Socio-Genomic Research Using Genome-Wide Molecular Data

Dalton Conley

Department of Sociology, Princeton University, Princeton, NJ 08544;
email: dconley@princeton.edu

Annu. Rev. Sociol. 2016. 42:275–99

First published online as a Review in Advance on May 23, 2016

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

This article's doi:
10.1146/annurev-soc-081715-074316

Keywords

socio-genomics, biosociology, race, behavior genetics, assortative mating, social stratification

Abstract

Recent advances in molecular genetics have provided social scientists with new tools with which to explore human behavior. By deploying genomic analysis, we can now explore long-term patterns of human migration and mating, explore the biological aspects of important sociological outcomes such as educational attainment, and, most importantly, model gene-by-environment interaction effects. The intuition motivating much socio-genomic research is that to have a more complete understanding of social life, scholars must take into consideration both nature and nurture as well as their interplay. Most promising is gene-by-environment research that deploys polygenic measures of genotype as a prism through which to refract and detect heterogenous treatment effects of plausibly exogenous environmental influences. This article reviews much recent work in this vein and argues for a broader integration of genomic data into social inquiry.

INTRODUCTION

Until recently, the study of human genetic variation had consisted mainly of behavioral genetic studies, which use twin and adoption designs to identify heritable, or genetic, variation in various traits (see, e.g., Björklund et al. 2006, Plomin 2009, Plomin et al. 1994, Plug 2004, Sacerdote 2007). Whether or not one believes the estimates of genetic influence on phenotypes such as IQ, income, or personality that emerge from such studies, the fact remains that they do not directly measure genotypes and are of limited utility for social scientists. Today, however, the costs of comprehensively genotyping subjects have fallen to the point where major funding bodies, including those in the social and behavioral sciences, can now begin to incorporate genetic and biological markers into major social surveys. For example, the Wisconsin Longitudinal Study (WLS) and the Health and Retirement Survey (HRS) have released data sets with comprehensively genotyped subjects. Similar efforts are also under way in Europe, for example with the Biobank Project in the United Kingdom (Ollier et al. 2005, Platt et al. 2010) and large-scale genotyping of subjects at several European twin registries (Rønningen et al. 2006). These samples contain large numbers of extensively genotyped individuals and thus provide new opportunities for social and behavioral scientists to ask questions that could not be explored until very recently.

The presence of data measuring the most common forms of human genetic variation, single nucleotide polymorphisms (SNPs; variations in base pairs at specific points along a chromosome that are present in at least 1% of a population) and copy number variants (CNVs; variations in patterns of nucleotide repeats[1]), allows for several lines of research that were basically infeasible under the old regime of twin-based imputed heritability analysis: (*a*) direct modeling of genotype as a moderator of the social influences on behavior; (*b*) assessment of genetic homophily in populations; and (*c*) accurate characterization of the continental ancestry (and admixture) of subpopulations and of the role of genetics in macro-level outcomes. The present review addresses each of these three strands of research in turn.

It bears mentioning that this article will not cover much ongoing research in the classical twin or adoption literature (i.e., nonmolecular approaches to social genetics) or extant candidate gene studies (i.e., research that focuses on one or a handful of genetic variants). Good reviews of the twin and adoption literature can be found elsewhere (see, e.g., Batouli et al. 2014, Silventoinen et al. 2010). These studies are controversial, and the assumptions underlying them have been questioned (e.g., Goldberger 1979; for a defense, see Conley et al. 2013, Scarr & Carter-Saltzman 1979). With respect to the candidate gene literature, most of these studies have failed to replicate, causing even the flagship journal *Behavior Genetics* to adopt a policy of not publishing such studies (Hewitt 2012; for a review of why false positives are rife with this approach, see Chabris et al. 2012). Here, I review work that employs genome-wide data that better allow for the assessment of polygenic effects on behavior as well as for the reduction of bias due to gene–environment correlation.

GENE–ENVIRONMENT INTERACTIONS

The move to studying SNPs and other genetic polymorphisms such as CNVs has opened up a particularly promising research program on genetic–(social) environmental interactions in human populations. The estimation of such interaction effects has long been a goal of social and behavioral scientists fond of expressing the dependence of genetic expression on social structure. Since at least the publication in *Science* (Caspi et al. 2002, 2003) of empirical evidence suggesting

[1]An example would be a case in which some individuals have a string of TTATTATTA, whereas others have five repeats of TTA.

gene–environment interactions (G×E interactions), there has been a growing interest in integrating biological and social science approaches, data, and models. Attempting to partially answer the question of why some individuals are resilient to stressors whereas others suffer deleterious psychological sequelae, Caspi et al. (2002, 2003) suggested an important genetic source of heterogeneity in response to adverse early-life events. Although these studies created substantial interest in potential G×E interactions, they also required replication and extension by researchers using alternative data sources. Indeed, there are now competing meta-analyses suggesting that the original results linking differential response to stress by *5-HTT* genotype either are reasonably robust (Karg et al. 2011) or lack consistent supporting replication (Risch et al. 2009).

The discussion generated by this line of research in the social science community has been productive mainly because it has led to a greater appreciation of the shortcoming of Caspi et al.'s research design—namely, that the alleles and the proposed environmental modifiers may not be randomly assigned in the population and may therefore be correlated with unobserved causal factors. For example, it may be the case that an observed interaction between a genetic variant and environmental exposure reflects a differential risk of exposure (e.g., genes selecting environments) rather than the genetic modification of exogenous environmental exposures. This is known as gene–environment correlation. In this way, measured environments—particularly when fashioned by parents who also pass on their genes to the respondents—may be correlated with unmeasured genetic variation and thus could be acting as proxies for a gene-by-gene interaction rather than a G×E interaction.

Most G×E studies that do manage to obtain adequate causal identification on the environmental side through sibling difference models or other natural experiments have all focused on the interaction of one SNP or CNV with a given exogenous environmental shock (cf. Conley & Rauscher 2013, Cook & Fletcher 2013, Fletcher 2012). This is problematic for at least two reasons: First, with data that contain only a few genetic markers, it is quite difficult to address the problem of population stratification. That is, whereas it is often possible that environmental measures are acting as proxies for unobserved genotypes, thus leading to biased estimates, it is also possible that apparent genetic effects are false positives, the result of the confounding of genotypes and environment through population stratification. This concept was popularized by Hamer & Sirota (2000), who used the example of a "chopstick gene" appearing because of data that mix Asians and Caucasians.

The above-mentioned studies all limit their samples to non-Hispanic whites; however, even within an ethnically homogenous population, genotypes may be acting as proxies for different places or social environments (Benjamin et al. 2012a, Cardon & Palmer 2003, Conley et al. 2014). By moving G×E research to analyses of genome-wide data, population stratification can be addressed by deploying controls for the principal components (PCs) of the variance-covariance structure of the genetic data (Price et al. 2006) and/or by modeling the error structure of the models based on the genetic relatedness matrix, an approach developed by Kang et al. (2010).

Further auguring a genome-wide approach is the fact that most complex social phenotypes are highly polygenic in nature. A polygenic score (PGS, formerly termed a polygenic risk score) is an attempt to use the SNP data in a given sample to construct a predictive equation for an outcome in that sample. By way of example of this latter approach, a recent study analyzed 126,558 individuals from 54 distinct cohorts to search for alleles that may be associated with educational attainment (Rietveld et al. 2013).

Rietveld et al. (2013) conducted what is called a genome-wide association study (GWAS), an atheoretical approach to gene discovery in which hundreds of thousands of SNPs are tested for association with an outcome of interest (McCarthy et al. 2008). In what follows, SNPs are indexed by j and individuals by i. Each individual SNP is tested for association by running a regression of

the sort shown in Equation 1,

$$y_i = \mu + \beta_j x_{ij} + Z_i \gamma_i + \epsilon_i, \quad 1.$$

where x_{ij} is the number of reference alleles that individual i is endowed with at SNP j, and Z is a vector of controls that include age, sex, and the first four PCs of the variance-covariance matrix of the genotypic data. The PCs are included to guard against the problem of population stratification—the tendency for allele frequencies to covary with unobserved environmental confounds.

Because the number of hypotheses that were tested is very large, it is common to declare a SNP association to be significant if it reaches a P value of 5×10^{-8}. Rietveld et al. (2013) identified three SNPs that reached this level of significance, and all three replicated in an independent sample. Further, versatile gene-based association study analyses, which pool SNPs into genes to increase the power to detect genes that may affect education, also found 17 genes that were significantly related to education, many of which have been associated with central nervous system processes—specifically, expression in the anterior caudate nucleus. The authors also deployed pathway analysis, or interval enrichment analysis (Lee et al. 2012), to find biological processes that were enriched in the SNP data. Finally, the greatest significance of this study is that it allows for the construction of a PGS for educational attainment. A common approach to constructing such a PGS, which is labeled $\hat{g}$, is to take a weighted sum of SNPs in which the weights are given by the estimated β_j coefficients from Equation 2,

$$\hat{g}_i = \sum_{j=1}^{J} x_{ij} \beta_j. \quad 2.$$

[For other examples of PGS deployment, see, e.g., Belsky et al. (2012; 2013a,b), Benjamin et al. (2012b), Purcell et al. (2009), Visscher et al. (2010), and Yang et al. (2010).]

Thus, although only three alleles reached what geneticists call genome-wide significance ($p < 5 \times 10^{-8}$) and replicated in the independent samples, these explained a trivial amount of the total variance in years of schooling or college attendance. Meanwhile, relaxing the threshold continually increases the predictive power of the genetic risk score up to the point at which all SNPs are taken into account regardless of significance level. This suggests that, to the extent that it is associated with genotype, educational attainment—as we might expect—is driven by many small effects across the entire genome.

The out-of-sample predictive power that can be obtained from considering the SNP data simultaneously is presently too small to be of practical use for most outcomes. For example, Rietveld et al. (2013) explain 2–3% of the variation in educational attainment in independent samples. Meanwhile, the International Schizophrenia Consortium (Purcell et al. 2009) reported an out-of-sample predictability of up to 3% from a predictive risk equation estimated in a total sample of 6,907 individuals. This predictive accuracy will, however, improve with larger samples and more comprehensive genotyping platforms (Chatterjee et al. 2013, Daetwyler et al. 2008, Dudbridge 2013).

Armed with PGSs, researchers have begun to assess social factors that may moderate their effects by making a PGS interact with a putatively exogenous form of environmental variation. There is a small but growing literature that has attempted to separate gene–environment correlation and interplay by making use of environmental variation that is not likely the result of (or correlated with) genetic variation in a genome-wide context where population stratification can be factored out through the use of PCs. For example, a study by Schmitz & Conley (2015) used the Vietnam era draft lottery as an instrumental variable to ask whether the risk of serving during the Vietnam War interacted with a smoking PGS to predict smoking status during adulthood. They found

evidence that veterans with a high genetic predisposition for smoking were more likely to become regular smokers, to smoke heavily, and to have a higher risk of being diagnosed with cancer or hypertension at older ages than nonveterans are.

Other recent studies have used birth cohort variation to assess how shifting macro-level environments affect the influence of genotype. Domingue et al. (2016) found that the genetic effects on smoking have increased over the course of the twentieth century as the deleterious health effects of tobacco use have become more widely appreciated in the US population. The implied story is as follows: Most American youth or young adults try cigarettes at some point in their lives; the question is whether they become addicted and/or are able to quit or whether they turn into lifelong smokers. In an environment where the perceived costs of smoking appear steep, only those with a genotype that makes nicotine dependence more severe end up as regular smokers, thereby increasing the genetic penetrance.

Rietveld et al. (2015) have applied this same sort of cohort approach to exogenous variation in the environmental landscape to the Swedish educational system. They find that a PGS for education (a second version of the original one) predicts educational attainment better in older birth cohorts than in younger respondents. They rule out competing explanations of mortality bias and ascertainment bias (due to mismatch between ages in the discovery or training samples and the replication sample), and instead hypothesize that policy reforms that opened up access and increased the level of compulsory schooling attenuated the effect of genotype across the twentieth century.

A classic area for this interrogation of G×E effects has been the claim that the heritability of IQ (and by extension educational attainment) varies by socioeconomic status (SES) and race. Turkheimer et al. (2003) have argued that when it comes to the effect of genes, there is little equality. Twins coming from the low end of the socioeconomic distribution demonstrate a low heritability of intelligence, whereas those at the top approach complete genetic penetrance. According to the authors' interpretation of these results, those born to families who are socially advantaged enjoy financial and nonfinancial resources that ensure that they reach their full genetic potential—at least when it comes to cognitive measures. By contrast, those who are socially disadvantaged face the deficit that comes from having to start their climb from the bottom of a deep hole, and the lack of resources they experience growing up also eliminates genetic distinctions that would have otherwise naturally emerged in an equal-opportunity society. [Guo & Stearns (2002) make a similar point to Turkheimer and colleagues (2003) with respect to race.]

This theory is appealing because it makes intuitive sense: Think of a discriminatory social environment where everyone who is black or who is poor ends up being unable to attend college or apply for good jobs. Such a dynamic was observed—with respect to race—as far back as 1967, when sociologists Blau & Duncan (1967) coined the term "perverse equality" in their book *The American Occupational Structure* to describe a situation where the class background of African Americans had relatively little influence on their occupational attainment. Discrimination held back the children of black doctors as much as it did the offspring of black ditch diggers. Meanwhile, tokenism ensured that for each generation there emerged a "talented tenth" (to use the language of W.E.B. Du Bois)—a cadre of black professionals—whose outcomes could not be well predicted by their backgrounds.

Although Turkheimer and colleagues (2003) focus on IQ and Blau & Duncan (1967) discuss race and the inheritance of occupation, their stories coalesce nicely. It is a politically appealing story as well, since it manages to incorporate the arguments and evidence that genetics does indeed matter—in contrast to more extreme theories that posit a blank slate or pure nurturance—and yet preserves a very important role for unequal environments. Even more appealing is the fact that such a pattern suggests that redistribution not only would serve to equalize outcomes (i.e., fill in

the hole in which low SES children find themselves), but it would also lead to greater economic efficiency by unleashing the underutilized, latent genetic talent at the bottom end. In this way, it may be cost neutral or even revenue positive from a public economics standpoint.

In fact, though some would see it as a dystopian nightmare, certain scholars have argued that we should actually strive for a world where socioeconomic measures of success are wholly genetically determined—i.e., where heritability is as close to 100% as possible. Any other social effects, especially those of family environment, are inefficient and unfair. Specifically, some sociologists have suggested that we should abandon raw or adjusted mobility rates (or intergenerational earnings elasticities) as measures of openness and meritocracy. Rather, Guo & Stearns (2002) and Nielsen (2006, 2008), among others, argue that we should compare the genetic component to the common environmental component of social status as determined by twin- and other kin-based variance decomposition models. In this paradigm, it is not the overall correlation between siblings, for instance, that measures the relative openness or closure of a stratification system (Björklund et al. 2002, Corcoran et al. 1992, Hauser & Sewell 1986, Hauser et al. 1999, Kuo & Hauser 1995, Olneck 1976, Page & Solon 2003, Warren & Hauser 1997, Warren et al. 2002), but rather the proportion of that correlation that is due to shared genotype. That is, fundamentally unjust societies exhibit low heritability estimates where the genetic potential of the population is not fully realized because social factors are primarily responsible for phenotypic variation.

In this view, a meritocratic society would display a high genetic component to achieved social position and a low common (read: familial) environmental component. According to this argument, policy should aim to enhance sorting on innate characteristics and not on the social advantages or disadvantages that may be conferred by birth and upbringing (Heath et al. 1985). In this framework, inequality and inefficiency are captured by the fact that low SES (and black) individuals demonstrate a greater environmental component and a smaller genetic one with respect to critical outcomes.

But are such empirical claims of a genotype–SES interaction accurate? Could these outcomes be attributed not to a stronger environmental effect within the lower end of the socioeconomic distribution but rather to a weaker genetic effect, perhaps because of a different distribution of genotypes due to a lesser degree of assortative mating? The observed pattern of results would be the same, but the implications would be entirely different and might suggest an intervention in the mating market rather than in the educational system.

With twin models, such as those used by Guo & Sterns (2002) or Turkheimer and colleagues (2003), it is hard to know what exactly is going on, because the genotype is inferred rather than measured. A next step, then, would be to measure the heretofore unmeasured genetic predictors of academic achievement and ask more directly whether the genetic predictors do a worse job predicting achievement in lower-SES households and a better job predicting achievement in higher-SES households. To measure this unmeasured variable—the underlying genetic architecture of academic achievement—Conley et al. (2015) deployed Rietveld and colleagues' (2013) PGS for education in two novel samples. They asked two related questions: First, does the variance of the polygenic genetic score differ by social class background (as measured by maternal and/or paternal education)? This would be a way to assess the possibility that in previous studies, it was the genetic landscape and not necessarily the environmental landscape that varied by SES. Second, did the impact of the education PGS vary by class background?

The answer to the first question was negative: The spread of scores did not vary by social class. In at least two data sets—the Framingham Heart Study and the Minnesota Twin Family Study—the standard deviations of the raw scores were the same for children of mothers (or fathers) who had only a high school education (or less) and for children whose parents had at least some additional schooling (since these were white samples, they did not directly address the race question). So far, these results supported the environmental explanation favored by Turkheimer

and colleagues (2003), because the results were consistent with the view that the outcomes of poor children were different not because of differences in genes but possibly because of differences in environments. But then Conley et al. (2015) tested whether the effect of the offspring score varied by parental education, and it did not. Though an imperfect test of the argument, the results suggested that individuals with the same genetic "risk" for low education were not affected by environmental factors of low versus high social class. In fact, the only variable that seemed to moderate the effect of the PGS in children was the mother's PGS for educational attainment. When genotypically educationally advantaged children had high-genotype mothers, they got an extra boost compared to those born to a genotypically average mother. Likewise, double genetic disadvantage had multiplicative rather than additive effects. In fact, this parental genetic measure was the only background variable that seemed to show a significant interaction with offspring genotype. No putatively social variable had any influence on the genotype–phenotype relationship for offspring.

Although the PGS approach enjoys the advantage of direct measurement, it suffers from the big disadvantage of capturing only a small portion of the putative genetic effect. Additionally, the PGS is calculated from a meta-analysis of cohorts across a wide range of environments, and thus it may be picking up the genetic effects most robust to environmental differences. Therefore, although the issue of what is going on across class lines is far from resolved, at the very least these newer, molecularly based results point out the potential foolishness of readily assuming that a difference in genetic effects along some population split—race, class, geography, or family type—reflects a true G×E interplay.

GENETIC HOMOPHILY AND HOMOGAMY

Not only has molecular data allowed social scientists to directly measure genotypes that explain a nontrivial amount of variation in behavioral phenotypes, but SNP chips have also allowed for classic heritability analysis that takes advantage of genetic similarity among nonkin to estimate additive genetic effects of social traits using an approach termed genomic-relatedness-matrix restricted maximum likelihood (GREML) [or, alternatively, modified Defries-Fulker regression (Bataille et al. 2002)]. Because this literature does not deal with measured genotype, I will not address it here; the number of outcomes that have been assessed using GREML grows weekly—see, e.g., Peyrot et al. (2015), Sieradzka et al. (2015). However, the approach of counting allele similarity between pairs of individuals to generate a distribution of genetic relatedness has also been used to study the degree to which friends and spouses are genetically assorting.

This question of genetic homophily is important to the estimation of network models and peer effects. It has long been a challenge to network researchers to separate out homophily from social contagion. Obtaining a sense of how much genetic assortment occurs in networks is, therefore, informative to this concern. First, the classic twin models of heritability (as well as other models) assume random mating with respect to genotypes. To the extent that mating deviates from this ideal, it suggests the models are flawed. Second, understanding the extent to which spouses sort on genotypes may also help us better understand the mechanisms of phenotypic assortative mating (for a good review of the assortative mating literature, see Schwartz & Han 2014). Finally, genetic sorting among reproductive mates coupled with differential fertility rates may change the genetic and phenotypic landscapes of subsequent generations.

In this vein, Fowler et al. (2011) found that friends (who were not relatives) were genetically the equivalent of fourth cousins. Not only did friends in their network study share more SNPs than nonfriends, but they also displayed an excess of opposite genotypes, loci that diverged more than they would by chance. When they investigated these respective patterns of homophily and heterophily, Fowler and colleagues found that the homophilic (i.e., like-likes-like) genes tended

to cluster in two biological pathways: linoleic acid metabolism and olfactory perception. These authors also found that there was a certain class of genes that were overrepresented in the heterophilous group: those related to immune function.

Such heterophily has long been theorized with respect to spouses, who are hypothesized to be discordant on their genotypes for a particular region of chromosome six that codes for immunological genes, called the major histocompatibility complex or human leukocyte antigens area. The theory is that evolutionary forces have pushed us to seek diversity in this genetic package that confers biological resistance to disease, so that if an epidemic hits a family or wider tribe, at least some individuals in the group will have native resistance and survive.

This example is an instance of metagenomics, in which the effect of our genes depends on the genetic context around us. In particular, it is a case of negative frequency dependency (known as apostatic selection). By extending the analysis of correlations between the genotypes of friends, other work in this area shows that schools (and, more generally, environments) shape the way these correlations are created. There is a role for social structure to produce correlated genotypes among friends, even if they do not actively seek out similar friends (Boardman et al. 2012).

Domingue et al. (2014) take a similar approach to spousal sorting among couples in the HRS. They tackled the issue of genetic assortative mating in several ways. They found that, overall, spouses were more genetically similar to each other than randomly paired individuals in the population, as shown in **Figure 1**, where the shaded area represents the genetic difference across all measured SNPs for spouses as compared to randomly paired individuals. Non-Hispanic white spouses in the HRS were, on average, not quite as genetically similar to each other as first cousins once removed, and more related than second cousins. Although this analysis was restricted to whites, the authors found that even factoring out population structure (i.e., historical ethnic intramarriage patterns) through deployment of PCs, genetic relatedness among spouses was still

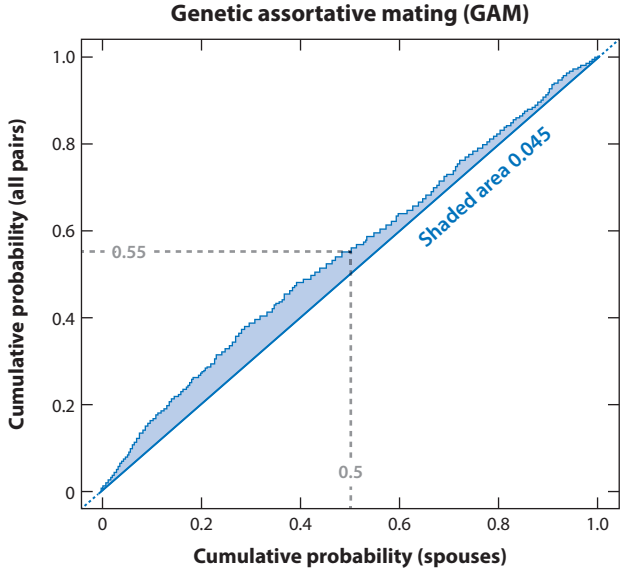

Figure 1

Genetically assortative mating (GAM). Spouses (*shaded area*) are related at a degree of 0.045, equivalent to first cousins once removed in the United States. When ethnicity is factored out through controls for principal components, the level of genetic similarity between spouses falls to that of second cousins.

equivalent to second-cousin status (between 2% and 3% genetic similarity). Indeed, being a standard deviation more similar genetically increased the probability that one would be married to that person by 15%.

RACE, CONTINENTS, AND DIVERSITY

This sorting on genotype in modern society may inspire the question of how genetic sorting, natural selection, population bottlenecks, and drift in premodern times may have generated meaningful genetic differences among populations today. The most important factor is the population bottleneck that occurred around 60,000 years ago, when a group of modern humans left Africa and fanned out through the rest of the world. Evidence suggests that the effective population size (i.e., the mating pool from which others have descended) for this group of out-of-Africa migrants reached a low of around one to two thousand individuals (Cavalli-Sforza et al. 1994).

However, evolution worked for many millennia to allow mutations to build up, causing a great degree of genetic variation in the populations that lived—then and now—in the cradle of human origins. Those who left northeast Africa, however, took with them only the genetic polymorphisms that they happened to have at the time—i.e., a subset of all the contemporaneous variations in the human species. Of course, new mutations have arisen in the 60,000 years since some humans left the Rift Valley, but these have occurred at the same rate inside and outside Africa.

From a socio-genetics perspective, a key result of this population bottleneck during the migration out of Africa is that the most fundamental difference in continental origin is the difference in genetic variation between the two groups—those of African descent and those not of direct African descent. This divergence is due to the fact that when there is a small effective mating population, polymorphisms are likely to die out through fixation; that is, absent selective pressures, the neutral theory suggests that genetic markers rise and fall (and may die out) by chance in a population. If a population is large, the chances that a given marker will disappear by chance in a given generation is quite low; but in a small population, this is much more likely to happen. Indeed, as humans fanned out across the world, genetic diversity became more and more limited: Migratory distance from East Africa becomes a very good proxy for genetic diversity in a population.

There are many ways to measure genetic diversity, but the simplest is the rate of heterozygosity (the frequency at which individuals do not have the same alleles at the same locus on both chromosomes), because if a particular polymorphism is 50/50 in the population, the rate of heterozygosity is going to be high (50% assuming random mating, or more precisely, that the alleles are in Hardy-Weinberg equilibrium); but if it is close to zero, the proportion of individuals who will be heterozygous at that locus is practically null.

Another way to assess genetic diversity in a group is the variation not in SNP frequencies (i.e., heterozygosity) but in copy number repeats. In addition to base pair substitutions, another common form of genetic variation is the number of times a given sequence (e.g., AGGTCT) repeats in a row. More diversity means more variation in the number of repeats of these repetitive sequences. We can see in **Figure 2** that the two measures track each other very well. Further, we can see that, with some exceptions, the African groups cluster at the top right (i.e., more diversity) of the graph, whereas the Native American and Pacific Islanders are in the lower left.

Figure 2*b* shows an enlarged version of panel *a*, focusing on the upper right portion of the graph. Note that some of the tribes labeled in this bottom illustration include North Carolina and Pittsburgh, with high degrees of genetic variation. These are samples of African Americans from these geographic localities, demonstrating that despite admixing (i.e., breeding) with people of European and Native American descent, and despite the potential founder effect (i.e., bottleneck)

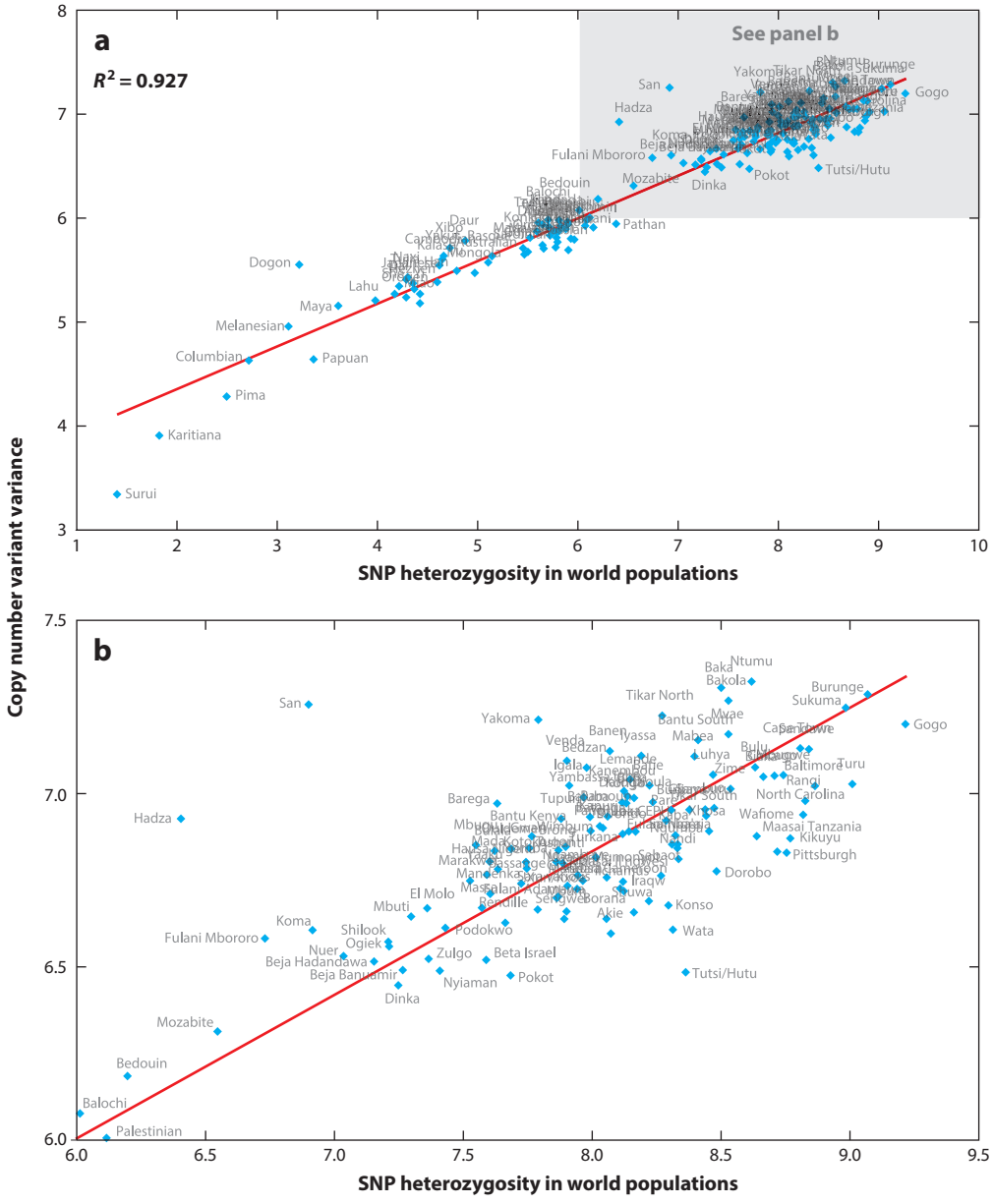

Figure 2

Measures of genetic diversity plotted against each other: SNP (single nucleotide polymorphism) heterozygosity (*x* axis) against copy number variant variance (*y* axis) for (*a*) world populations and (*b*) a subset from the upper right quadrant of the world population plot. Adapted with permission from Tishkoff et al. (2009, supplemental figure S4).

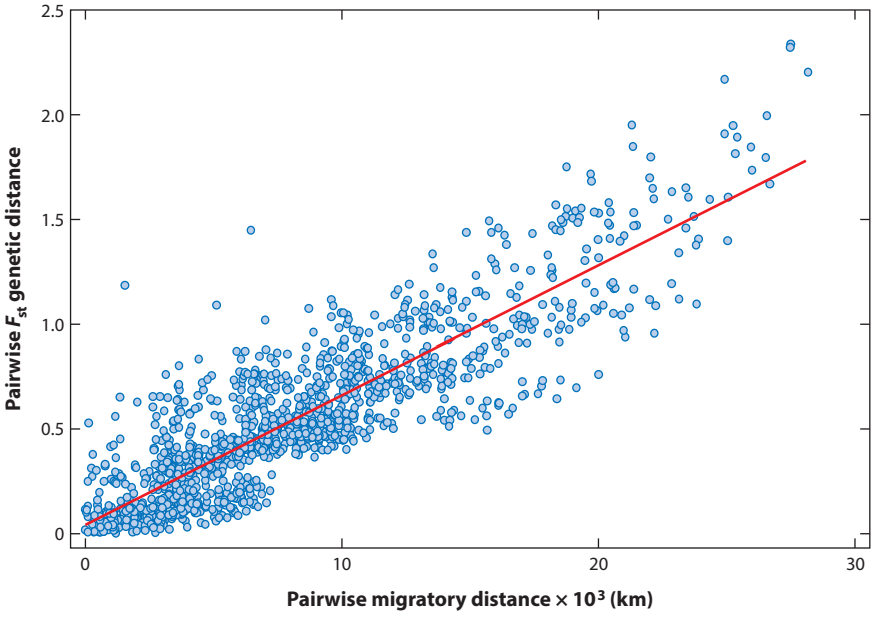

Figure 3

Plot of the F-statistic (F_{st}), a measure of genetic diversity, against migratory distance from East Africa. Adapted with permission from Ashraf & Galor (2013, figure B1).

of the Middle Passage in the slave trade, black Americans preserved a relatively high level of within-group genetic variation.

This greater genetic variation among individuals with African origins subsists even though modern humans outside of Africa actually mated with nonhumans (i.e., Neanderthals and Denisovians) and indeed still bear the genetic signature of these cross-species sexual encounters (see, e.g., Sankararaman et al. 2014). (Among Europeans or those of European descent, for example, 1–3% of the genome is of Neanderthal ancestry.) This very strong pattern is more clearly illustrated by plotting the F-statistic against migratory distance from East Africa, as shown in **Figure 3**. (The F-statistic is an alternative measure of genetic diversity that is essentially the difference in heterozygosity from what would be expected from allele frequencies in Hardy-Weinberg equilibrium and the observed rate of heterozygosity in a population or across populations.)

The result of all these patterns is a situation in which the genetic distance between ethnic groups does not align with folk conceptions of race. For example, if we examine the unrooted genetic tree mapped by Tishkoff et al. (2009) and shown in **Figure 4**, we can see that the genetic distance between groups that were sampled within Africa is as great as the genetic distance between some very racially divergent groups in the rest of the world. For example, if we trace the path from East Asian to European, our finger traverses a distance that is less than the one connecting the Hazda in North Central Tanzania to the Fulani shepherds of West Africa (who live in present-day Mali, Niger, Burkina Faso, and Guinea).

In the United States we have historically classified race based on the rule of hypodescent—i.e., the one drop rule—by which any amount of African heritage makes someone black, and we further divide the population into Asian, Native American, and white races. We can clearly see that these categories do not match the genetic distances naturally found in the population (or the admixture across ancestry groups; see Guo et al. 2014). Furthermore, it becomes all the more ironic that

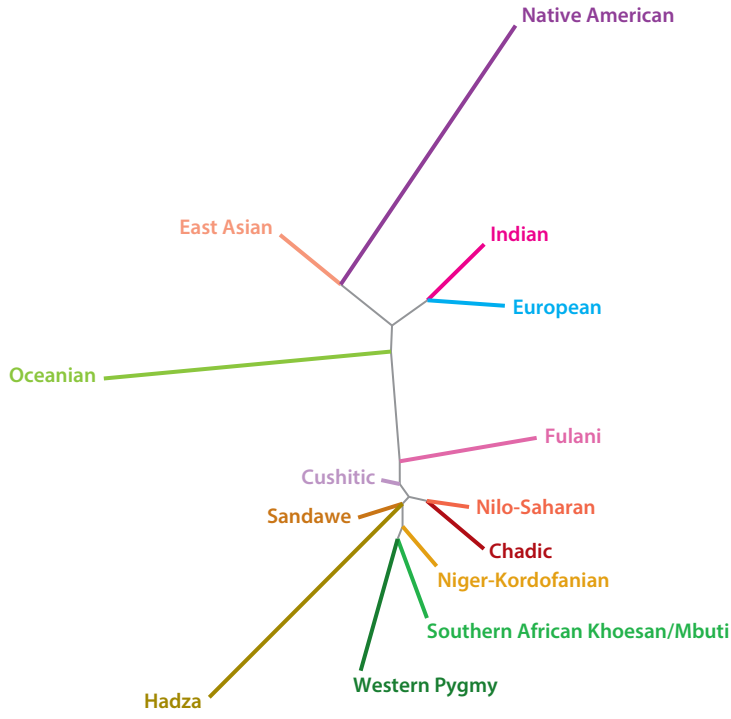

Figure 4

Unrooted ancestry tree showing divergent associated ancestral clusters. Adapted with permission from Tishkoff et al. (2009, supplemental figure S14).

the racial-ethnic system in the United States has eliminated ethnic identity from the racial group where ethnic differences actually have the greatest genetic meaning and significance—i.e., African Americans.

Whereas other groups—Latinos, Native Americans, Asians, and whites—all have ethnic groups within their ranks (in the case of Native Americans these are nations or tribes; in the other cases they are usually associated with countries of origin), black Americans have no ethnic subgroups. The ethnic homogenization of blacks resulted from the slave owners' deliberate mixing of slave populations from different geographic and tribal origins to break down solidarity among the enslaved as one of many methods to prevent revolt.

The mélange of tribal-ethnic origins of African Americans, combined with the complete cutoff from the land of origin and the need to construct new cultural practices once here (and to adapt and mix old ones), means that African Americans were effectively stripped of their ethnic honor—i.e., the pride of belonging to a group with its own history, traditions, and nationhood that exists outside the borders of the immigrant society that is the United States. That is, blacks do not get a St. Patrick's Day or a Cinco de Mayo or a Bastille Day. Indeed, if US celebratory holidays were allocated based on genetic distinctiveness, we would have multiple holidays for each of the several tribes in Kenya and drop St. Patrick's Day altogether.

(DUBIOUS) RACE CLAIMS ABOUND

There have been many myths about continental genetic variation that have been promulgated by both the left and the right. On the left, one of the favorite approaches to discredit the notion that

genetic differences underlie phenotypic differences among human population groups is to point out that there is more genetic variation within these groups than between them. A second approach is to cite the fact that all humans are 99.9% genetically identical and that no group of humans has a gene (i.e., a coded-for protein) that another group lacks. Both of these arguments are canards. After all, we are also more than 98% identical to chimps and 99.7% similar to Neanderthals. Overall genetic variation tells us less than those specific differences that matter do.

Imagine a group of humans that had a mutation in the *FOXP2* gene—often called the language gene—such that this transcription factor (a gene that helps stimulate the expression of select other genes) was nonfunctional. This group of humans would lack the ability to communicate through language. (In fact, this gene's significance was first discovered by the study of an English family in which half the members across three generations suffered from severe developmental verbal dyspraxia—i.e., they could not communicate verbally.)

Moreover, the fact that all humans share the same genes, even if their morphology may differ, ignores that much of evolutionary change and biological difference is about the regulation of those genes' expression rather than the development of novel proteins. In fact, when the human genome project first began, the number of human protein-coded genes was anticipated to be on the order of 100,000 or more. After all, we are certainly more complex that *Zea mays* (i.e., corn) with its 32,000 genes—are we not? As it turns out, we had a mere 20,000; so, most of human difference is driven by the turning on and off of those 20,000 genes in specific tissues and at particular times. The upshot of this is that the simple fact of sharing those same 20,000 genes does not mean that we cannot have very different phenotypic differences based on differences in the regulatory regions of the genome—promoters, enhancers, micro-RNAs, and other molecular switches.

A better question to ask than whether we have different proteins is whether we have different alleles. When we ask whether there are alleles in one population that are not seen in any other human population—the parallel question to the unique genes inquiry—the answer turns out to be affirmative. As shown in **Figure 5**, it is African populations that have the most private (i.e., nonshared) alleles. This is, of course, a reflection of the greater wellspring of diversity in

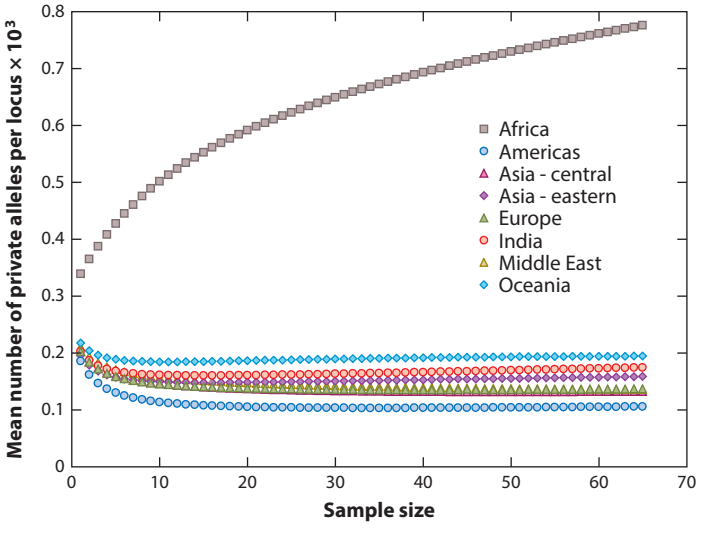

Figure 5

Private alleles (i.e., unique to that population) per locus with variation versus number of alleles (in thousands) sampled by ancestral group. Adapted with permission from Tishkoff et al. (2009, supplemental figure S6).

sub-Saharan Africa compared to the reduction of genetic diversity suffered by some groups as a consequence of the population bottleneck in the migration from Africa to the rest of the world. The point, however, is that there is no a priori reason to discount the potential impact of these private alleles on group differences.

Another argument that the left makes to discredit any genetic basis for observed group differences is that there has not been enough time, evolutionarily speaking, for meaningful differences to emerge. Stephen J. Gould is famously quoted as saying, "There's been no biological change in humans in 40,000 or 50,000 years. Everything we call culture and civilization we've built with the same body and brain" (Gould 2000). According to this viewpoint, human evolution basically ended with the emergence of anatomically modern humans in the Rift Valley.

This position makes sense at first blush. Sixty thousand years is but the blink of an eye compared to the history of life or even of hominids; when we get to parsing differences between groups outside of Africa, that time span drops even more dramatically. However, important group differences can emerge not only through positive selection for novel mutations but also through purifying selection on traits that are highly polygenic in the first place and for which there is plenty of genetic variation already in the genome on which those traits can selectively sort and reproduce.

We already know that height and cognitive ability are highly polygenic, influenced by thousands or perhaps millions of small differences in the human genome. If the tallest individuals bred at higher rates than shorter individuals, an overall genetic shift in the height distribution could be achieved in a matter of a few generations if the reproductive and survival gradient in height (or any trait) were steep enough. Sixty thousand years in this view is not a blink but an eternity. So if there were really different premiums to survival on different behavioral traits—not just IQ but also trust, grit, self-regulation, and so on—we could easily witness genetic divergence over the millennia.

Indeed, this is exactly what controversial anthropologists Gregory Cochran and Henry Harpending have argued in their book *The 10,000 Year Explosion* (Cochran & Harpending 2009). The Neolithic revolution and the rise of sedentary civilizations has led to a condition, they argue, in which human social arrangements, as opposed to the natural landscape, have become the primary driver of changes in population genetics. The result is that many human differences we witness today can be traced to this accelerated, intense selective pressure that agrarian society introduced to favor mental traits like advanced planning at the expense of physical endurance and other traits that would be more favorable for hunter-gatherers. According to these authors, the time since the development of agriculture is a good predictor of how much the genetic landscapes of different populations have adapted to these changed demands for fecundity. They go so far as to argue that the Industrial Revolution was spurred on by genetic changes in Europe, at least in part. Indeed, there is now evidence that selection has been continuing in modern humans (Milot et al. 2011). That evidence aside, however, Cochran & Harpending's case—though plausible—has not been made with the data at hand but rather represents a narrative that ties together much circumstantial evidence.

Those who seek genetic explanations suffer from their own misconceptions, however. For example, scholars like Nicholas Wade, author of the infamous book *A Troublesome Inheritance* (Wade 2014), often focus on genotype differences at one locus that shows significant geographic/ethnic differences as a way to explain huge differences in group outcomes. For example, Wade and others often discuss the *MAO-A* copy number variant as the "warrior gene," because early candidate gene studies showed that this gene predicted violent behavior and other related phenotypes. They then point out that the "violent" allele is found at higher frequencies in the black population. However, as mentioned above, these candidate gene studies—and this one in particular—have not stood up to replication tests that better control for population stratification. And even if they did, they explain a trivial amount of the variation in the measured outcomes, so they are hardly a solid foundation on which to build a genetic model of group differences in behavior.

A second mistake of those who promote genetic explanations is to give too much credit to natural selection and too little to genetic drift. Of course we can see that genetic change (i.e., purifying selection) has occurred to accommodate the variegated environmental landscapes that humans have encountered as they fanned out across the globe. Very obvious examples include the prevalence of the sickle cell genotype—with its protective effect against malaria—only in West African populations, where the malaria incidence is among the highest in the world. A similar example is the clear gradation in dermal melanin expression (i.e., skin tone) as predicted by distance from the equator and its intense sun exposure; or even body morphology as evidenced by Allen's rule, which suggests that in colder climates warm-blooded organisms will tend to have shorter, stockier builds to preserve heat, whereas in hotter climes, big ears, noses, and limbs allow for better heat loss through a greater surface area to body mass ratio. This relationship does indeed obtain in humans as well as in other endotherms.

The mistake that many genetic determinists make is assuming that because we can observe this clear environmental-genetic relationship in some physical characteristics, we can unproblematically expand it to highly complex human behaviors and mental characteristics. The fact that we can see selective pressures at work in generating phenotypic differences in traits that rely on a small number of genes—such as skin tone and eye color or lactose tolerance—does not easily translate to a clear relationship between a highly polygenic trait like, say, cognitive ability and the social or physical landscape. Even in relation to body size, we can observe limb length variation in human populations as predicted by Allen's rule, but limb size is much less polygenic (controlled largely by a series of *HOX* genes) than is overall height; and indeed, height fails to show the latitude-phenotype relationship as clearly. Consider Pygmies versus Bantus (who occupy a similar relation to the equator) or Inuit versus Swedes (who also live at more or less the same latitude). This, of course, does not mean that a highly polygenic trait cannot be subject to intense selective pressure, as much factory-farmed livestock has been for the past six decades (Zuidhof et al. 2014).

Add to the polygenicity of behavioral traits the observed rapid change in economic fortunes—which are meant to be the outgrowth of this rapid selective pressure—over the world during the last 50 years, and a genetic explanation of relative success by ethnic groups in the modern world becomes all the more dubious. Whereas the last 10,000 years is certainly a plausible amount of time for racial or geographic differences in the genetic architecture of social life to emerge, 200 years is probably not, and 50 years is most definitely not [especially a half-century in which the reproductive–class gradient is such that the poor have more surviving offspring than the rich (i.e., successful) within and between countries].

Yet 200 years ago the median wealth of all nations was relatively equal, despite some important differences that were starting to emerge. Meanwhile, since World War II, the genetic landscape of Japan and Taiwan has not changed, but their levels of social and economic development have shot up. (These changes cannot even be attributed to selective migration—as the story of China's rise perhaps can, whereby Shanghai now attains a level of income equal to Italy whereas rural Western areas are more like some African countries.) Even within Europe we have seen huge changes in development over the postwar period—think Ireland or Spain, for example. Thus, there are likely better accounts than genetic ones for explaining geographic variation in standards of living and associated social outcomes, like rule of law, social capital, and so on. For now it suffices to say that genetic differences are a potential, but highly unlikely, explanation for national, racial, or ethnic differences in behavior and socioeconomic success.

Whereas precious little can be said about the role of genetic differences in explaining racial phenotypic differences, there are some areas where the integration of genetic information into traditional analyses of disparities is helpful. For example, Daw (2015) shows that one very serious consequence of the greater genetic diversity among populations of African descent is that it is more

difficult to find organ matches for black Americans than it is for white Americans. The greater degree of genetic variability within the black population means that even a mother-son dyad, or a brother-sister one, is less likely to be a viable match than the same pairing for a white donor-recipient dyad. Thus, purely social explanations like institutional discrimination in hospitals on the one hand, or lack of black familial donors (due to less integrated family structures) on the other, do not alone account for the race gap in waiting time for a kidney; rather, genotype matters.

GENES AND MACRO-LEVEL OUTCOMES

Genetic variation within societies may have important social consequences beyond the health care system. For example, Cook (2015) has shown that populations with immune systems that are genetically diverse have had a health advantage in the premodern period. The idea is that pathogens evolve to target specific immune function weaknesses, and populations with limited genetic diversity (and hence limited diversity in immune response) are at particular risk of infectious pathogens spreading and reducing the health of wide swaths of the population. However, at the population level, genetic diversity can be beneficial in inoculating against such widespread health insults by constraining epidemic spreads of illness.

Indeed, Cook (2015) finds that increases in population-level immune genetic diversity (through the human leukocyte antigen system) lead to increases in country-level life expectancies. He further documents this causal argument by showing that the invention and widespread use of modern vaccinations and other medical technologies has led to a decline in the genetic advantage. That is, modern science and medicine is substituting for natural (genetic) defenses against illnesses at the population level, and in doing so is promoting convergence in life expectancy, and eventually growth and income, across rich and poor countries, yielding another example of interplay between genetics and environments at the population level over historical time.

In earlier times (in disease-rich environments with a lack of medications), genetic variation acted as a buffer against disease, leading to country-level differences in life expectancy based in part on genetic differences. But now that the environment has changed, with new medications and vaccinations, the previous genetic advantages have been largely eliminated, and population genetic factors have important interactions with the larger environment in producing outcomes. These genes only confer advantages in environments that have the ability to foster agriculture. With no cows, goats, or other domesticable mammals, the gene confers no population advantage.

With respect to the importance of milk from livestock, Cook (2014) has shown that the (genetic) ability to digest milk after weaning that appeared early in human history conferred large advantages in population density around 1500 CE (a 10% increase in the beneficial genetic variant in the population was associated with a ~15% increase in population density). Given that other studies have shown that economic development differences in history have been remarkably persistent, the implication is that (relatively) small changes in the genome, at the right time and in the right place (during the Neolithic Revolution in areas able to raise cattle), can lead to large, persistent, and accumulating differences in economic development across countries.

Other recent work has focused on genetic diversity as an overall explanation to growth. Ashraf & Galor (2013) marshal evidence that a Goldilocks level—i.e., not too low and not too high—of genetic diversity within countries might lead to higher incomes and better growth trajectories. The authors discuss the observation that there are many countries with low genetic diversity (e.g., countries predominantly comprised of Native Americans, like present-day Bolivia) as well as populations with high genetic diversity (e.g., many sub-Saharan African countries) that have experienced low economic growth, whereas many countries with an intermediate (Goldilocks) level of diversity (European and Asian countries) have experienced development in the precolonial as

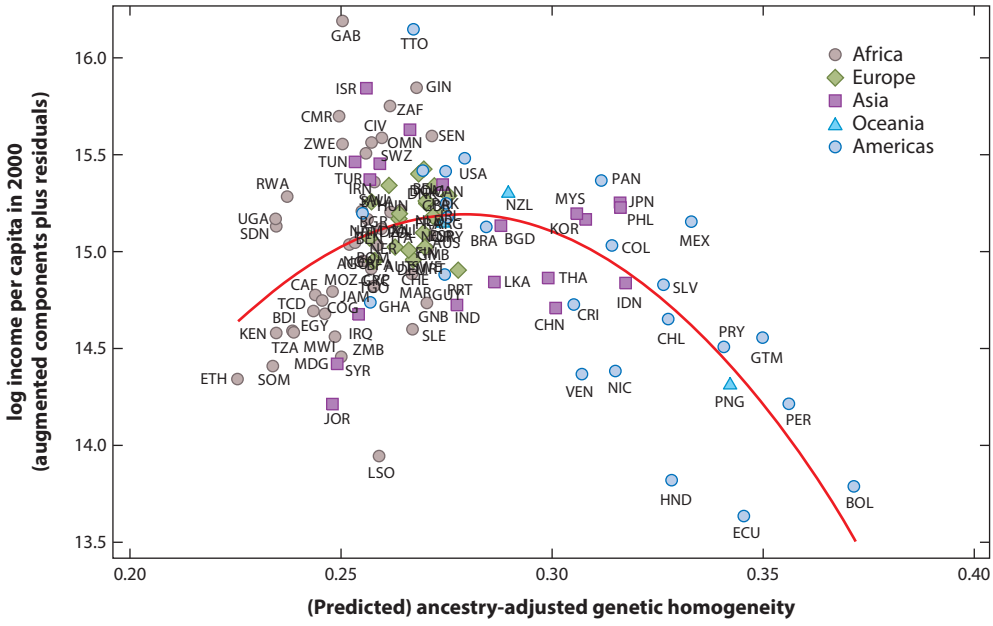

Figure 6

Ancestry-adjusted genetic diversity and country-level income per capita. The three-letter labels are abbreviations of country names. Adapted with permission from Ashraf & Galor (2013, figure 5).

well as the modern era. **Figure 6**, which is adapted from Ashraf & Galor (2013), plots country-level income per capita (on a log scale) versus a measure of country-level genetic diversity. The top of the hump is the Goldilocks level of genetic diversity vis-à-vis economic success.

Some observers have noted how convenient it is that a genetic explanation of country-level differences in development would point to low progress in Africa and high progress in Europe and the United States. Added to this concern is the implication that historical conquest by Europeans would have a positive effect on the development of the native populations by increasing their genetic diversity.

These inflaming descriptions and predictions call for great caution in the current analysis and in future efforts to build on this emerging field of study: Some of the cards seem stacked to repeat some of the past scientifically embarrassing and harmful outcomes of integrating biology/genetics into economic and social explanations of development. But let us consider the key conceptual ideas offered by these articles that begin to be entertained in social science.

The basic theoretical idea (and hypothesis) is that populations/countries face a tradeoff with respect to population-level genetic diversity. Ashraf and Galor argue that higher diversity "enhances society's capability to integrate advanced and more efficient production methods, expanding the economy's production possibility frontier and conferring the benefits of improved productivity" (Ashraf & Galor 2013, p. 3)—an idea motivated by Darwin's theory of evolution by natural selection. These conflicting costs and benefits of genetic diversity lead to the proposition that a middle level, one that optimally trades off the pros and cons of genetic diversity, will lead to the highest growth and development patterns.

Indeed, in an empirical analysis of over 140 countries around the world, Ashraf and Galor find that, for countries that have low genetic diversity (e.g., Bolivia), small increases of 1% in diversity could have increased population density (a measure of economic development) by 58%

by the year 1500; likewise, countries with current high genetic diversity (e.g., Kenya) that reduced this diversity by 1% could have seen an increase in population density of 23%. Fast-forwarding to present-day outcomes, Ashraf and Galor find that a 1% increase in genetic diversity would have raised the income of a homogenous country like Bolivia by 30% by the year 2000, and that reducing genetic diversity by 1% in an already diverse country like Ethiopia would have raised its income by 21%. Obviously, these effects are large.

The claim that diversity has beneficial effects in productive endeavors has a long history in development economics, though the extension of this claim to the benefits of (genetic) diversity has been the subject of very few empirical tests. Here researchers may be at risk of extending, and overextending, the intuition from one domain (the gains in labor productivity of having novel ideas that complement one another) to an expanded domain (the existence of genetic sources of these complementarities). Even if there are findings that could scaffold the two ideas—for example, Alesina & La Ferrara (2005) review the literature that uses measures of ethnic diversity, fragmentation, and heterogeneity to examine aggregate outcomes and find some evidence of both costs and benefits of this type of diversity—like many claims in this nascent literature, the ideas and hypotheses could be true but are as of yet quite untested.

Indeed, new studies would be needed to further show—at the micro, subcountry, and perhaps individual industry, factory, and production team levels—that these effects are detectable and real at units below the country level. New research by Cook & Fletcher (2015) has recently begun to make these extensions. In their article, instead of comparing countries in terms of economic development and genetic diversity, the authors compare (much) smaller populations of high school graduates—indeed, the authors use high school information from a single state (Wisconsin, via the WLS collected in 1957).

By focusing on this single state and on white populations only, the authors can eliminate two specific potential problems with the original Ashraf-Galor country-level analysis: (*a*) the possibility that the countries may differ on other confounding variables that are statistically related to both genetic diversity and economic success but are not measured by the analysts, and (*b*) the fact that race differences in genetic variants may be related both to genetic diversity measures and perhaps to economic performance (through the history of discrimination against specific racial/ethnic groups, for example)—i.e., population stratification. Surprisingly, the authors find a very similar hump-shaped pattern of results linking genetic diversity with wealth measures at the high school level.

Many new studies have sprung up in quick succession extending the Ashraf & Galor (2013) hypothesis on genetic diversity to new areas and new outcomes. Spolaore & Wacziarg (2009), colleagues and coauthors of Ashraf and Galor, have done just that by examining whether population-level (country-level) genetic diversity is related to the likelihood of conflicts and wars with other countries. After all, one of the main mechanisms by which high genetic diversity is hypothesized to hinder growth is through conflict. They examine interstate conflicts and wars between 1816 and 2001 for more than 175 countries and ask whether countries that are less similar in their genetics (i.e., that have higher genetic distance from one another) are more likely to engage in conflicts and wars. The hypothesis is related to much research in economics, political science, conflict studies, and international relations that shows increased conflict among countries and populations that are different from one another—and therefore potentially see "the other" as a potential conquest.

Counter to the general hypothesis by the extant literature on conflict that dissimilar groups are more likely to have conflicts, the new findings suggest that (genetically) similar groups were more likely to have a conflict or war over this time period. The authors take further steps to rule out some obvious counter-explanations for this finding. First, genetically similar populations are likely to live next door to one another, potentially generating wars and conflicts because of this

proximity. The research adjusts the analysis for geographic distance between countries, and the genetic distance still matters. Another potential explanation relies on histories of conquest, trade, and democratization efforts and resulting counterrevolutions and coups. Again, the researchers make adjustments for these (and other) predictors of conflict and war and still see a remaining effect of genetic similarity.

More recently, the same authors have proposed genetic distance between countries as a general measure (a summary statistic) of cultural similarity between nations, taking another important and controversial step toward combining traditional social scientific measures of culture and norms with the explosion of genetic data from around the world. They show that measures of genetic distance across country populations are statistically related to other measures of distance between populations, such as languages, religions, and values as reported in surveys about norms (i.e., having traditional family values or agreeing with notions of gender equality).

CONCLUSION

The interface between the social sciences and genetics has been a growing field over the last decade. At the annual conference held by the University of Colorado at Boulder on "Integrating Genetics and Social Science," attendance has climbed from 27 the first year it was held (in 2009) to 93 in 2015. This review has covered only a small portion of the work that is being produced at this intersection. For example, a growing body of work now uses particular genetic variants that have known effects on phenotypes (such as FTO polymorphisms on body mass index or alcohol dehydrogenase variants on alcohol consumption behavior) as instrumental variables to estimate completely social relationships (e.g., between ADHD and school performance or between body size and wages). This methodology, as I describe elsewhere (Conley 2009), is ill-advised due to unknown pleiotropic effects that violate the instrumental variable/two-stage least-squares exclusion restriction (absent a placebo test in a population in which the $Z \rightarrow X \rightarrow Y$ pathway is environmentally blocked). Likewise, there is much work that I have not reviewed on the social genetics of political behavior and attitudes, on the genetics of personality, on ancestry and gene flow, and on the social regulation of gene expression through epigenetic and other mechanisms.

The appeal of genotype data to social scientists is manifold, but to conclude, it is worth emphasizing the intuition that drives many scholars working in the area. For most of its history, social science has been concerned with average or level effects—be that a regression coefficient on schooling in a wage equation, an early childhood behavioral intervention, or alternatives to incarceration. However, most average treatment effects may mask enormous heterogeneity in elasticity. By applying the prism of $G \times E$ models, it is hoped that the white light of average effects will be refracted into a rainbow of genetically mediated responses that are made clear to the scholar interested in describing human behavior. Likewise, even if a scholar does not care at all about genetic main effects or moderation, controlling for genotype may help reduce standard errors on social variables that researchers typically care about. After all, we cannot continue to ignore mounting evidence that genetic influences explain a third or more of variation in many important social outcomes.

APPENDIX: HUMAN MOLECULAR GENETICS PRIMER FOR THE SOCIAL SCIENTIST

The central dogma of molecular biology is $DNA \rightarrow RNA \rightarrow$ protein. DNA provides the blueprint, which aside from de novo mutations (the sort that sometimes lead to cancer) or mosaicism (when, for example, some of the cells in an individual's body are of a different origin, such as maternal

or fraternal) is identical in every cell of the body. The genome (i.e., the DNA) is stored in the nucleus of each cell—with the exception of red blood cells, which have no nucleus—in 23 pairs of chromosomes as well as in the mitochondria (power plants) of each cell. Mitochondrial DNA (mtDNA) is inherited only from the mother because it arises from the ovum, though there is some debate as to whether some mitochondria from the sperm cell penetrate the egg and survive in the development of the fertilized zygote. The nuclear DNA is inherited from both parents, one of each pair of the 22 autosomal chromosomes coming from each progenitor. As for the sex chromosomes, under typical circumstances, the mother always provides an X (female) chromosome. The father provides an X (making the offspring female) or a Y (making the offspring male). Thus, analysis of mtDNA allows us to peer back through the enate line, whereas Y chromosome analysis allows for characterization of the agnate line.

All in all, if we unfurled the 46 chromosomes and lined them end-to-end, they would be six feet in length, containing three billion base pairs. There are four bases: adenine (A), thymine (T), guanine (G), and cytosine (C). They have specific complementarity so that the double helix, phosphate backbones can be joined only by A-T or by C-G. Among the three billion of these pairings known as alleles, there is variation in about 1% (high-end estimates put this at 4%) of these loci or locations—yielding a figure of three million single base differences and the commonly cited notion that we are 99.9% or 99.7% genetically identical. If we also consider another common form of difference, CNVs, we are an estimated 99.5% similar. Other forms of variation include structural variation in chromosomes such as insertions or deletions (indels). The figures about similarity are rather misleading, because small differences can lead to huge phenotypic differences.

In regions of the genome that encode for messenger RNA (mRNA) that transmits the template for proteins to ribosomes, where proteins are assembled, triplets of bases known as codons specify which amino acid is called for in the assemblage of the protein (which are chains of usually 100 or more amino acids strung together like beads). There are also codons for "start" and "stop." If there is a change of nucleotide in the third position in the codon (say from CTA to CTG), this is usually known as a silent or synonymous mutation, because it does not change the amino acid called for and thus does not affect the protein's makeup, though it can affect the efficiency of production. A change to either of the first two nucleotides in the triplet is called nonsynonymous and leads to a structural change such as an amino acid substitution (missense) or stoppage of transcription (nonsense).

After an mRNA has been transcribed from DNA, it is edited by biochemical machinery that snips out introns and leaves exons that will go on to be translated into proteins. Another change worth noting is that T (thymine) in DNA becomes U (uracil) in RNA. The term gene generally refers to a protein-coding stretch of DNA, including not only the part that gets transcribed but also the promoter region (i.e., the part before the start of the coding region where the transcriptome attaches to begin its work) and other regulatory regions known as enhancers (commonly found within the first intron but also sometimes present thousands of base pairs away) or the 5′ (pronounced five-prime) UTR (untranslated region), which comes after the stop codon.

There are only about 20,000 or so genes (i.e., protein-coding regions) that can each produce about three different proteins on average by alternate splicing or pruning of introns. That figure is much smaller than what most geneticists had expected (for example, rice has about 46,000 genes). This is important because it reveals the importance of gene regulation: That is, because every cell contains the same genetic blueprint, the differences between a neuron, a hepatocyte, and an epithelial cell all derive from which genes are expressed and when. Likewise, differences among humans are largely due not to different protein structures but to the fine-tuning of gene expression at critical points in development. This realization comes hand-in-hand with the recognition that much of the non-protein-coding part of the genome is hardly junk DNA but rather is critical

to conducting this symphony. For example, a form of RNA called micro-RNA (miRNA), often encoded into the 5′ UTR of genes, serves an important role in regulating the process of translation. Other areas of the genome produce not full proteins but peptides, which are short rather than long strings of amino acids and can form a certain class of hormones as well as some neurotransmitters such as endorphins (endogenous opioids).

Variation in gene expression is controlled by a number of factors, some of which are collectively called epigenetics. Epigenetics has become a field of great excitement within the social sciences, possibly because of the notion that it reverses the causal arrow of traditional genetic analysis, pointing it in a direction in which sociologists, for one, feel much more comfortable: from environment to genome. Namely, whereas traditional genetic analysis of behavior examines variations in the nucleotides that are fixed at conception and that have effects that ripple out across the life course, social epigenetics often examines how the environment affects gene expression through processes such as histone acetylation [addition of a $COCH_3$ group to one of the proteins (histone) around which the DNA is coiled when stored] or DNA methylation (the addition of a CH_3 group to a CG sequence) that influence whether or not a particular gene gets turned on or off in a given tissue at a given time. Some scholars are particularly excited by the notion that such environmentally sensitive epigenetic marks may, in fact, be inherited transgenerationally. If so, this would suggest that part of the biological inheritance has environmental roots, and that social factors such as wealth/poverty, incarceration, slavery, family processes, and the like can all be incorporated into the genome. It should be noted, however, that whereas intergenerational associations have been shown in, for example, DNA methylation patterns, other mechanisms have not been ruled out. At the same time, some experimental evidence from animals is providing fodder for theories that some epigenetic marks that are conditioned by stimuli may, in fact, survive in offspring. The evidentiary bar for transgenerational epigenetic memory is rightly set very high, because the current thinking is that the vast majority (if not all) of epigenetic marks are erased during reproduction (meiosis, specifically) to produce an omnipotent stem cell capable of becoming all cell types in the developing embryo (whereas epigenetic marks tend to constrain pathways of development). Meanwhile, there are many other pathways in addition to fixed DNA or epigenetic marks by which information about the environment can be transmitted to offspring. Transgenerational epigenetics promises to be an exciting field for social scientists to watch in the next decade or two, regardless of whether it proves to be a revolution in our understanding of heredity and the nature-nurture dichotomy; at the very least, however, molecular biologists have complicated their central dogma and now recognize many ways in which causal arrows go forward, backward, and loop around the DNA-RNA-protein nexus. Social scientists ignore this genomics revolution at their peril if they seek a complete understanding of human behavior.

DISCLOSURE STATEMENT

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

LITERATURE CITED

Alesina A, La Ferrara E. 2005. Ethnic diversity and economic performance. *J. Econ. Lit.* 43(3):762–800

Ashraf Q, Galor O. 2013. The "out of Africa" hypothesis, human genetic diversity, and comparative economic development. *Am. Econ. Rev.* 103(1):1–46

Bataille V, Snieder H, MacGregor AJ, Sasieni P, Spector TD. 2002. The influence of genetics and environmental factors in the pathogenesis of acne: a twin study of acne in women. *J. Investig. Dermatol.* 119(6):1317–22

Batouli SAH, Trollor JN, Wen W, Sachdev PS. 2014. The heritability of volumes of brain structures and its relationship to age: a review of twin and family studies. *Ageing Res. Rev.* 13:1–9

Belsky DW, Moffitt TE, Baker TB, Biddle AK, Evans JP, et al. 2013b. Polygenic risk and the development progression to heavy, persistent smoking and nicotine dependence: evidence from a 4-decade long longitudinal study. *JAMA Psychiatry* 70(5):534–42

Belsky DW, Moffitt TE, Houts R, Bennett GG, Biddle AK, et al. 2012. Polygenic risk, rapid childhood growth, and the development of obesity: evidence from a 4-decade longitudinal study. *Arch. Pediatr. Adolesc. Med.* 166(6):515–21

Belsky DW, Sears MR, Hancox RJ, Harrington H, Houts R, et al. 2013a. Polygenic risk and the development and course of asthma: an analysis of data from a four-decade longitudinal study. *Lancet Respir. Med.* 1(6):453–61

Benjamin DJ, Cesarini D, Chabris CF, Glaeser EL, Laibson DI. 2012a. The promises and pitfalls of genomics. *Annu. Rev. Econ.* 4:627–62

Benjamin DJ, Cesarini D, van der Loos MJHM, Dawes CT, Koellinger PD, et al. 2012b. The genetic architecture of economic and political preferences. *PNAS* 109(21):8026–31

Björklund AM, Eriksson T, Jäntti M, Raaum O, Österbacka E. 2002. Brother correlations in earnings in Denmark, Finland, Norway and Sweden compared to the United States. *J. Pop. Econ.* 15(4):757–72

Björklund AM, Lindahl M, Plug E. 2006. The origins of intergenerational associations: lessons from Swedish adoption data. *Q. J. Econ.* 121(3):999–1028

Blau PM, Duncan OD. 1967. *The American Occupational Structure*. New York: Free Press

Boardman JD, Domingue BW, Fletcher JM. 2012. How social and genetic factors predict friendship networks. *PNAS* 109(43):17377–81

Cardon LR, Palmer LJ. 2003. Population stratification and spurious allelic association. *Lancet* 361(9357):598–604

Caspi A, McClay J, Moffitt TE, Mill J, Martin J, et al. 2002. Role of genotype in the cycle of violence in maltreated children. *Science* 297(5582):851–54

Caspi A, Sugden K, Moffitt TE, Taylor A, Craig IW, et al. 2003. Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. *Science* 301(5631):386–89

Cavalli-Sforza LL, Menozzi P, Piazza A. 1994. *The History and Geography of Human Genes*. Princeton, NJ: Princeton Univ. Press

Chabris CF, Hebert BM, Benjamin DJ, Beauchamp JP, Cesarini D, et al. 2012. Most reported genetic associations with general intelligence are probably false positives. *Psychol. Sci.* 23(11):1314–23

Chatterjee N, Wheeler B, Sampson J, Hartge P, Chanock SJ, Park JH. 2013. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat. Genet.* 45(4):400–5

Cochran G, Harpending H. 2009. *The 10,000 Year Explosion: How Civilization Accelerated Human Evolution*. New York: Basic Books

Conley D. 2009. The promise and challenges of incorporating genetic data into longitudinal social science surveys and research. *Biodemography Soc. Biol.* 55(2):238–51

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? *Soc. Sci.* 2:82–105

Conley D, Rauscher E. 2013. Genetic interactions with prenatal social environment effects on academic and behavioral outcomes. *J. Health Soc. Behav.* 54(1):109–27

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

Conley D, Siegal ML, Domingue BW, Harris KM, McQueen MB, Boardman JD. 2014. Testing the key assumption of heritability estimates based on genome-wide genetic relatedness. *J. Hum. Genet.* 59(6):342–45

Cook CJ. 2014. The role of lactase persistence in precolonial development. *J. Econ. Growth* 19(4):369–406

Cook CJ. 2015. The natural selection of infectious disease resistance and its effect on contemporary health. *Rev. Econ. Stat.* 97(4):742–57

Cook CJ, Fletcher JM. 2013. *Understanding heterogeneity in the effects of birth weight on adult outcomes*. Work. Pap. 20895, Natl. Bur. Econ. Res., Cambridge, MA

Cook CJ, Fletcher JM. 2015. *Genetic diversity and economic outcomes: a novel test of a controversial hypothesis*. Work. Pap., Univ. Wisconsin–Madison
Corcoran M, Gordon R, Laren D, Solon G. 1992. The association between men's economic status and their family and community origins. *J. Hum. Resour.* 27(4):575–601
Daetwyler HD, Villanueva B, Woolliams JA. 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLOS ONE* 3(10):e3395
Daw J. 2015. Explaining the persistence of health disparities: social stratification and the efficiency-equity trade-off in the kidney transplantation system. *Am. J. Sociol.* 120(6):1595–640
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
Dudbridge F. 2013. Power and predictive accuracy of polygenic risk scores. *PLOS Genet.* 9(3):e1003348
Fletcher JM. 2012. Why have tobacco control policies stalled? Using genetic moderation to examine policy impacts. *PLOS ONE* 7(12):e50576
Fowler JH, Settle JE, Christakis NA. 2011. Correlated genotypes in friendship networks. *PNAS* 108(5):1993–97
Goldberger AS. 1979. Heritability. *Economica* 46(184):327–47
Gould SJ. 2000. The spice of life: an interview with Stephen Jay Gould. *Lead. Lead.* 15:14–19
Guo G, Fu Y, Lee H, Cai T, Harris KM, Li Y. 2014. Genetic bio-ancestry and social construction of racial classification in social surveys in the contemporary United States. *Demography* 51(1):141–72
Guo G, Stearns E. 2002. The social influences on the realization of genetic potential for intellectual development. *Soc. Forces* 80(3):881–910
Hamer D, Sirota L. 2000. Beware the chopsticks gene. *Mol. Psychiatry* 5(1):11–13
Hauser RM, Sewell WH. 1986. Family effects in simple models of education, occupational status, and earnings: findings from the Wisconsin and Kalamazoo studies. *J. Labor Econ.* 4(3):83–115
Hauser RM, Sheridan JT, Warren JR. 1999. Socioeconomic achievements of siblings in the life course: new findings from the Wisconsin longitudinal study. *Res. Aging* 21(2):338–78
Heath AC, Berg K, Eaves LJ, Solaas MH, Corey LA, et al. 1985. Education policy and the heritability of educational attainment. *Nature* 314:734–36
Hewitt JK. 2012. Editorial policy on candidate gene association and candidate gene-by-environment interaction studies of complex traits. *Behav. Genet.* 42(1):1–2
Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, et al. 2010. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* 42(4):348–54
Karg K, Burmeister M, Shedden K, Sen S. 2011. The serotonin transporter promoter variant (5-HTTLPR), stress, and depression meta-analysis revisited: evidence of genetic moderation. *Arch. Gen. Psychiatry* 68(5):444–54
Kuo HD, Hauser RM. 1995. Trends in family effects on the education of black and white brothers. *Soc. Educ.* 68(2):136–60
Lee PH, O'Dushlaine C, Thomas B, Purcell SM. 2012. INRICH: interval-based enrichment analysis for genome-wide association studies. *BMC Bioinformatics* 28(13):1797–99
McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, et al. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* 9(5):356–69
Milot E, Mayer FM, Nussey DH, Boisvert M, Pelletier F, Ráele D. 2011. Evidence for evolution in response to natural selection in a contemporary human population. *PNAS* 108(41):17040–45
Nielsen F. 2006. Achievement and ascription in educational attainment: genetic and environmental influences on adolescent schooling. *Soc. Forces* 85(1):193–216
Nielsen F. 2008. The nature of social reproduction: two paradigms of social mobility. *Sociologica* 3:1–35
Ollier W, Sprosen T, Peakman T. 2005. UK biobank: from concept to reality. *Pharmacogenomics* 6(6):639–46
Olneck MR. 1976. *On the use of sibling data to estimate the effects of family background, cognitive skills, and schooling: results from the Kalamazoo brothers study*. Work. Pap., Inst. Res. Poverty, Univ. Wisconsin-Madison, Madison, WI

Page MR, Solon G. 2003. Correlations between brothers and neighboring boys in their adult earnings: the importance of being urban. *J. Labor Econ.* 21(4):831–55

Peyrot WJ, Lee SH, Milaneschi Y, Abdellaoui A, Byrne EM, et al. 2015. The association between lower educational attainment and depression owing to shared genetic effects? Results in ~25,000 subjects. *Mol. Psychiatry* 20(6):735–43

Platt A, Pullum S, Lewis D, Hall A, Ollier W. 2010. The UK DNA banking network: a "fair access" biobank. *Cell Tissue Bank.* 11(3):241–51

Plomin R. 2009. The nature of nurture. In *Experience and Development: A Festschrift in Honor of Sandra Wood Scarr*, ed. K McCartney, RA Weinberg, pp. 61–80. New York: Taylor & Francis

Plomin R, Owen MJ, McGuffin P. 1994. The genetic basis of complex human behaviors. *Science* 264(5166):1733–39

Plug E. 2004. Estimating the effect of mother's schooling on children's schooling using a sample of adoptees. *Am. Econ. Rev.* 94(1):358–68

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38(8):904–9

Purcell SM, Wray NM, Stone JL, Visscher PM, O'Donovan PM, et al. 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 460(7256):748–52

Rietveld CA, Hessels J, van der Zwan P. 2015. The stature of the self-employed and its relation with earnings and satisfaction. *Econ. Hum. Biol.* 17:59–74

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

Risch N, Herrell R, Lehner T, Liang KY, Eaves L, et al. 2009. Interaction between the serotonin transporter gene (*5-HTTLPR*), stressful life events, and risk of depression: a meta-analysis. *JAMA* 301(23):2462–71

Rønningen KS, Paltiel L, Meltzer HM, Nordhagen R, Lie KL, et al. 2006. The biobank of the Norwegian mother and child cohort study: a resource for the next 100 years. *Eur. J. Epidemiol.* 21(8):619–25

Sacerdote B. 2007. How large are the effects from changes in family environment? A study of Korean American adoptees. *Q. J. Econ.* 122(1):119–57

Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J, et al. 2014. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* 507:354–57

Scarr S, Carter-Saltzman L. 1979. Twin method: defense of a critical assumption. *Behav. Genet.* 9(6):527–42

Schmitz L, Conley D. 2015. The long-term consequences of Vietnam-era conscription and genotype on smoking behavior and health. *Behav. Genet.* 46:43–58

Schwartz CR, Han H. 2014. The reversal of the gender gap in education and trends in marital dissolution. *Am. Sociol. Rev.* 79(4):605–29

Sieradzka D, Power RA, Freeman D, Cardno AG, Dudbrige F, Ronald A. 2015. Heritability of individual psychotic experiences captured by common genetic variants in a community sample of adolescents. *Behav. Genet.* 45(5):493–502

Silventoinen K, Rokholm B, Kaprio J, Sørensen TIA. 2010. The genetic and environmental influences on childhood obesity: a systematic review of twin and adoption studies. *Int. J. Obes.* 34(1):29–40

Spolaore E, Wacziarg R. 2009. *War and relatedness*. Work. Pap. 15095, Natl. Bur. Econ. Res., Cambridge, MA

Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, et al. 2009. The genetic structure and history of Africans and African Americans. *Science* 324(5930):1035–44

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

Visscher PM, Yang J, Goddard ME. 2010. A commentary on "Common SNPs explain a large proportion of the heritability for human height" by Yang et al. 2010. *Twin Res. Hum. Genet.* 13(6):517–24

Wade N. 2014. *A Troublesome Inheritance: Genes, Race and Human History*. London: Penguin

Warren JR, Hauser RM. 1997. Social stratification across three generations: new evidence from the Wisconsin longitudinal study. *Am. Sociol. Rev.* 62(4):561–72

Warren JR, Hauser RM, Sheridan JT. 2002. Occupational stratification across the life course: evidence from the Wisconsin longitudinal study. *Am. Sociol. Rev.* 67(3):432–55

Yang J, Banyamin B, McEvoy BP, Gordon S, Henders AK, et al. 2010. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* 42(7):565–69

Zuidhof MJ, Schneider BL, Carney VL, Korver DR, Robinson FE. 2014. Growth, efficiency, and yield of commercial broilers from 1957, 1978, and 2005. *Poultry Sci.* 93(12):2970–82