Kemper KE, Nivard MG, et al. 2019. Genetic correlates of social stratification in Great Britain. *Nat. Hum. Behav.* 3:1332–42

Adams PM, Albert MS, Albin RL, Apostolova LG, Arnold SE, et al. 2016. Assessment of the genetic variance of late-onset Alzheimer's disease. *Neurobiol. Aging* 41:200.e13–e20

Arnau-Soler A, Adams MJ, Clarke T-K, MacIntyre DJ, Milburn K, et al. 2019. A validation of the diathesis-stress model for depression in Generation Scotland. *Transl. Psychiatry* 9:25

Barban N, Jansen R, De Vlaming R, Vaez A, Mandemakers JJ, et al. 2016. Genome-wide analysis identifies 12 loci influencing human reproductive behavior. *Nat. Genet.* 48(12):1462–72

Barr PB, Kuo SI, Aliev F, Latvala A, Viken R, et al. 2019. Polygenic risk for alcohol misuse is moderated by romantic partnerships. *Addiction* 114(10):1753–62

Baselmans BML, Jansen R, Ip HF, van Dongen J, Abdellaoui A, et al. 2019. Multivariate genome-wide analyses of the well-being spectrum. *Nat. Genet.* 51(3):445–51

Belsky DW, Caspi A, Arseneault L, Corcoran DL, Domingue BW, et al. 2019. Genetics and the geography of health, behaviour and attainment. *Nat. Hum. Behav.* 3(6):576–86

Belsky DW, Domingue BW, Wedow R, Arseneault L, Boardman JD, et al. 2018. Genetic analysis of social-class mobility in five longitudinal studies. *PNAS* 115(31):E7275–84

Belsky J, Pluess M. 2009. Beyond diathesis stress: differential susceptibility to environmental influences. *Psychol. Bull.* 135:885–908

Bliss C. 2018. *Social by Nature: The Promise and Peril of Sociogenomics*. Stanford, CA: Stanford Univ. Press

Boardman JD. 2009. State-level moderation of genetic tendencies to smoke. *Am. J. Public Health* 99(3):480–86

Boardman JD, Blalock CL, Pampel FC. 2010. Trends in the genetic influences on smoking. *J. Health Soc. Behav.* 51(1):108–23

Boardman JD, Blalock CL, Pampel FC, Hatemi PK, Heath AC, Eaves LJ. 2011. Population composition, public policy, and the genetics of smoking. *Demography* 48(4):1517–33

Boardman JD, Daw J, Freese J. 2013. Defining the environment in gene-environment research: lessons from social epidemiology. *Am. J. Public Health* 103(10):64–72

Bowden J, Davey Smith G, Haycock PC, Burgess S. 2016. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genet. Epidemiol.* 40(4):304–14

Boyle EA, Li YI, Pritchard JK. 2017. An expanded view of complex traits: from polygenic to omnigenic. *Cell* 169(7):1177–86

Breen R, Jonsson JO. 2005. Inequality of opportunity in comparative perspective: recent research on educational attainment and social mobility. *Annu. Rev. Sociol.* 31:223–43

Bronfenbrenner U, Ceci SJ. 1994. Nature-nurture reconceptualized in developmental perspective: a bioecological model. *Psychol. Rev.* 101:568–86

Burgess S, Thompson SG. 2017. Interpreting findings from Mendelian randomization using the MR-Egger method. *Eur. J. Epidemiol.* 32(5):377–89

Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, et al. 2018. The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562(7726):203–9

Caspi A. 2002. Role of genotype in the cycle of violence in maltreated children. *Science* 297(5582):851–54

Chabris CF, Lee JJ, Benjamin DJ, Beauchamp JP, Glaeser EL, et al. 2013. Why it is hard to find genes associated with social science traits: theoretical and empirical considerations. *Am. J. Public Health* 103(S1):S152–66

Cho Y, Haycock PC, Sanderson E, Gaunt TR, Zheng J, et al. 2020. Exploiting horizontal pleiotropy to search for causal pathways within a Mendelian randomization framework. *Nat. Comm.* 11:1010

Conley D. 2016. Socio-genomic research using genome-wide molecular data. *Annu. Rev. Sociol.* 42:275–99

Conley D. 2017. The challenges of GxE: commentary on "Genetic endowments, parental resources and adult health: evidence from the Young Finns Study." *Soc. Sci. Med.* 188:201–3

Conley D, Domingue BW, Cesarini D, Dawes C, Rietveld CA, Boardman JD. 2015. Is the effect of parental education on offspring biased or moderated by genotype? *Sociol. Sci.* 2:82–105

Conley D, Fletcher J. 2017. *Genome Factor. What the Social Genomics Revolution Reveals About Ourselves, Our History and the Future.* Princeton, NJ: Princeton Univ. Press

Conley D, Rauscher E, Dawes C, Magnusson PKE, Siegal ML. 2013. Heritability and the equal environments assumption: evidence from multiple samples of misclassified twins. *Behav. Genet.* 43(5):415–26

Conley D, Zhang S. 2018. The promise of genes for understanding cause and effect. *PNAS* 115(22):5626–28

Davey Smith G. 2005. What can Mendelian randomisation tell us about modifiable behavioural and environmental exposures? *BMJ* 330(7499):1076–79

Davies G, Lam M, Harris SE, Trampush JW, Luciano M, et al. 2018a. Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function. *Nat. Commun.* 9(1):2098

Davies NM, Dickson M, Smith GD, Van Den Berg GJ, Windmeijer F. 2018b. The causal effects of education on health outcomes in the UK Biobank. *Nat. Hum. Behav.* 2:117–25

Day FR, Bulik-Sullivan B, Hinds DA, Finucane HK, Murabito JM, et al. 2015. Shared genetic aetiology of puberty timing between sexes and with health-related outcomes. *Nat. Commun.* 6(1):8842

de Vlaming R, Okbay A, Rietveld CA, Johannesson M, Magnusson PKE, et al. 2017. Meta-GWAS Accuracy and Power (MetaGAP) calculator shows that hiding heritability is partially due to imperfect genetic correlations across studies. *PLOS Genet.* 13(1):e1006495

Demange PA, Malanchini M, Mallard TT, Biroli P, Cox SR, et al. 2020. Investigating the genetic architecture of non-cognitive skills using GWAS-by-subtraction. bioRxiv 905794. **https://doi.org/10.1101/2020.01.14.905794**

Ding X, Barban N, Mills MC. 2019. Educational attainment and allostatic load in later life: evidence using genetic markers. *Prev. Med.* 129:105866

DiPrete TA, Burik CAP, Koellinger PD. 2018. Genetic instrumental variable regression: explaining socioeconomic and health outcomes in nonexperimental data. *PNAS* 115(22):E4970–79

Domingue BW, Belsky DW. 2017. The social genome: current findings and implications for the study of human genetics. *PLOS Genet.* 13(3):e1006615

Domingue BW, Belsky DW, Fletcher JM, Conley D, Boardman JD, Harris KM. 2018. The social genome of friends and schoolmates in the National Longitudinal Study of Adolescent to Adult Health. *PNAS* 115(4):702–7

Domingue BW, Belsky DW, Harrati A, Conley D, Weir DR, Boardman JD. 2017a. Mortality selection in a genetic sample and implications for association studies. *Int. J. Epidemiol.* 46(4):1285–94

Domingue BW, Conley D, Fletcher J, Boardman JD. 2016. Cohort effects in the genetic influence on smoking. *Behav. Genet.* 46(1):31–42

Domingue BW, Fletcher J, Conley D, Boardman JD. 2014. Genetic and educational assortative mating among US adults. *PNAS* 111(22):7996–8000

Domingue BW, Liu H, Okbay A, Belsky DW. 2017b. Genetic heterogeneity in depressive symptoms following the death of a spouse: polygenic score analysis of the U.S. Health and Retirement Study. *Am. J. Psychiatry* 174(10):963–70

Dudbridge F. 2013. Power and predictive accuracy of polygenic risk scores. *PLOS Genet.* 9(3):e1003348

Duncan LE, Pollastri AR, Smoller JW. 2014. Mind the gap: why many geneticists and psychological scientists have discrepant views about gene-environment interaction (G×E) research. *Am. Psychol.* 69(3):249–68

Durkheim E. 1938 (1895). *The Rules of Sociological Method.* Chicago: Univ. Chicago Press

Duster T. 2006. *Backdoor to Eugenics.* New York: Routledge

Eckland BC. 1967. Genetics and sociology: a reconsideration. *Am. Sociol. Rev.* 32(3):173–94
Elwood JM. 2013. Commentary: on representativeness. *Int. J. Epidemiol.* 42(4):1014–15
Engzell P, Tropf FC. 2019. Heritability of education rises with intergenerational mobility. *PNAS* 116(51):25386–88
Erikson R, Goldthorpe JH. 1993. *The Constant Flux: Study of Class Mobility in Industrial Societies*. Oxford, UK: Clarendon Press
Erlich Y, Shor T, Pe'er I, Carmi S. 2018. Identity inference of genomic data using long-range familial searches. *Science* 362(6415):690–94
Euesden J, Lewis CM, O'Reilly PF. 2014. PRSice: Polygenic Risk Score software. *Bioinformatics* 31(9):1466–68
Figlio DN, Freese J, Karbownik K, Roth J. 2017. Socioeconomic status and genetic influences on cognitive development. *PNAS* 114(51):13441–46
Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, et al. 2015. Partitioning heritability by functional category using GWAS summary statistics. *Nat. Genet.* 47:1228–35
Fletcher JM. 2012. Why have tobacco control policies stalled? Using genetic moderation to examine policy impacts. *PLOS ONE* 7(12):e50576
Freese J. 2008. Genetics and the social science explanation of individual outcomes. *Am. J. Sociol.* 114(S1):S1–35
Freese J. 2018. The arrival of social science genomics. *Contemp. Sociol.* 47(5):524–36
Freese J, Domingue B, Trejo S, Sicinski K, Herd P. 2019. Problems with a causal interpretation of polygenic score differences between Jewish and non-Jewish respondents in the Wisconsin longitudinal study. SocArXiv. **https://doi.org/10.31235/osf.io/eh9tq**
Freese J, Shostak S. 2009. Genetics and social inquiry. *Annu. Rev. Sociol.* 35:107–28
Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, et al. 2017. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *Am. J. Epidemiol.* 186(9):1026–34
Galton F. 1869. *Hereditary Genius: An Inquiry into Its Laws and Consequences*. New York: Macmillan
Ganna A, Verweij KJH, Nivard MG, Maier R, Wedow R, et al. 2019. Large-scale GWAS reveals insights into the genetic architecture of same-sex sexual behavior. *Science* 365(6456):eaat7693
Ganzeboom HBG, Treiman DJ, Ultee WC. 1991. Comparative intergenerational stratification research: three generations and beyond. *Annu. Rev. Sociol.* 17:277–302
Gaydosh L, Belsky DW, Domingue BW, Boardman JD, Harris KM. 2018. Father absence and accelerated reproductive development in non-Hispanic white women in the United States. *Demography* 55(4):1245–67
Gillborn D. 2016. Softly, softly: genetics, intelligence and the hidden racism of the new geneism. *J. Educ. Policy* 31(4):365–88
Grotzinger AD, Rhemtulla M, de Vlaming R, Ritchie SJ, Mallard TT, et al. 2019. Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nat. Hum. Behav.* 3(5):513–25
Guo G. 2006. The linking of sociology and biology. *Soc. Forces* 85(1):145–49
Guo G, Li Y, Wang H, Cai T, Duncan GJ. 2015. Peer influence, genetic propensity, and binge drinking: a natural experiment and a replication. *Am. J. Sociol.* 121(3):914–54
Guo G, Tong Y, Cai T. 2008. Gene by social-context interactions for number of sexual partners among white male youths: genetics-informed sociology. *AJS* 114(Suppl.):S36–66
Gurdasani D, Carstensen T, Fatumo S, Chen G, Franklin CS, et al. 2019. Uganda genome resource enables insights into population history and genomic discovery in Africa. *Cell* 179(4):984–1002.e36
Hamer D, Sirota L. 2000. Beware the chopsticks gene. *Mol. Psychiatry* 5(1):11–13
Harden KP, Domingue BW, Belsky DW, Boardman JD, Crosnoe R, et al. 2019. Genetic associations with mathematics tracking and persistence in secondary school. *Sci. Learn.* 5:1
Harris SE, Hagenaars SP, Davies G, Hill WD, Liewald DCM, et al. 2016. Molecular genetic contributions to self-rated health. *Int. J. Epidemiol.* 46(3):994–1009
Heckman JJ. 1979. Sample selection bias as a specification error. *Econometrica* 47(1):153–61
Henrich J, Heine SJ, Norenzayan A. 2010. Most people are not WEIRD. *Nature* 466:29

Hewitt J. 2012. Editorial policy on candidate gene association and candidate gene-by-environment interaction studies of complex traits. *Behav. Genet.* 41(1):1–2

Hill WD, Davies NM, Ritchie SJ, Skene NG, Bryois J, et al. 2019a. Genome-wide analysis identifies molecular systems and 149 genetic loci associated with income. *Nat. Commun.* 10(1):5741

Hill WD, Hagenaars SP, Marioni RE, Harris SE, Liewald DCM, et al. 2016. Molecular genetic contributions to social deprivation and household income in UK biobank. *Curr. Biol.* 26(22):3083–89

Hill WD, Weiss A, Liewald DC, Davies G, Porteous DJ, et al. 2019b. Genetic contributions to two special factors of neuroticism are associated with affluence, higher intelligence, better health, and longer life. *Mol. Psychiatry.* **https://doi.org/10.1038/s41380-019-0387-3**

Hudson KL, Holohan MK, Collins FS. 2008. Keeping pace with the times—the Genetic Information Nondiscrimination Act of 2008. *N. Engl. J. Med.* 358(25):2661–63

Jaffee SR, Lombardi CM, Coley RL. 2013. Using complementary methods to test whether marriage limits men's antisocial behavior. *Dev. Psychopathol.* 25(1):65–77

Jensen AR. 1968. Social class, race, and genetics: implications for education. *Am. Educ. Res. J.* 5(1):1–42

Joshi PK, Fischer K, Schraut KE, Campbell H, Esko T, Wilson JF. 2016. Variants near *CHRNA3/5* and *APOE* have age- and sex-related effects on human lifespan. *Nat. Commun.* 7(1):11174

Karavani E, Zuk O, Zeevi D, Atzmon G, Barzilai N, et al. 2019. Screening human embryos for polygenic traits has limited utility. *Cell* 179(6):1424–35

Karlsson Linnér R, Biroli P, Kong E, Meddens SFW, Wedow R, et al. 2019. Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. *Nat. Genet.* 51(2):245–57

Keller MC. 2014. Gene × environment interaction studies have not properly controlled for potential confounders: the problem and the (simple) solution. *Biol. Psychiatry* 75(1):18–24

Keyes KM, Westreich D. 2019. UK Biobank, big data, and the consequences of non-representativeness. *Lancet* 393(10178):1297

Khan R, Mittelman D. 2018. Consumer genomics will change your life, whether you get tested or not. *Genome Biol.* 19(1):120

Khramtsova EA, Davis LK, Stranger BE. 2019. The role of sex in the genomics of human complex traits. *Nat. Rev. Genet.* 20(3):173–90

Kim J, Edge MD, Algee-Hewitt BFB, Li JZ, Rosenberg NA. 2018. Statistical detection of relatives typed with disjoint forensic and biomedical loci. *Cell* 175(3):848–58.e6

Kim MS, Patel KP, Teng AK, Berens AJ, Lachance J. 2018. Genetic disease risks can be misestimated across global populations. *Genome Biol.* 19(1):179

Kohler H-P, Rodgers JL, Christensen K. 1999. Is fertility behavior in our genes? Findings from a Danish twin study. *Popul. Dev. Rev.* 25(2):253–88

Kong A, Thorleifsson G, Frigge ML, Vilhjalmsson B, Young AI, et al. 2018. The nature of nurture: effects of parental genotypes. *Science* 359:424–28

Kranzler HR, Zhou H, Kember RL, Vickers Smith R, Justice AC, et al. 2019. Genome-wide association study of alcohol consumption and use disorder in 274,424 individuals from multiple populations. *Nat. Commun.* 10(1):1499

Lee JJ, Wedow R, Okbay A, Kong E, Maghzian O, et al. 2018. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* 50(8):1112–21

Lee SH, Wray NR, Goddard ME, Visscher PM. 2011. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* 88(3):294–305

Liu H. 2018. Social and genetic pathways in multigenerational transmission of educational attainment. *Am. Sociol. Rev.* 83(2):278–304

Liu M, Jiang Y, Wedow R, Li Y, Brazel DM, et al. 2019. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat. Genet.* 51(2):237–44

Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, et al. 2015. Genetic studies of body mass index yield new insights for obesity biology. *Nature* 518(7538):197–206

Maney DL. 2017. Polymorphisms in sex steroid receptors: From gene sequence to behavior. *Front. Neuroendocrinol.* 47:47–65

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. 2009. Finding the missing heritability of complex diseases. *Nature* 461(7265):747–53

Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, et al. 2017. Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet.* 100(4):635–49

Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. 2019. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* 51(4):584–91

Martschenko D, Trejo S, Domingue BW. 2019. Genetics and education: recent developments in the context of an ugly history and an uncertain future. *AERA Open* 5(1):233285841881051

Mathieson I, Day FR, Barban N, Tropf FC, Brazel DM, et al. 2020. Genome-wide analysis identifies genetic effects on reproductive success and ongoing natural selection at the *FADS* locus. bioRxiv. **https://doi.org/10.1101/2020.05.19.104455**

McBride CM, Koehly LM, Sanderson SC, Kaphingst KA. 2010. The behavioral response to personalized genetic information: Will genetic risk profiles motivate individuals and families to choose more healthful behaviors? *Annu. Rev. Public Health* 31:89–103

Mills MC. 2019. How do genes affect same-sex behavior? *Science* 365(6456):869–70

Mills MC, Barban N, Tropf F. 2018. The sociogenomics of polygenic scores of reproductive behavior and their relationship to other fertility traits. *RSF Russell Sage Found. J. Soc. Sci.* 4(4):122–36

Mills MC, Barban N, Tropf FC. 2020a. *An Introduction to Statistical Genetic Data Analysis*. Cambridge, MA: MIT Press

Mills MC, Rahal C. 2019. A scientometric review of genome-wide association studies. *Commun. Biol.* 2:9

Mills MC, Rahal C. 2020. The GWAS Diversity Monitor tracks diversity by disease in real-time. *Nat. Genet.* 52:242–43

Mills MC, Tropf FC. 2015. The biodemography of fertility: a review and future research frontiers. *Kölner Z. Soz. Sozialpsychol.* 67(Suppl. 1):397–424

Mills MC, Tropf FC, Brazel DM, van Zuydam N, Vaez A, et al. 2020b. Identification of 370 loci for age at onset of sexual and reproductive behavior, highlighting common aetiology with reproductive biology, externalizing behavior and longevity. bioRxiv. **https://doi.org/10.1101/2020.05.06.081273**

Monroe SM, Simons AD. 1991. Diathesis-stress theories in the context of life stress research: implications for the depressive disorders. *Psychol. Bull.* 110:406–25

Moore R, Casale FP, Jan Bonder M, Horta D, Franke L, et al. 2019. A linear mixed-model approach to study multivariate gene-environment interactions. *Nat. Genet.* 51(1):180–86

Morris TT, Davies NM, Smith GD. 2019. Can education be personalised using pupils' genetic data? bioRxiv 645218. **https://doi.org/10.1101/645218**

Mostafavi H, Berisa T, Day FR, Perry JRB, Przeworski M, Pickrell JK. 2017. Identifying genetic variants that affect viability in large cohorts. *PLOS Biol.* 15(9):e2002458

Mostafavi H, Harpak A, Agarwal I, Conley D, Pritchard JK, Przeworski M. 2020. Variable prediction accuracy of polygenic scores within an ancestry group. *eLife* 9:e48376

Mullins N, Bigdeli TB, Børglum AD, Coleman JRI, Demontis D, et al. 2019. GWAS of suicide attempt in psychiatric disorders and association with major depression polygenic risk scores. *Am. J. Psychiatry* 176(8):651–60

Nelson A. 2016. *The Social Life of DNA*. Boston: Beacon Press

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, et al. 2008. Genes mirror geography within Europe. *Nature* 456(7218):98–101

Okbay A, Baselmans BML, De Neve J-E, Turley P, Nivard MG, et al. 2016a. Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nat. Genet.* 48(6):624–33

Okbay A, Beauchamp JP, Fontana MA, Lee JJ, Pers TH, et al. 2016b. Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* 533(7604):539–42

Panofsky A, Bliss C. 2017. Ambiguity and scientific authority. *Am. Sociol. Rev.* 82(1):59–87

Perry JRB, Day FR, Elks CE, Sulem P, Thompson DJ, et al. 2014. Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature* 514(7520):92–97

Pilling LC, Kuo C-L, Sicinski K, Tamosauskaite J, Kuchel GA, et al. 2017. Human longevity: 25 genetic loci associated in 389,166 UK biobank participants. *Aging* 9(12):2504–20

Plomin R. 2018. *Blueprint: How DNA Makes Us Who We Are*. Cambridge, MA: MIT Press

Plomin R, DeFries JC, Loehlin JC. 1977. Genotype-environment interaction and correlation in the analysis of human behavior. *Psychol. Bull.* 84(2):309–22

Plomin R, von Stumm S. 2018. The new genetics of intelligence. *Nat. Rev. Genet.* 19(3):148–59

Polderman TJC, Benyamin B, de Leeuw CA, Sullivan PF, van Bochoven A, et al. 2015. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat. Genet.* 47(7):702–9

Popejoy AB, Fullerton SM. 2016. Genomics is failing on diversity. *Nature* 538(7624):161–64

Reardon J. 2017. *The Postgenomic Condition: Ethics, Justice & Knowledge After the Genome*. Chicago: Univ. Chicago Press

Rice F, Lewis G, Harold GT, Thapar A. 2013. Examining the role of passive gene-environment correlation in childhood depression using a novel genetically sensitive design. *Dev. Psychopathol.* 25(1):37–50

Rietveld CA, Medland SE, Derringer J, Yang J, Esko T, et al. 2013. GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* 340(6139):1467–71

Rimfeld K, Krapohl E, Trzaskowski M, Coleman JRI, Selzam S, et al. 2018. Genetic influence on social outcomes during and after the Soviet era in Estonia. *Nat. Hum. Behav.* 2(4):269–75

Robinson GE, Grozinger CM, Whitfield CW. 2005. Sociogenomics: social life in molecular terms. *Nat. Rev. Genet.* 6(4):257–70

Robinson MR, Kleinman A, Graff M, Vinkhuyzen AAE, Couper D, et al. 2017. Genetic evidence of assortative mating in humans. *Nat. Hum. Behav.* 1:0016

Rothman K, Gallacher J, Hatch E. 2013. Why representativeness should be avoided. *Int. J. Epidemiol.* 42(4):1012–14

Sanchez-Roige S, Fontanillas P, Elson SL, Gray JC, de Wit H, et al. 2019. Genome-wide association studies of impulsive personality traits (BIS-11 and UPPSP) and drug experimentation in up to 22,861 adult research participants identify loci in the *CACNA1I* and *CADM2* genes. *J. Neurosci.* 39:2562–72

Savage JE, Jansen PR, Stringer S, Watanabe K, Bryois J, et al. 2018. Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat. Genet.* 50(7):912–19

Scarr-Salapatek S. 1971. Race, social class, and IQ. *Science* 174(4016):1285–95

Schmitz LL, Conley D. 2017. The effect of Vietnam-era conscription and genetic potential for educational attainment on schooling outcomes. *Econ. Educ. Rev.* 61:85–97

Schwartz CR. 2013. Trends and variation in assortative mating: causes and consequences. *Annu. Rev. Sociol.* 39:451–70

Selzam S, Ritchie SJ, Pingault J-B, Reynolds CA, O'Reilly PF, Plomin R. 2019. Comparing within- and between-family polygenic score prediction. *Am. J. Hum. Genet.* 105(2):351–63

Shanahan M, Hofer S. 2005. Social context in gene-environment interactions: retrospect and prospect. *J. Gerontol. B Psychol. Sci. Soc. Sci.* 60(Special Issue 1):65–76

Sorokin P. 1927. *Social Mobility*. New York: Harper and Brothers

South SC, Krueger RF. 2013. Marital satisfaction and physical health: evidence for an orchid effect. *Psychol. Sci.* 24:373–78

Stolk L, Perry JRB, Chasman DI, He C, Mangino M, et al. 2012. Meta-analyses identify 13 loci associated with age at menopause and highlight DNA repair and immune pathways. *Nat. Genet.* 44(3):260–68

Stulp G, Barrett L, Tropf FC, Mills M. 2015. Does natural selection favour taller stature among the tallest people on earth? *Proc. R. Soc. B Biol. Sci.* 282(1806):20150211

Trejo S, Belsky DW, Boardman JD, Freese J, Harris KM, et al. 2018. Schools as moderators of genetic associations with life course attainments: evidence from the WLS and Add Health. *Sociol. Sci.* 5:513–40

Trejo S, Domingue BW. 2019. Genetic nature or genetic nurture? Quantifying bias in analyses using polygenic scores. bioRxiv 524850. **https://doi.org/10.1101/524850**

Tropf FC, Hong Lee S, Verweij RM, Stulp G, van der Most PJ, et al. 2017. Hidden heritability due to heterogeneity across seven populations. *Nat. Hum. Behav.* 1:757–65

Tropf FC, Stulp G, Barban N, Visscher P, Yang J, et al. 2015. Human fertility, molecular genetics, and natural selection in modern societies. *PLOS ONE* 10(6):e0126821

Tucker-Drob EM. 2017. Measurement error correction of genome-wide polygenic scores in prediction samples. bioRxiv 165472. **https://doi.org/10.1101/165472**

Tucker-Drob EM, Bates TC. 2016. Large cross-national differences in gene × socioeconomic status interaction on intelligence. *Psychol. Sci.* 27(2):138–49

Turkheimer E, Haley A, Waldron M, D'Onofrio B, Gottesman II. 2003. Socioeconomic status modifies heritability of IQ in young children. *Psychol. Sci.* 14(6):623–28

Turley P, Walters RK, Maghzian O, Okbay A, Lee JJ, et al. 2018. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.* 50(2):229–37

Udry JR. 1995. Sociology and biology: What biology do sociologists need to know? *Soc. Forces* 73(4):1267–78

van Rheenen W, Peyrot WJ, Schork AJ, Lee SH, Wray NR. 2019. Genetic correlations of polygenic disease traits: from theory to practice. *Nat. Rev. Genet.* 20(10):567–81

Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, et al. 2015. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* 97(4):576–92

Visscher PM, Hill WG, Wray NR. 2008. Heritability in the genomics era—concepts and misconceptions. *Nat. Rev. Genet.* 9(4):255–66

Ware EB, Schmitz LL, Faul JD, Gard AM, Mitchell C, et al. 2018. Heterogeneity in polygenic scores for common human traits. bioRxiv 106062. **https://doi.org/10.1101/106062**

Watanabe K, Stringer S, Frei O, Umićević Mirkov M, de Leeuw C, et al. 2019. A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* 51(9):1339–48

Wedow R, Zacher M, Huibregtse BM, Mullan Harris K, Domingue BW, Boardman JD. 2018. Education, smoking, and cohort change: forwarding a multidimensional theory of the environmental moderation of genetic effects. *Am. Sociol. Rev.* 83(4):802–32

Wertz J, Belsky J, Moffitt TE, Belsky DW, Harrington HL, et al. 2019a. Genetics of nurture: a test of the hypothesis that parents' genetics predict their observed caregiving. *Dev. Psychol.* 55(7):1461–72

Wertz J, Moffitt TE, Agnew-Blais J, Arseneault L, Belsky DW, et al. 2019b. Using DNA from mothers and children to study parental investment in children's educational attainment. *Child Dev*. In press

Wood AR, Esko T, Yang J, Vedantam S, Pers TH, et al. 2014. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* 46(11):1173–86

Wray NR, Maier R. 2014. Genetic basis of complex genetic disease: the contribution of disease heterogeneity to missing heritability. *Curr. Epidemiol. Rep.* 1(4):220–27

Wright KM, Rand KA, Kermany A, Noto K, Curtis D, et al. 2019. A prospective analysis of genetic variants associated with human lifespan. *G3* 9(9):2863–78

Yang J, Lee SH, Goddard ME, Visscher PM. 2011. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 88(1):76–82

Yengo L, Sidari M, Verweij KJH, Visscher PM, Keller MC, Zietsch BP. 2019. No evidence for social genetic effects or genetic similarity among friends beyond that due to population stratification: a reappraisal of Domingue et al. (2018). *Behav. Genet.* 50:67–71

Young AI, Benonisdottir S, Przeworski M, Kong A. 2019. Deconstructing the sources of genotype-phenotype associations in humans. *Science* 365(6460):1396–400

Zeng Y, Nie C, Min J, Chen H, Liu X, et al. 2018. Sex differences in genetic associations with longevity. *JAMA Netw. Open* 1(4):e181670

Zenin A, Tsepilov Y, Sharapov S, Getmantsev E, Menshikov LI, et al. 2019. Identification of 12 genetic loci associated with human healthspan. *Commun. Biol.* 2:41