Biodemography and Social Biology, 55:238–251, 2009

ISSN: 1948-5565 print / 1948-5573 online
DOI: 10.1080/19485560903415807

Routledge
Taylor & Francis Group

The Promise and Challenges of Incorporating Genetic Data into Longitudinal Social Science Surveys and Research

DALTON CONLEY

New York University, New York, New York

In this paper, I argue that social science and genomics can be integrated; however, the way this marriage is currently occurring rests on spurious methods and assumptions and, as a result, will yield few lasting insights. However, recent advances in both econometrics and in developmental genomics provide scientists with a novel opportunity to understand how genes and environment interact to produce social outcomes. Key to any causal inference about the interplay between genes and social environment is that either genotype be exogenously manipulated (i.e. through sibling fixed effects) while environmental conditions are held constant, and/or that environmental variation is exogenous in nature, i.e. experimental or arising from a natural experiment of sorts. Further, initial allele selection should be motivated by findings from genetic experiments in model animal studies linked to orthologous human genes. Likewise, genetic associations found in human population studies should then be tested through knock-out and over-expression studies in model organisms.

Introduction

Studying genetic-environmental (GE) interactions has long been a goal of social scientists fond of stressing the dependence of genetic expression on social structure. However, how do we get from the adage that "a gene for aggression lands you in prison if you're from the ghetto, but in the boardroom if you're to the manor born" to a serious empirical research program on the study of GE interactions? Even if we are only interested in "pure" environmental or genetic effects, how does one deal empirically with the fact that environmental conditions may affect gene expression and that genes may influence environments? This article will address best practices for integrating genetic markers into social science research with a particular emphasis on what social scientists can draw from existing genetic models in non-human organisms and on what biologists can learn from social scientists.

Gene markers in nationally representative social surveys can be deployed for at least three important uses: assessing the direct impact of specific genetic influences on socioeconomic and behavioral outcomes; modeling genetic-environmental interactions; and tracing genealogies across time and space. In this article, I will address only the first two issues given that the last is a topic that has been thoroughly addressed by the human genetics literature elsewhere (for an example using Icelandic data, please see Price et al. 2009). I conclude the article with a discussion of some of the data quality and sensitivity concerns that surround DNA collection and analysis. Further, it is important to note that because this article is being written for a social scientific audience primarily, I will focus on

Address correspondence to Dalton Conley, New York University, 6 Washington Square North, Room 20, New York, NY 10003. E-mail: conley@nyu.edu

explanations of genetics at the expense of survey or sampling methodology. However, the flow of knowledge across these disciplines should not be one-way. As we shall see, coming out of a medical science tradition of case-control studies on small samples, human geneticists have generally paid insufficient attention to issues of sampling, measurement of behavioral outcomes, and survey design. The findings from human geneticists would be strengthened if adequate attention were paid to the expertise developed by in survey methodology that has been developed in the social sciences as well as the techniques of inference from observational data pioneered most notably by applied econometricians. Though the current article will not go into depth regarding sampling and survey design, I will discuss econometric techniques as they can be applied to genomic data, which should be of utility to those in the biological sciences doing observational studies (as opposed to experiments).

Recently there has been intense interest in collecting biomarkers, in general, and genetic data in particular, among social scientists—particularly those conducting panel studies. Within the United States, the National Longitudinal Survey of Adolescent Health (Add Health) has been a pioneer in the collection of biological data, including DNA markers from a sample of monozygotic and dizygotic twins. Add Health has collected markers that indicate zygosity as well as information on alleles for seven putatively important genes, six on autosomal chromosomes (ones for which each individual has two copies) and one marker on the X-chromosome. Investigators associated with the study plan to expand this number greatly in future waves. In 2006, along with other biomarkers such as levels of high-density lipoprotein cholesterol in blood, the Health and Retirement Survey began collecting saliva samples to extract DNA for sequencing and analysis. The Wisconsin Longitudinal Survey is also collecting DNA samples, and the Panel Study of Income Dynamics (PSID) is considering adding such a module as well. This last possibility is particularly promising because none of the earlier studies have the degree of intergenerationality of the PSID, nor are they nationally representative samples of the entire adult population across the age spectrum. In many ways, the United States is a laggard in collecting such data—possibly due to the increased salience of privacy concerns as compared to other societies. Iceland's Decode project has DNA data on almost the entire citizen population. The United Kingdom has launched an ambitious study that will attempt to collect genetic data on 500,000 respondents. And the Scandinavian countries already have genetic samples that can be linked to rich administrative datasets.

Genetic Effects on Behavior

For a long time, modeling the effect of genes on social outcomes among human populations was the province of behavioral geneticists who relied on adoption and twin studies as indicators of unmeasured genetic background. These methods often rested on a number of critical assumptions that have been challenged elsewhere (see, e.g., Goldberger 1979). More recently, genetic markers on specific loci—such as single-nucleotide polymorphisms (SNPs)—have seemed to offer hope for those interested in an explicit research program aimed at specifying and measuring gene-specific effects for complex traits such as behavioral phenotypes (what geneticists call *quantitative traits*). Polymorphisms are genetic variants that occur in at least 1 percent of the population. They could include base-pair substitutions that may affect the amino acid produced out of that codon if the polymorphism is in an open reading frame (ORF) of a gene (i.e., the protein-related coding region) and is nonsynonymous; they may truncate the protein by causing the transcription machinery to stop there (by producing a stop codon); or they may do nothing (what are called silent or synonymous mutations) because multiple three-letter codes may result in the same

amino acid being produced (though, perhaps at different efficiency levels, something called *codon bias*). Hence, these nonlethal polymorphisms, which result from mutations, may present an opportunity to study how specific alleles may result in different outcomes.

The basic logic is the following: A certain proportion of a population sample is found to have a variant of a particular allele. *If* this allele is shown to be randomly distributed across demographic subgroups (or, for example, *within* a particular subgroup such as ethnic group), and, likewise, it is found to be associated with a specific social outcome or tendency (such as addictiveness, shyness, or schizophrenia, to name a few) *within* that same population (or subgroup as the case may be), then researchers may try to look for specific outcomes that seem to covary with the presence or absence of that particular allele. This has been the approach of most work to date in both the social and biological sciences that have worked with population (nonexperimental) data. However, the immediate problem is that alleles are not necessarily distributed randomly across subpopulations, thus potentially biasing the observed phenotypic associations with those alleles.

The location of the genetic effect in specific places on the genome is seen as a key step forward from earlier behavioral genetics (BG) research. (Recent models also allow for genetic dominance—that is, nonlinear interactions between alleles.) However, because the object of study is typically just one allele, such analysis tells us little about the overall genetic heritability of an outcome.

To complicate matters even further, absent genetic experiments that knock out or overexpress specific genes, we can never be sure that the allele in question is what is causing any observed effect (irrespective of environmental interactions) thanks to the possibility of genetic linkage mentioned above. Namely, genes are "shuffled" across the chromosomes of a parent during the recombination period of meiosis. (Meiosis results in the formation of the 1N gamete—i.e., the sperm or egg.) However, two alleles are more likely to stay paired together in a given gamete the closer they are to each other on the chromosome—hence the term *linkage* or *linkage disequilibrium*—because they are more likely to be found on the same pieces of DNA that are exchanged. A helpful analogy is the shuffling of a deck of cards: It is more likely that cards right next to each other will not get separated in the shuffling process than it is for cards separated by a longer "distance." So even when we know that a given gene is associated with a quantitative trait, we cannot be 100 percent sure (absent genetic experiments on non-humans) that said gene is causally responsible. The best we can say is that that area of the genome is associated with the phenotype under study. If we allow for different degrees of genetic linkage of particular genes with other genes by population, then we cannot even plausibly say (for sure) that a given gene is responsible for the outcome in two different populations even if we observe the same marker-phenotype association (never mind GE interactions). And indeed, microsatellites (groups of genetically linked genes) have been shown to vary across conventionally defined population groups (such as our folk-racial categories). Further, the lengths of microsatellite repeats (also known as *simple sequence repeats*) of DNA base pair motifs are, in fact, one way that geneticists identify human population origins because such repeats are frequently occurring (i.e., the DNA replication machinery makes this sort of coding error more frequently than other types) and because these repeats do not appear to be under any selective pressure (at least about which we know; however, some recent work on dogs suggests that these repeats may, in fact, face selection pressures, particularly when they occur in a coding region). For the most part, these repeated sequences appear to be junk DNA in noncoding regions that produce neither protein products nor peptides nor other important forms of RNA such as micro-RNAs, transfer RNAs, or ribosomal RNAs. However, they may influence the degree to which important parts of the genome (such as genes themselves) are separated—and thus

linked or delinked—during recombination. (As a side note, this also means that the assumption of a *complete* lack of selective pressure on such microsatellite repeats may be incorrect to the extent that these repeats fall between genes [or other important DNA products] that interact with each other in functionally important ways.)

The real rub is that, because we can plausibly postulate second-, third-, fourth-, and, ultimately, *N*th-order interactions across alleles, there simply would not be enough degrees of freedom in the approximately 7 billion human beings currently living to properly test a fully specified model (21,000! = 9.58 E 81648 > 7,000,000,000). The discovery of about 21,000 genes—a figure much lower than originally hypothesized—is good news in that it is a tractable number of alleles for geneticists to study. However, the irony lies in the fact that, if this lowly number of genes explains the development of human beings in all their glorious forms, then gene-gene interactions are probably quite important. There has also been a recent explosion of discoveries relating to the important role of microRNAs in affecting how messenger RNAs are spliced (and therefore can produce multiple products) and whether or not they get translated at all (as well as increased interest in other nonprotein products of DNA once considered "junk").

Of course, in order to mitigate the possibility of admixture or linkage confounding the results, one should begin such a project with a theory about why the expression of a given gene (that is, the causal pathway from gene to protein to outcome) would covary with an important socioeconomic outcome (such as education or income)—rather than just going in with a fishing net to troll for associations. The way to accomplish this is through reference to experimental literature based on animals. For example, Add Health, as mentioned above, has collected and sequenced DNA from a twin subsample of its respondents. The gene regions that were analyzed were picked based on results from genetic experiments among mammals. Take the monoamine oxidase A (MAOA) gene, for instance. Cases et al. (1995) and Shih and Thompson (1999) studied knockout mice (those with the MAOA gene removed) and found that they had increased dopamine, serotonin, and norepinephrine levels and increased aggression among males. Likewise, some polymorphisms in the dopamine receptor allele DRD4 have been linked to attention deficit hyperactivity disorder (ADHD) in humans through associational study (Brookes et al. 2006) and by virtue of experimental studies in animals. DRD4 is a g-coupled protein receptor that forms part of a signaling pathway in neurons in certain brain circuits responsible for pleasure. (The activated conformation of the receptor inhibits the activity of the enzyme adenylyl cyclase, which, in turn, lowers the concentration of cyclic AMP, an important intracellular signaling molecule.) Finally, Murphy et al. (2001) studied mice with a disrupted 5-HTT (serotonin transporter) gene and found that those with risky alleles were more fearful and had higher stress hormone levels in response to stress but no differences by genotype without environmental stress. Research on rhesus macaques found different biological reactions depending on 5-HTT genotype for those raised in stressful environments but no differences among those raised normally (Bennett et al. 2002). Krishnan and Nestler (2007) conducted a study on inbred (genetically identical) mice in a carefully controlled environment to try to explain differences in resilience to stressful life events. They found differences in stress-induced outcomes even controlling environment. Previous studies attributed such outcomes to environmental or early development differences (e.g., Wong et al., 2005; Peaston and Whitelaw, 2006).

Though these studies all relied on genetic experiments in animals for guidance, a number of counterexamples can be found where associational fishing expeditions have led to more tenuous findings that have not withstood the rigors of replication. One notable example can be found in the so-called gay gene. In 1993, Hamer, Hu, Magnuson, Hu, and Pattatucci published an article in *Science* showing an association between a microsatellite

on the X-chromosome (called Xq28) and homosexuality in men. The conclusion rested on the greater propensity of gay brothers to share genetic markers at this locus as well as pedigree analysis that showed a greater likelihood of gay men to have other gay male relatives on their maternal side (because the X that males receive always comes from their mother). Later work (see Rice, Anderson, Risch, and Ebers 1999) failed to replicate the findings among a similar sample of Canadian brothers and a heated debate ensued. Hamer et al.'s study is among the better of the associational studies given its pedigree-based analysis, but like many others in the field it relies on a small, nonrepresentative sample and purports to explain a complicated phenotype: *stated* sexual orientation. I emphasize *stated* for a reason: Even if the results could be routinely replicated, it may be the case that the Xq28 locus is associated with willingness to reveal homosexuality in a survey rather than to homosexuality itself, given its sometimes stigmatizing status in North American culture.

Multiple hypothesis testing—with so many potential genetic loci of study—is of major concern here. Luckily, biologists have elaborated on the Bonferroni correction to produce a series of ways to approach the problem of false positives (see, e.g., Thornton and Jensen 2007). However, as the cost of sequencing continues to drop, the temptation for social (and biological) scientists will be to conduct genome-wide association studies (GWAS) with little regard for theory and experimental evidence about target genes of interest. The result, I fear, will be many association studies for complex quantitative traits that result from a mix of environmental and genetic influences and interactions. Such analyses will inevitably produce a number of false positives that survive even the most conservative false discovery rate threshold and which, in turn, will send researchers down many fruitless paths.

However, if allele variation of experimentally verified gene loci is studied within families (i.e., across siblings or conditional on parents' genotype), then such markers measured in social surveys do indeed offer a potential way to measure specific genetic influences with some certainty (Allison 1997; and Allison, Mooseong, Kaplan, and Martin 1999 offer mixed models for such analysis). One would then compare the expression of that allele—compared to the sibling without the polymorphism, for example—using fixed or random effects. One illustration of this approach is provided by Ding et al. (2006), who use sibling fixed effects to identify "random" genetic variation within families and thereby hold constant parental genotype as well as shared environment.

Another example is provided by Fletcher et al. (2008), who, rightly, examined animal-identified genes—the same ones Caspi et al. (2002; 2003) used plus two dopamine receptor alleles DRD4 and DRD2 and the dopamine transporter gene DAT1. Indeed, they do find effects of some of the genes of interest on behavioral phenotypes (such as depression and attention-deficit hyperactivity disorder [ADHD] as well as obesity); however, the authors then pushed the data too far. They asserted that these randomized genes can be used as instrumental variables (Z) in order to predict such behavioral outcomes (X) and, in turn, instrumented behavior (X*) can be used to generate unbiased estimates of the effects of child behavior on schooling outcomes. Of course, though the genes-as-instruments meet the first qualification of a valid instrument—that Z predicts X strongly enough (otherwise known as the weak instrument test)—they fail the second requirement, the exclusion restriction (namely, that Z has no effect on Y net of X). In other words, for genes to be used as instrumental variables (IVs), they must not only be randomized within a population (such as between nonidentical twin siblings), they must have no other effect on the ultimate outcome of interest other than through their causal impact on the intermediary phenotype measured. Does DRD4 only affect school performance through the pathway of diagnosed ADHD? Of course not. ADHD is a complicated syndrome that involves lots of measurement error and thus most likely reflects a whole host of other unmeasured traits.

And even if ADHD were measured perfectly by the researchers, there may be other effects of the genes in question on educational outcomes through any number of mechanisms, thereby violating the exclusion restriction necessary for unbiased IV estimation.

By way of example, a quick check on the National Center for Biotechnology Information Expressed Sequence Tags database shows that DRD4 is also expressed in the kidney, in various components of the eye (including the lens), and in ovarian tumors. This list, of course, is incomplete and will inevitably grow longer the more the gene is studied. So any claim that polymorphisms in this gene are causing ADHD through a specific pathway in the brain may be called into question. Similarly, any second-order effects of DRD4-related ADHD on other outcomes (as well as putative environmental interactions with DRD4 or gene-gene interactions) may be underidentified due to the fact that DRD4 may be having other, unmeasured effects on phenotype through its actions in the kidneys or eyes (or elsewhere). In other words, if DRD4 were observed to lead to ADHD and, in turn, if ADHD were associated with poor academic outcomes only in students who are in classrooms with more than 25 students, we could not be sure whether it was genetically caused ADHD that was interacting with class size through a brain-behavior mechanism or whether larger classes merely placed these students further from the blackboard and that a DRD4 effect on eyesight was responsible for poorer academic performance. Or worse yet, perhaps the associated ADHD diagnosis itself was attributable to the eyesight effect that led to a lack of concentration in school. In other words, by virtue of its multiple occasions of expression, multiple causal pathways are possible, throwing into jeopardy a social scientist's claims.

Conversely, running DRD4 through a web-interface that searches for similar genes yields a total of 172 homologous genes in the human species alone, including dopamine receptors (see Figure 1). Obtaining such results suggests—but by no means proves—that DRD4 might be "redundant" in the human body. Indeed, its change through mutation to a

gi|4503389 dopamine receptor D4; D(2C) dopamine receptor [H... 832 0.0 gi|4503387 dopamine receptor D3 isoforma [Homo sapiens] 256 3e-68 gi|17986270 dopamine receptor D2 isoformshort [Homo sapiens] 243 1e-64 gi|4503385 dopamine receptor D2 isoformlong [Homo sapiens] 234 6e-62 gi|15718670 alpha-2A-adrenergic receptor; alpha-2AAR subtyp... 207 1e-53 gi|4501967 alpha-2C-adrenergic receptor; alpha2-AR-C4 [Homo... 207 1e-53 gi|4501965 alpha-2B-adrenergic receptor; alpha-2-adrenergic... 206 3e-53 gi|4504531 5-hydroxytryptamine (serotonin) receptor 1A [Hom... 188 5e-48 gi|16445396 dopamine receptor D3 isoforme [Homo sapiens] 173 2e-43 gi|16445402 dopamine receptor D3 isoforme [Homo sapiens] 173 2e-43 gi|4557265 beta-1-adrenergic receptor [Homo sapiens] 169 3e-42 gi|4557267 adrenergic, beta-3-, receptor [Homo sapiens] 166 2e-41 gi|4504539 5-hydroxytryptamine (serotonin) receptor 2B [Hom... 156 3e-38 gi|6005782 histamine receptor H3; G protein-coupled recepto... 151 7e-37 gi|16445396 dopamine receptor D3 isoformb [Homo sapiens] 150 1e-36 gi|16445400 dopamine receptor D3 isoformd [Homo sapiens] 147 1e-35 gi|4502821 cholinergic receptor, muscarinic4; muscarinica... 145 7e-35 gi|4504533 5-hydroxytryptamine (serotonin) receptor 1B; 5-H... 145 7e-35 gi|4502815 cholinergic receptor, muscarinic1; muscarinica... 142 4e-34 gi|4501957 alpha-1D-adrenergic receptor; adrenergic, alpha-... 135 4e-32 gi|10880131 5-hydroxytryptamine receptor 7 isoformb; serot... 129 3e-30 gi|10880129 5-hydroxytryptamine receptor 7 isoformd; serot... 129 3e-30 gi|4504547 5-hydroxytryptamine receptor 7 isoforma; seroto... 129 3e-30 gi|10835175 5-hydroxytryptamine (serotonin) receptor 2A [Ho... 129 5e-30 gi|4502817 cholinergic receptor, muscarinic2; muscarinica... 127 1e-29 gi|4501959 alpha-1B-adrenergic receptor; adrenergic, alpha ... 126 3e-29 gi|15451761 alpha-1A-adrenergic receptor isoform4; adrener... 123 2e-28 gi|15451759 alpha-1A-adrenergic receptor isoform2; adrener... 123 2e-28 gi|15451757 alpha-1A-adrenergic receptor isoform3; adrener... 123 2e-28 gi|4501961 alpha-1A-adrenergic receptor isoform1; adrenerg... 123 2e-28 gi|4504541 5-hydroxytryptamine (serotonin) receptor 2C [Hom... 120 1e-27 gi|29727439 similar to 5-HT5B serotonin receptor [Homo sapi... 118 7e-27 gi|4503383 dopamine receptor D1 [Homo sapiens] 114 1e-25 gi|30150689 similar to D(1B) dopamine receptor (D(5) dopami... 114 1e-25 gi|4504545 5-hydroxytryptamine (serotonin) receptor 6 [Homo... 111 1e-24 gi|4503391 dopamine receptor D5; dopamine receptor D1B; D1b... 110 1e-24 gi|13236497 5-hydroxytryptamine (serotonin) receptor 5A [Ho... 110 2e-24 gi|4504537 5-hydroxytryptamine (serotonin) receptor 1E [Hom... 100 2e-21 gi|30148127 similar to D(1B) dopamine receptor (D(5) dopami... 99 7e-21 gi|4504535 5-hydroxytryptamine (serotonin) receptor 1D [Hom... 99 7e-21 gi|30148324 similar to D(1B) dopamine receptor (D(5) dopami... 97 2e-20 gi|13435405 histamine receptor H2; gastric receptor 1 [Homo... 96 5e-20 gi|10835197 5-hydroxytryptamine (serotonin) receptor 1F; 5-... 96 6e-20 gi|4502819 cholinergic receptor, muscarinic3; muscarinica... 95 8e-20 gi|11321563 5-hydroxytryptamine (serotonin) receptor 4 [Hom... 94 2e-19 gi|17484441 similar to Alpha-1D adrenergic receptor (Alpha ... 92 5e-19 gi|7108336 cholinergic receptor, muscarinic5; muscarinica... 91 1e-18 gi|4501969 adrenergic, beta-2-, receptor, surface; beta-2 a... 89 4e-18 gi|4505925 putative neurotransmitter receptor [Homo sapiens] 88 1e-17 gi|21361153 cholecystokininB receptor [Homo sapiens] 86 6e-17 gi|19923823 neuromedinU receptor 2 [Homo sapiens] 86 6e-17 gi|28173558 trace amine receptor 4 [Homo sapiens] 85 8e-17 gi|4504491 histamine receptor H1; histamine receptor, subcl... 85 8e-17 gi|24432089 G protein-coupled receptor 66 [Homo sapiens] 85 8e-17 gi|28875799 cholecystokininB receptor [Homo sapiens] 85 1e-16 gi|4557637 orexinreceptor 1 [Homo sapiens] 83 3e-16 gi|4501951 adenosine A2b receptor [Homo sapiens] 82 7e-16 gi|28173562 trace amine receptor 3 [Homo sapiens] 82 7e-16 gi|21264324 trace amine receptor 1 [Homo sapiens] 82 7e-16 gi|5921992 adenosine A2a receptor; adenosine A2 receptor; a... 82 9e-16 gi|4557865 somatostatinreceptor 5 [Homo sapiens] 81 1e-15 gi|9506747 G protein-coupled receptor 27; super conserved r... 79 5e-15 gi|4501947 adenosine A1 receptor [Homo sapiens] 78 1e-14 gi|5453668 G protein-coupled receptor 19 [Homo sapiens] 77 2e-14 gi|4557863 somatostatinreceptor 4 [Homo sapiens] 75 7e-14 gi|16751917 G protein-coupled receptor 102; trace amine rec... 75 8e-14 gi|16876435 G protein-coupled receptor 101 [Homo sapiens] 73 4e-13 gi|4503903 galaninreceptor 1; Galanin receptor [Homo sapiens] 72 6e-13 gi|4557857 somatostatinreceptor 1 [Homo sapiens] 72 6e-13 gi|4503907 galaninreceptor 3; galaninreceptor, family mem... 72 7e-13 gi|4557861 somatostatinreceptor 3 [Homo sapiens] 72 7e-13 gi|11545887 neuropeptideFF 1; RFamide-related peptide rece... 72 7e-13 gi|4501953 adenosine A3 receptor [Homo sapiens] 72 9e-13 gi|4758468 G protein-coupled receptor 50 [Homo sapiens] 71 1e-12 gi|14211847 G protein-coupled receptor 54; G protein-couple... 71 1e-12 gi|4758474 G protein-coupled receptor 10; prolactin-releasi... 71 1e-12 gi|4557639 orexinreceptor 2 [Homo sapiens] 71 1e-12 gi|4505513 opiate receptor-like 1; opioidreceptor-like 1; ... 70 2e-12 gi|7657142 G protein-coupled receptor 58 [Homo sapiens] 70 2e-12 gi|4503905 galanin receptor 2 [Homo sapiens] 70 3e-12 gi|4503639 coagulation factor II (thrombin) receptor-like 3... 69 6e-12 gi|15150803 opsin4 (melanopsin); melanopsin [Homo sapiens] 69 6e-12 gi|14251205 histamine H4 receptor [Homo sapiens] 68 1e-11 gi|29729009 similar to seven transmembranehelix receptor [... 68 1e-11 gi|27734717 opioidreceptor, delta 1 [Homo sapiens] 68 1e-11 gi|5174593 melatonin receptor 1A; melatonin receptor type 1... 68 1e-11 gi|30151598 similar to seven transmembranehelix receptor [... 68 1e-11 gi|17437027 similar to seven transmembranehelix receptor [... 68 1e-11 gi|4758434 growth hormone secretagoguereceptor [Homo sapiens] 67 2e-11 gi|4505515 opioidreceptor, mu 1 [Homo sapiens] 66 5e-11 gi|9507143 super conserved receptor expressed in brain 3 [H... 65 9e-11 gi|7705316 CC chemokinereceptor 10 [Homo sapiens] 64 2e-10 gi|4557859 somatostatinreceptor 2 [Homo sapiens] 64 2e-10 gi|14210484 G protein-coupled receptor slt; melanin-concent... 64 2e-10 gi|16804258 G protein-coupled receptor 74; neuropeptideG p... 64 2e-10 gi|4758820 G protein-coupled receptor 74; neuropeptideG pr... 64 2e-10 gi|4505447 neuropeptideY receptor Y2 [Homo sapiens] 64 2e-10 gi|24476016 G-protein coupled receptor [Homo sapiens] 64 3e-10 gi|6677701 G-protein coupled receptor [Homo sapiens] 64 3e-10 gi|5174595 melatonin receptor 1B; melatonin receptor type 1... 63 3e-10

Figure 1. Homologous genes of DRD4 in *Homo sapiens*.
Source: Jeong, Msason, Barabási, and Oltvai (2001).

shorter version is nonlethal (hence our ability to study human variation). Perhaps one of these other 172 genes is overexpressed to compensate for a deficient DRD4 allele. Thus we face a similar problem of inference as with the expression data: Is any observed association with quantitative traits the direct result of DRD4 changes or the indirect affects of "compensation" in other parts of the genetic network?

Gene-Environment Interactions

Despite the formidable complications described above, I argue that it is, in fact, possible to obtain empirically robust estimates of genetic environmental interaction effects. However, the strategy needed to parameterize such effects relies on the proper estimation of truly exogenous, causal environmental effects. Once an exogenous source of environmental variation has been identified, it is possible to look for differential treatment effects based on genotypical characteristics—polymorphisms, haplotypes (groups of polymorphisms that cluster uniquely together), and the like—that vary randomly within a given subpopulation (family, ethnic group, and so on). So, in short, the first task at hand for the social scientist who desires to show environmental-genetic interactions is the same task facing all social scientist who seek to rule out genetic (or other unobserved) factors when assessing causal, environmental effects.

There are a number of statistical approaches that economists have pioneered to obtain causal estimates. First, there are IV strategies (also called two-stage least squares) discussed above in relation to the work of Fletcher et al. (2009), which deploy a source of exogenous variation (i.e., the instrument, Z) to predict the covariate of interest (X) and then use the predicted covariate (X*) to model the outcome. (For a general review see Winship and Morgan 1999.) A particularly notable example of instrumental variable estimation is provided by Angrist (1990), who estimated the effect of military service during the Vietnam War period on subsequent earnings, using the draft lottery as a source of exogenous variation in veteran status. Another example is provided by Conley and Glauber (2006) who estimated effects of sibship size on parental educational investment, using the sex mix of the first two children born into a family to instrument whether or not parents have a third child (the sex of a child depends on the random segregation of X and Y chromosomes in the paternal gametes and U.S. parents are more likely to have a third child if the first two are of the same sex). More recently, economists have deployed regression discontinuity (RD) designs (see, e.g., van der Klaaw 2002, on the effects of financial aid on college enrollment decisions) where researchers compare subjects that fall just on either side of an otherwise arbitrary cutoff point—such as those who score a few points above or below an admissions test. And then, of course, there is actual experimentation in which researchers determine what sorts of conditions subjects are exposed to (see, for example, research on the randomized housing program, Moving to Opportunity; Katz, Kling, and Liebman 2001). In any of these cases, if genetic information were available for respondents, researchers could have estimated GE interactions—because they had properly estimated the "E" part in a way that we could be sure was uncorrelated with G. Another benefit of having genetic information is that researchers can demonstrate that a given genetic trait is not correlated with the presumed exogenous variation (e.g., the instrument or the randomized experiment) and that it is randomly distributed across at least measurable social categories. The major problem with the natural experiment approach, however, is that IV and RD approaches typically require huge sample sizes because they are inefficient estimation strategies. These are precisely the data sources—Social Security records, census samples, to name a couple—that are not likely to have genetic information. But there are other forms of putatively exogenous variation in social conditions that require

smaller sample sizes akin to that of social surveys such as the PSID. One such example is provided by the work of Strully (2009), who examined the health effects of job loss by comparing the impact of plausibly exogenous employment shocks (such as plant closings) to outcomes resulting from putatively endogenous sources of unemployment (such as dismissal for cause) using the PSID. If Strully had enjoyed access to genetic markers within the PSID, she may have been able to estimate a GE interaction with some confidence using her approach, even given the relatively small sample size (~1,500 persons).

Once we have an exogenous source of variation in, let us say, schooling, then we can identify an interaction effect between years of schooling and some genetic marker in looking at outcomes such as income, criminality, shyness, and so on. Let us take the example of Lleras-Muney (2005): She estimated the mortality returns to an additional year of high school by focusing on educational variation generated by changes in compulsory schooling laws during the first half of the twentieth century. These changes in state laws generated an exogenous change in the environmental characteristics of schooling because they affected everyone, regardless of genetic makeup or other characteristics. If she had enjoyed access to genetic information in her sample (which she did not, having used the U.S. Census as her data source), she would have been able to interact instrumented years of schooling (predicted based on these exogenous law changes and individual-level characteristics) with a given genetic marker when estimating the mortality effects of schooling (assuming that the genetic marker was not significantly associated with education and was randomly distributed across existing population divisions—such as race and socioeconomic status). In this way, she would have been able to tell whether certain genetic profiles receive larger health benefits from additional schooling than other genotypes—controlling for population-level stratification of alleles.

The linchpin to the approach, of course, is the validity of the claim to exogeneity of the environmental shock. By way of example, a paper by Caspi et al. (2002) that has become a classic in this area of research claims to have uncovered a GE interaction by comparing male children who have a particular functional polymorphism in the MAOA gene (monoamine oxidase A)—an enzyme that breaks down various neurotransmitters once they are chaperoned out of the synaptic cleft—with those who do not among a longitudinal sample of 1,037 white Australians followed from ages 3 to 26. Those individuals who showed a variable number tandem repeat (VNTR) in the promoter region of the gene (the area that precedes the actual coding portion but that is important to transcriptional activation and regulation) putatively transcribe (and by extension translate) MAOA at a lower rate than those without this polymorphism on their X-chromosomes. In turn, MAOA activity as indicated by this genetic difference was interacted with degree of maltreatment the respondents experienced between the ages of 3 and 11 to predict an index of antisocial behavior that included four measures ranging from criminal convictions to antisocial personality disorder criteria of the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (American Psychiatric Association 1994). They argued that though there do exist other MAO genes that may compensate for deficiencies in MAOA (in particular MAOB), these are not yet fully expressed among children, thus making MAOA particularly important with respect to moderating the effect of maltreatment during early childhood.

Eight percent of the sample experienced severe maltreatment, 28 percent experienced "probable" maltreatment, and 64 percent experienced no maltreatment. In a multiple regression context, the main effect of maltreatment level on the antisocial behavior index was significant, whereas the main effect of MAOA activity level was not, but an interaction effect between the two measures was statistically significant at the $\alpha = .01$ level. Caspi et al. (2002) argued that this is a true genetic-environmental interaction effect because

the MAOA genotypes were not significantly differently distributed across maltreatment levels—suggesting that this genotype did not itself influence exposure to maltreatment (i.e., the environment is not standing in for the genotype).

In a follow-up study (Caspi et al. 2003), they use the same cohort to examine the interaction of stressful life events with alleles of the serotonin transporter gene (5-HTT)-linked promoter region (5-HTTLPR). Specifically, individuals who have a short 5-HTT (i.e., upstream) promoter may show more propensity toward depression than those with a long promoter. However, previous studies found conflicting results; namely, many replications have failed to produce results claimed in earlier linkage studies. Some researchers had despaired that psychiatric and other behavioral phenotypes were controlled by so many quantitative trait genes that modeling genetic effects in a robust, direct way would not be possible and/or would account for little of the variation (see, e.g., Hamer et al. 1993). Caspi et al. argued instead that rather than complicated gene-gene interactions, the muddle of results could be resulting from GE interactions. This muddle motivates their search for an interaction effect of stressful life events and the 5-HTTLPR allele.

5-HTTLPR is an autosomal gene, so each individual has two copies. Thus, Caspi et al. (2003) compared three groups of individuals: those who were homozygous for the short alleles; those who were homozygous for the long alleles; and the heterozygotes who had one of each. They found that in the subsample who had experienced no stressful life events between ages 21 and 26, there was no difference between the three genotypes in the propensity to depression. However, as the number of self-reported stressful life events increased, the genotypes diverged with respect to their likelihood of clinical depression at age 26. They interpreted this as a GE interaction.

However, it could still be possible that what Caspi et al. (2002; 2003) were uncovering was actually a gene-gene interaction in both studies, because they did not have an exogenous source of environmental variation. In the latter case, those with the "at-risk," short alleles were, in fact, more likely to report stressful events than those who had long alleles. We may conclude, then, that measured genotype did influence the measured environmental factor. The researchers tried to get around this by reversing the time order: measuring stressful life events between ages 21 and 26 and measuring depression at age 21 (i.e., prior to the stressful life event). When they did this, they did not find the significant interaction that emerged in the "correctly" ordered model. However, it still may be the case that depression was induced by a gene-gene interaction because it may be an underlying unmeasured gene that causes the phenotype of "negative life events" to emerge in one's early 20s: Imagine a gene that causes excessive thrill-seeking and risk-taking, which, in turn, manifests as negative events during one's early adulthood. As for the MAOA interaction, we face the same issue: Though measured maltreatment did not vary by MAOA status, it could very well have varied by other genes (present in the parents and potentially passed on to the children). Thus, it would not be the maltreatment that interacted with MAOA status but rather the underlying, unmeasured genotype, which, in combination with given MAOA alleles, causes both parents and offspring to act antisocially.

In the same vein, Guo, Roettger, and Cai (2008) conducted a gene-environment interaction study, incorporating genotype into a social-control life-course model of delinquency. They use the Add Health sibling sample, excluding females, resulting in a sample of 1,100 males. They found interaction effects between MAOA, DAT1, and DRD2 genotype and family, school, or friendship network processes on self-reported delinquent behavior. In all cases of risky genotypes, increased social control from family, school, or peers reduced the genetic effect on delinquency, whereas reduced social control amplified it. They innovatively included

peer network information, including peer delinquency, network density, centrality, and popularity, and found a significant gene-environment interaction only for peer delinquency. However, like Caspi et al. (2002, 2003), they did not randomize environmental factors. Those with risky genotypes could select into delinquent peer groups, for example. Thus, like other gene-environment research, Guo et al. did not rule out a gene-gene interaction.

In fact, supporting the notional importance of gene-gene interactions (and offering a competing model to GE interactions) is recent genetics research that has shown that among the genes studied in humans (or other model organisms such as the fruit fly, *Drosophila melanogaster*, or the nematode worm, *Caenorhabditis elegans*), the vast majority of known genes are linked in a single network component when measured by either protein-protein interactions, regulatory relationships, or phenotypic covariation (as illustrated by the human case shown in Figure 2). This suggests that, indeed, one cannot conceptualize the perturbation of one gene as unrelated to the impact of other genes. Conversely, the embeddedness of this network suggests that genomic systems are highly redundant and robust and that other genes may be up- (or down-) regulated to compensate when a given gene is nonfunctional (or hypertrophic). For example, Isalan et al. (2008) have shown that even random rewiring of 598 promoter-gene relationships (one at a time) has little effect on phenotypic outcome or expression levels in *Escherichia coli* compared to wild-type bacteria of the same initial strain (95% survivorship among altered organisms)—suggesting that networks that are highly robust to failure.

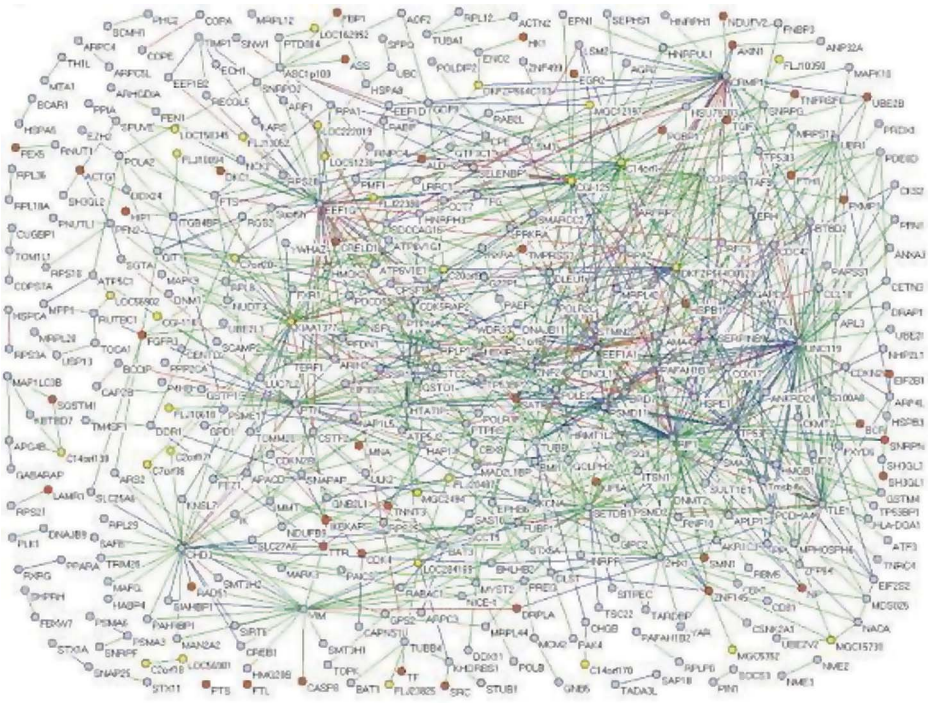

Figure 2. Protein-protein network in humans.
Source: Stelzl et al. (2005).

In sum, in order to investigate GE interactions, we need some source of exogeneity on the environmental side as a lever for estimation as well as carefully selected candidate genes that are chosen a priori based on their status as homologues to those manipulated in experimental animal studies (and evidence that these markers are not significantly associated with plausible subgroups in our sample or statistical controls for such possible associations). In fact, the *ideal* study would randomize genes (within families) as well as environment.

Discussion

As the preceding discussion I hope has made clear, doing sociogenomics is difficult but not impossible. There is much that human geneticists can learn from social scientists—particularly applied econometricians—who have long wrestled with questions of exogeneity and selection that are similar to complications facing human geneticists surrounding population admixture (i.e., allele stratification), linkage disequilibrium, and so forth. Likewise, social scientists who venture into this emerging subfield would be wise to team up with animal-based experimental geneticists. The interplay of human survey analysis (replete with DNA marker data) and animal experiments (such as genetic knock-outs or knock-ins) should be pursued in much the same way that rich ethnography confirms, refutes, elaborates, or stimulates quantitative analysis within the social sciences. Biologists who are performing genetic experiments might do well to read the social science literature for candidate homologous loci to manipulate in their own experimentation. A future of complementary wet-lab and survey research center collaborations is not too far-fetched to imagine.

After all, there is no reason why social scientists should be left out of the gold rush of analysis that is ensuing from the decoding of the human genome. Of course, there are potential risks, however, to the entire PSID effort if privacy or other concerns cause increased attrition (particularly when PSID is now collecting data only every other year). That said there is not much research to date on attrition in longitudinal panels that have collected DNA data in particular (or even biomarkers in general),

In sum, the factors that appear to affect participation in population studies that collect genetic data are no different than those that social scientists are used to. Wrote Harge (2006):

> The salience of the topic exerts the strongest effect on willingness. Incentives enhance response, especially incentives given early in the recruitment process. Personality, training, and experience of the recruiter have major effects, whereas demographic attributes have lesser effects, depending more on the specific setting. House-to-house or other in-person approaches typically (but not uniformly) elicit higher response than initial telephone contacts, but they are more expensive and harder to monitor for quality assurance. (p. 253)

With respect to DNA collection in particular, we are lucky that the field has coalesced around a collection protocol that seems to yield both the highest participation rate as well as very high-quality data: saliva samples. These are easy to collect—ideally face to face to increase participation although collection can be done by mail if costs are prohibitive—and they are the least invasive. Earlier approaches used either blood, which provided high-quality data but depressed response rates for obvious reasons, or buccal (inner cheek) swabs, which evinced higher participation but provided poor-data quality (only 31% of samples could be amplified in one study compared to 100% for blood and about three quarters of saliva samples; Hansen, Simonsen Finn, Nielsen, and Andersen Hundrup 2007). Saliva collection protocols such as the Oragene proprietary system appear to provide the

best of both worlds, with 80 percent participation rates in one study compared to 31 percent for blood and 76 percent for buccal swabs (*ibid*). (Hansen et al. 2007) What is more, they cost only a few dollars a sample, and prices are likely to fall over time. Finally, the saliva samples can be stored at room temperature for extended periods without significant degradation of DNA quality. This is particularly handy for researchers who face delays between collection and sequencing due to funding or other issues.

A perhaps more subtle (and thus more worrisome) concern is that willingness to participate (in DNA collection) varies by the polymorphism of interest (or by hidden paternity status), which would, obviously, lead to serious bias for all endeavors (Harge 2006). As such, it may be reassuring that the first study that examined whether willingness to participate in same study varied by haplotype found no impact (Bhatti, Sigurdson, Wang, et al. 2005) That said, the rub for social scientists not faced by medical geneticists is that if alleles matter to the social outcomes we care about, they are more likely to be related to willingness to participate ipso facto. Thus, statistical models should take care to include nonrespondents on DNA collection into their models, perhaps treating the nonparticipation flag as equivalent to a genotype itself.

There are also concerns on the "output" side. Namely, particularly in the United States, genetics and race have a particularly dubious intellectual co-history. Most recently, *The Bell Curve* (Herrnstein and Murray 1994) made a case (based on many dubious assumptions) that racial and class stratification in the contemporary United States was primarily a result of genetic differences in ability. Though the optimist may claim that such unfounded claims are the very reason we need to collect "real" genetic data, surely some who analyze the putative markers will be doing so with a political agenda. So it would behoove social scientists running such surveys that integrate socioeconomic and behavioral outcomes, demographic data, and genetic markers to be extra careful in how such data is managed and released. For example, perhaps only genes that have been well established using animal models (and/or other human associational studies) should be released to researchers to prevent "data mining" exercises that may lead to lots of initial controversial findings that later turn out to fail the test of replication. This would minimize the level of "unnecessary" political controversy over findings that turn out to be for naught.

Therefore, once coded, such data should be treated with the same level of security (or perhaps more) as the geo-coded sensitive data file now is. One model might be a Luxembourg Income Study model where the genetic data are housed at the University of Michigan's Institute for Social Research (ISR) and approved researchers have to go there to run models (and/or submit code over a secure web interface). Another model for data delivery might be the Census research data center model, which requires a background check and a research proposal before allowing access to researchers from contributing institutions. Add Health currently requires project descriptions of much shorter length and has a much speedier approval process (though a much less rigorous review). These models convey the advantage of screening for potentially hot-button papers before analysis is performed (a post-analysis, prepublication screen represents another option). The concern is to balance the need for openness of scientific research to any well-formed investigative endeavor with the concerns about misuse of the data for political ends.

In a similar vein, subjects themselves should be assured that even if they desire the genetic data, they will not be able to obtain them. Perhaps this is Pollyanna-ish given some research demonstrating that when subjects are told the results of their DNA analysis, their anxiety levels decrease or remain constant—at least for certain a certain alleles such as the hereditary hemochromatosis marker (Picot et al. 2009). However, providing subjects with information on a marker for a "medical" condition is quite different from informing them

about loci that may be linked (or erroneously linked) to socially sensitive outcomes such as academic achievement, depression, or delinquency that may be subject to strong Pygmalion effects. There is also, of course, subject concerns about revealed paternity and other kinship-relatedness discrepancies that might even selectively depress participation if absolute data anonymity is not assured.

References

Allison, D. B. 1997. Transmission-disequilibrium tests for quantitative traits. *Am. J. Hum. Genet.* 60:676–690.

Allison, D. B., H. Mooseong, N. Kaplan, and E. R. Martin. 1999. Sibling-based tests of linkage and association for quantitative traits. *Am. J. Hum. Genet.* 64:1754–1764.

American Psychiatric Association. 1994. *Diagnostic and Statistical Manual of Mental Disorders*. Fourth Edition. Washington, DC: APA.

Angrist, J. D. 1990. Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records. *Am. Econ. Rev.* 80(3): 313–336.

Bennett, A. J., K. P. Lesch, A. Heils, J. C. Long, J. G. Lorenz, S. E. Shoaf, M. Champoux, S. J. Suomi, V. Linnoila, and J. D. Higley. 2002. Early experience and serotonin transporter gene variation interact to influence primate CNS function. *Mol. Psychiatr.* 7:118–122.

Bhatti, P., A. J. Sigurdson, S. S. Wang, J. Chen, N. Rothman, P. Hartge, A. W. Bergen, and M. T. Landi. 2005. Genetic variation and willingness to participate in epidemiologic research: data from three studies. *Canc. Epidemiol. Biomarkers Prev.* 14:2449–2453.

Brookes, K., X. Xu, W. Chen, K. Zhou, B. Neale, N. Lowe, R. Aneey, et al. 2006. The analysis of 51 genes in DSM-IV combined type attention deficit hyperactivity disorder: Association signals in DRD4, DAT1, and 16 other genes. *Mol. Psychiatr.* 11:934–953.

Cases, O., I. Self, J. Grimsby, P. Gaspar, K. Chen, S. Pournin, U. Müller, et al. 1995. Aggressive-behavior and altered amounts of brain-serotonin and norepinephrine in mice lacking MAOA. *Science* 268:1763–1766.

Caspi, A., J. McClay, T. E. Moffitt, J. Mill, J. Martin, I. W. Craig, A. Taylor, and R. Poulton. 2002. Role of genotype in the cycle of violence in maltreated children. *Science* 297:851–854.

Caspi, A., K. Sugden, T. E. Moffitt, A. Taylor, I. W. Craig, H. L. Harrington, J. McClay, et al. 2003. Influence of life stress on depression: Moderation by a polymorphism in the 5-HTT gene. *Science* 301:386–389.

Conley, D., and R. Glauber. 2006. Parental Educational Investment and Children's Academic Risk: Estimates of the Impact of Sibship Size and Birth Order from Exogenous Variation in Fertility. *J. Hum. Resour.* 41(4):722–737.

Ding, W., S. F. Lehrer, J. N. Rosenquist, and J. Audrain-McGovern. 2009. The Impact of Poor Health on Academic Performance: New Evidence Using Genetic Markers. *J. Health Econ.* 28(3):578-597.

Fletcher, J. M., and S. F. Lehrer. 2009. *Using genetic lotteries within families to examine the causal impact of poor health on academic achievement.* Paper presented at the NBER 2007 Summer Institute. Available at: http://post.queensu.ca/~lehrers/genelotto.pdf (accessed).

Fowler et al. 2008. Genetic variation in political participation. *Am. Polit. Sci. Rev.* 102(2):233–248.

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

Guo, G., M. E. Roettger, and T. Cai. 2008. The interaction of genetic propensities into social-control models of delinquency and violence among male youths. *Am. Socio. Rev.* 73:543–568.

Hamer, D. H., S. Hu, V. L. Magnuson, N. Hu, and A. M. Pattatucci. 1993. A linkage between DNA markers on the X chromosome and male sexual orientation. *Science* 261:321.

Hansen, T., M. K. Simonsen Finn, C. Nielsen, and Y. Andersen Hundrup. 2007. Collection of blood, saliva, and buccal cell samples in a pilot study on the Danish Nurse Cohort: Comparison of the response rate and quality of genomic DNA. *Canc. Epidemiol. Biomarkers Prev.* 16:2072.

Harge, P. 2006. Participation in population studies. *Epidemiology* 17:252–254.

Herrnstein, R. J., and C. Murray.1994. *The Bell Curve: Intelligence and Class Structure in American Life*. New York: Free Press.

Isalan, M., C. Lemerle, K. Michalodimitrakis, P. Beltrao, C. Horn, E. Raineri, M. Garriga-Canut, and L. Serrano. 2008. Evolvability and hierarchy in rewired bacterial gene networks. *Nature* 452:840–845.

Jeong, H., S. P. Mason, A.-L. Barabási, and Z. N. Oltvai. 2001. Lethality and centrality in protein networks *Nature* 411:41–42.

Katz, L. F., J. R. Kling, and J. B. Liebman. 2001. Moving to opportunity in Boston: Early results of a randomized mobility experiment. *Q. J. Econ.* 116:607–654.

Krishnan, V., and E. J. Nestler. 2007. The molecular neurobiology of depression. *Nature* 455:894–902.

Lleras-Muney, A. 2005. The relationship between education and adult mortality in the U.S. *Rev. Econ. Stud.* 72(1):189–221.

Murphy, D. L., Q. L. S. Engel, C. Wichems, A. Andrews, K.-P. Lesch, and G. Uhl. 2001. Genetic perspectives on the serotonin transporter. *Brain Res. Bull.* 56(5):487–494.

Peaston, A. E. and E. Whitelaw. 2006. Epigenetics and Phenotypic Variation in Mammals. *Mammalian Genome* 17:365–374.

Picot J., J. Bryant, K. Cooper, A. Clegg, P. Roderick, W. Rosenberg, and C. Patch. 2009. Psychosocial Aspects of DNA Testing for Hereditary Hemochromatosis in At-Risk Individuals: A Systematic Review. *Genet. Test. Mol. Biomarkers* 13(1):7-14.

Price, A. L., A. Helgason, S. Palsson, H. Stefansson, D. St. Clair, O. A. Andreassen, D. Reich, A. Kong, and K. Stefansson. 2009. The impact of divergence time on the nature of population structure: An example from Iceland. *PLoS Genet.* 5(6):e1000505.

Rice, G., C. Anderson, N. Risch, and G. Ebers. 1999. Male homosexuality: Absence of linkage to microsatellite markers at Xq28. *Science* 284:665.

Shih, J. C., and R. F. Thompson. 1999. Monoamine oxidase in neuropsychiatry and behavior. *Am. J. Hum. Genet.* 65:593–598.

Stelzl, U., U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksöz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach, and E. E. Wanker. 2005. A human protein-protein interaction network: A resource for annotating the proteome. *Cell* 122:957–968.

Strully, K.W. 2009. Job Loss and Health in the U.S. Labor Market. *Demography* 46:221-246.

Thornton, K. R., and J. D. Jensen. 2007. Controlling the false-positive rate in multilocus genome scans for selection. *Genetics* 175:737–750.

van der Klaauw, W. 2002. Estimating the effect of financial aid offers on college enrollment: A regression-discontinuity approach. *Int. Econ. Rev*. 43:1249–1287.

Winship, C., and S. L.Morgan. 1999. The estimation of causal effects from observational data. *Annu. Rev. Sociol.* 25:659–707.

Wong, A. H., I. I. Gottesman, and A. Petronis. 2005. Phenotypic differences in genetically identical organisms: The Epigenetic Perspective. *Hum. Mol. Genet.* 14(1):R11–R18.