White Paper for NSF Workshop on Genes, Cognition, and Social Behavior

Daniel J. Benjamin[1], Assistant Professor

Economics Department, Cornell University

Due to time and space constraints, I will restrict my discussion to molecular genetics in the social sciences, an enterprise which I am optimistic will eventually be transformative.

Broadly speaking, I believe that molecular genetics will ultimately contribute to social science research in three main ways. First, genetic information could eventually be useful for targeting social-scientific interventions, much like it is beginning to be useful for targeting medical interventions. For example, if dyslexia can eventually be predicted sufficiently well by genetic screening, children with dyslexia-susceptibility genes could be taught differently how to read from a very young age.[2]

Second, social scientists could use genotypic data to learn about the biological mechanisms that lead to behaviors of interest. One possibility is that the genetic data bear on existing hypotheses. For example, experiments in which humans were exposed to the neuropeptide oxytocin suggest that oxytocin causes trusting behavior.[3] This suggests the hypothesis that variation in the gene *OXTR*, which encodes the receptor for oxytocin, may be related to variation in trust-related behaviors.[4] Even more intriguingly, the genetic data might suggest new hypotheses. If a genetic marker is unexpectedly found to associate with some behavior, then the marker's biological pathway is implicated in that behavior.

Third, social scientists might be able to use genotypic data to more effectively address social science questions---questions that, in themselves, may have nothing to do with genetics. Rather mundanely, genotype data can be used simply as control variables to increase power in otherwise-standard statistical analyses. Most intriguingly---and most speculatively for reasons explained below---social scientists may be able to use genetic markers as "instrumental variables (IVs)" to infer the causal effect of (non-genetic) factor *X* on (non-genetic) factor *Y* using observational data. Among the several economics papers that already attempt to use this strategy, Fletcher and Lehrer (2009) study the effect

---

[1] I take sole responsibility for the opinions expressed here, but many of my views and most of my knowledge is the result of countless conversations with friends and collaborators, including Jonathan Beauchamp, David Cesarini, Christopher Chabris, Ed Glaeser, Ben Hebert, and David Laibson.

[2] See Schumacher et al (2007) for a recent review of genetic predictors of dyslexia.

[3] Kosfeld et al (2005).

[4] Indeed, Israel et al (2009) report an association between OXTR and dictator game giving---but Apicella et al (forthcoming) fail to replicate it in a much larger sample.

of mental health (*X*) on academic achievement (*Y*).[5] In effect, the idea is to use the fact that genotypes affecting mental health are randomly assigned among siblings within a family as a natural experiment. Under the assumption that the genetic marker IVs affect academic achievement only via their effect on mental health, the estimated causal effect of the genetic markers on academic achievement can be rescaled appropriately to infer the magnitude of the causal effect of mental health on academic achievement.

Despite the recent explosion in the number of papers reporting genotype-behavior associations, I am pessimistic that any of these potentially transformative contributions can be convincingly realized within the next 10 years. The most urgent problem---discussed below---is that genotype-behavior associations have tiny effect sizes, so current research designs in the social sciences are woefully underpowered. However, even once this problem has been solved, there are further obstacles that must be overcome before the contributions can commence.

For one thing, the biological-mechanisms and genes-as-IVs contributions require uncovering the causal effect of particular genetic markers on behavior, but most existing research designs focus on detecting correlations. There are myriad confounds to a causal interpretation, e.g.: genotypes are correlated with ethnicity which is correlated with behavior; an individual's genotype is correlated with his parent's genotype which is correlated with his family environment; and each genotype is highly correlated with many nearby genotypes that are in "linkage disequilibrium" with it. Ultimately, convergent evidence for a causal relationship will come from large family samples, where behavioral differences across siblings can be attributed to Mendelian random assignment of genotypes; modeling, measurement, and estimation of environmental factors and gene-environment interactions; experimental evidence from animal models where genes are selectively "knocked out"; and biological evidence on the function of protein products of the gene.

There is a further obstacle to using the IV strategy credibly. For IV estimation to be valid, not only must the genetic markers have sufficient predictive power for the *X* variable, but the causal effects of the genes must be understood well enough to rule out alternative pathways (besides *X*) by which the genes could affect outcome *Y*. Since the proteins produced by genes generally appear to have multiple effects, most of which we have barely begun to understand, it seems unlikely that we can be confident about all of the consequences of any particular genotype in the foreseeable future.

Targeting interventions is probably the potential contribution closest at hand because the genetic markers can be merely predictive, rather than established to be causal, and because an index composed of many markers can be used, which may in the aggregate have sizeable predictive power even if any constituent marker in the index has little. However, while I expect eventual successes, it will likely be slow and challenging to find sufficient predictive power even from an index. In medical genetics, with the exception of a few rare single-gene disorders, there has been a general failure to find

[5] The other papers are Ding, Lehrer, Rosenquist, and Audrain-McGovern (2009); Norton and Han (2009); and von Hinke Kessler Scholder, Smith, Lawlor, Propper, and Windmeijer (2010).

sizeable aggregate predictive power---a problem now called the "missing heritability" puzzle.[6] Consider height, a highly-studied physical trait that is both measured with much less error than most behavioral traits and is more heritable, with behavioral genetics studies on twins and other relatives indicating that about 80% of the variability in height is due to genetic factors. Yet the aggregate predictive power from known genotypes is only about 5%, with 0.3% being the largest R^2 that has been found out of the 44 genotypes so far found to be associated.[7] Given the failure to find sizeable predictable power in physical traits, the challenge is likely to be at least as large for behavioral traits where the causal mechanisms are arguably more complex.

The most urgent problem, however, is that most efforts in the social sciences to discover genetic associations are underpowered. Fundamentally, there are two reasons. First, with the exception of rare mutations, almost every true genotype-behavior correlation is probably very small. To take a social science example that seems fairly typical, a meta-analysis of 46 studies concluded that variation in the *COMT* gene explains 0.1% of variance in cognitive ability.[8] Second, while my collaborators and I were initially encouraged by the large number of associations reported regularly, we have now come to the view that the usual concerns about publication bias---the tendency for findings, as opposed to non-findings, to be selectively reported by researchers and selectively published by journals---are magnified in genetic association work because the typical dataset has many behavioral measures and many genetic markers. In order to account for publication bias and multiple hypothesis testing, it is important to adopt stricter statistical significance thresholds than usual, further reducing the power of a study with any given sample size to detect a true association.

If studies are underpowered, then the rate of false positives will be high. In the medical genetics community, it is now widely accepted that most published associations are not reproducible.[9] In my own social science work and the work of my close collaborators, we have been disappointed by our failure to replicate initially promising associations between genetic polymorphisms and economic phenotypes, despite samples of several thousand individuals.[10] Consequently, we have begun to systematically test existing candidate genes. In ongoing work using the Wisconsin Longitudinal Study (WLS), my collaborators and I attempted to replicate previously-reported associations of 13 genetic markers with cognitive ability. We can reject the hypothesis that the mean effect of those markers is larger than a tiny R^2 = .05%---and given our sample size of 5,413 individuals, we have essentially 100% power to detect effects of that size.[11] Also using the WLS, Freese et al (2010) attempt to replicate

[6] See, e.g., Sklar, Purcell, et al (2009).

[7] Wheedon and Frayling (2008).

[8] Barnett, Scoriels, and Munafò (2008).

[9] See Ioannidis et al (2001) and Hirschhorn et al (2002).

[10] Beauchamp et al (2010) and Benjamin et al (2009).

[11] Chabris et al (2010).

associations reported in the literature between *Taq*1a and educational attainment, voting, partisanship, organization memberships, socializing, tobacco use, and alcohol use, and conclude that none of the associations replicate.

To get a sense for the magnitude of the problem, consider a researcher studying a particular candidate genetic marker. To simplify, suppose there are only two alleles for the marker, with carriers of the High variant, as opposed to carriers of the Low variant, hypothesized to have a higher value for the phenotype of interest. To further simplify, suppose there are only two possibilities: either there is a true association, or there is not. Imagine the phenotype of interest is distributed normally. Suppose it is known that, if there is an association, then the genotype of interest explains R^2 = 0.1%---a rather large effect size for a single marker. For illustrative purposes, suppose any given sample has an equal number of High and Low carriers; in the usual case of asymmetric frequencies, the same amount of statistical power may require a much larger sample size. Finally, suppose that in a sample of size *N*, a researcher observes a statistically significant association at the standard .05 significance level. How large does *N* have to be in order for this result to constitute substantial evidence about whether there is an association? Table 1 shows how a researcher's posterior belief (after having seen the data) that there is a true association depends on the researcher's prior belief and on *N*.

Table 1. Posterior probability of a true association as a function of prior probability and sample size.

		Sample size				
		N = 100 (power = .06)	*N* = 1,000 (power = .17)	*N* = 5,000 (power = .61)	*N* = 10,000 (power = .89)	*N* = 30,000 (power = .99)
Prior probability of true association	.01%	.01%	.03%	.12%	.18%	.20%
	1%	1%	3%	11%	15%	17%
	10%	12%	27%	58%	66%	69%

Notes: Entries calculated by the author as described in the text. Power is calculated using Purcell, Cherny, and Sham's (2003) online tool: http://pngu.mgh.harvard.edu/~purcell/gpc/qtlassoc.html. Posterior probabilities are then calculated by Bayes' Rule:
Pr(true|significant) = (power × prior) / ((power × prior) + (.05 × (1-prior))).

Of course, it is difficult to know what an appropriate prior belief is, but for a typical candidate marker, it is probably much less than 10%. In any event, the clear message from these calculations is that a researcher should conclude almost nothing about a genotype-behavior relationship from a sample size in the hundreds, and sample sizes must number in the several thousands before non-negligible inferences are appropriate.

Relative to complex behavioral phenotypes, the power challenge is less daunting for intermediate phenotypes, such as functional Magnetic Resonance Imaging (fMRI) data, but adequately-powered research still requires sample sizes much larger than is currently typical. For instance, suppose it is known that, if there is an association, then the genotype of interest explains R^2 = 3%. Under the

same optimistic assumptions as above, for the conventional 80% power level, a sample size of *N* = 258 is required. In contrast, due to the cost of using the fMRI scanner, a typical large fMRI study currently has a sample size of *N* = 100.

Over the next few years, due to the plummeting cost of genome-wide scans, virtually all association studies will move from being candidate gene studies to being Whole-Genome Association Studies (GWAS). This switch is scientifically appropriate: Existing candidate genes were initially studied primarily because those genetic markers were technologically feasible to genotype. There is every reason to believe that markers elsewhere on the genome will be more strongly associated with behavioral phenotypes than the tiny fraction of all markers that happened to be available to researchers first. However, concerns about power are many times more severe in GWAS. Current GWAS platforms genotype about 2 million markers, and future platforms will genotype far more, so the prior probability on any particular marker must be miniscule, probably much smaller than .01%.

In my view, if a funding agency were to fund genetics research in the social sciences, the clear top priority is to put together datasets that are large enough to have adequate power to detect genotype-phenotype relationships in GWAS. Over the past several years, the medical genetics community has paved the way, forming large consortia of data providers, the most famous example being the Wellcome Trust Case Control Consortium. The resulting samples on the order of *N* = 20,000-30,000 are sufficiently large to detect alleles with modest effect sizes. These studies also tend to be very stringent in their hypothesis testing, thereby reducing the risk of false positives. Indeed, the findings that have emerged from these cooperative studies appear to be more likely to survive the challenges of replication.

I am optimistic that social scientists can follow suit within the next few years. For one thing, although existing medical consortia mainly study disease phenotypes, most of these datasets contain basic markers for socioeconomic outcomes as well. Furthermore, there are a number of large-scale social science datasets that have begun genotyping participants or plan to do so in the near future. The cost and ease of genotyping is plummeting: A commercial whole-genome scan of an individual (which measures about 2 million markers) currently costs less than $500, and since 1990, the price has been falling by half every 1-2 years.[12] Consequently, it seems likely to become standard for large-scale social science data providers from all over the world to genotype their participants, with an aggregate sample size of several hundred thousand.

Should a funding agency put money behind genetics research in the social sciences? I believe the answer is yes *if* (1) the research will have adequate power, and (2) the researchers are held to an unusually high standard of accurately communicating their results. More insistently than for other research, funding agencies should *require* grant proposals to include power calculations. Unfortunately, underpowered research has negative value-added because it generates false positives; some researchers will squander resources pursuing a dead end, and others will spend resources undoing the

[12] http://singularityhub.com/2008/12/30/whole-genome-sequencing-to-cost-only-1000-by-end-of-2009/ as accessed on June 9, 2010.

damage by publishing non-replications. Moreover, due to the media attention any gene-behavior association work will surely attract, even adequately-powered research runs the risk of exposing the general public to a rollercoaster ride of frequently-reported genetic associations with important social behaviors that subsequently turn out to be false positives. Funding agencies should pay attention not only to whether researchers are capable of carrying out the scientific work, but also whether the researchers are committed to highlighting the limitations of the work, such as the possibility of a false positive, and the appropriate interpretation of the work, namely tiny predictive power from any given genetic marker and inevitability of gene-environment interaction. In addition, funding agencies should encourage grant proposals that aim to replicate previously-obtained results and encourage publication, even/especially if the attempted replications fail.

If these conditions are met, then I think funding agencies should view molecular genetics research as having an attractive risk-return profile. It is high risk because it is possible that the enterprise as a whole may fail; there may be too many genetic markers with effects that are too small and too complex, and hence researchers may never be able to pin down causal relationships that have non-tiny predictive power in the aggregate. Even if the research is successful, it will be slow-going over many years, literally with the character of trying to find needles in a haystack, one at a time. However, the ultimate contributions to social science are potentially quite large, and molecular genetics research is rapidly becoming remarkably inexpensive.

Another funding priority of at least equal importance---about which I write less only because I know less---is research on the economics, politics, and ethics of using genetic information by both public agents (like governments) and private agents (like therapists and insurance companies). The cheap and plentiful availability of genetic data outside the scientific community in the near future will raise enormous social challenges. Research that studies these challenges may anticipate and offer policy suggestions to reduce potential risks.

References

Apicella, Coren L., David Cesarini, Magnus Johanneson, Christopher T. Dawes, Paul Lichtenstein, Bjorn Wallace, Jonathan Beauchamp, and Lars Westberg (in press). "No association between oxytocin receptor (OXTR) gene polymorphisms and experimentally elicited social preferences." *PLoS One*.

Barnett, J.H., L. Scoriels, and M.R. Munafò (2008). "Meta-analysis of the cognitive effects of the catechol-O-methyltransferase gene Val158/108Met polymorphism." *Biological Psychiatry*, 64, 137-144.

Beauchamp, Jonathan, David Cesarini, Niels Rosenquist, James H. Fowler, and Nicholas Christakis (2009). "A Genome Wide Association Study of Educational Attainment." Harvard University Working Paper.

Benjamin, Daniel J., Christopher F. Chabris, Edward L. Glaeser, Vilmundur Gudnason, Tamara B. Harris, David I. Laibson, Lenore Launer, and Shaun Purcell (2009). "Genetic Influences on Economic Behavior." Presentation at IZA/Volkswagen Foundation Workshop, November.

Chabris, Christopher F., Benjamin M. Hebert, Daniel J. Benjamin, Jonathan Beauchamp, Craig Atwood, Jonathan Freese, Taissa S. Hauser, Robert M. Hauser, and David I. Laibson (2010). "Most published SNP associations with general cognitive ability are probably false positives." Presented at the Integrating Genetics and the Social Sciences Conference, Boulder, Colorado, June 2–3.

Ding, Weili, Steven Lehrer, Niels Rosenquist, and Janet Audrain-McGovern (2009). "The Impact of Poor Health on Academic Performance: New Evidence Using Genetic Markers." *Journal of Health Economics,* 28(3), May, 578-597.

Fletcher, Jason, and Steven Lehrer (2009). "The Effects of Adolescent Health on Educational Outcomes: Causal Evidence Using Genetic Lotteries between Siblings." *Forum for Health Economics & Policy*, 12(2), Health and Education, Article 8.

Freese, Jeremy, Amelia R. Branigan, Craig S. Atwood, Taissa S. Hauser, Daniel J. Benjamin, Christopher F. Chabris, David Laibson, and Robert M. Hauser (2010). "*Taq*1a and College Attendance, Partisanship, Voting, and Other Outcomes: Replication Attempts Using The Wisconsin Longitudinal Study." Northwestern University Working Paper.

Hirschhorn, J.N., K. Lohmueller, E. Byrne, and K. Hirschhorn (2002). "A comprehensive

review of genetic association studies." *Genetics in Medicine*, 4, 45-61.

Ioannidis, J.P.A., E.E. Ntzani, T.A. Trikalinos, and D.G. Contopoulos-Ioannidis (2001). "Replication validity of genetic association studies." *Nature Genetics*, 29, 306-309.

Israel, Salomon, Elad Lerer, Idan Shalev, Florina Uzefovsky, Mathias Riebold, Efrat Laiba, Rachel Bachner-Melman, Anat Maril, Gary Bornstein, Ariel Knafo, and Richard Ebstein (2009). "The oxytocin receptor (OXTR) contributes to prosocial fund allocations in the dictator game and the social value orientations task." *PLoS One*, 4, e5535.

Kosfeld, M., M. Heinrichs, P.J. Zak, U. Fischbacher, and E. Fehr (2005). "Oxytocin increases trust in Humans." *Nature*, 435, 673–676.

Norton, Edward C., and Euna Han (2008). "Genetic Information, Obesity, and Labor Market Outcomes." *Health Economics*, 17(9), 1089-1104.

Purcell, Shaun, S.S. Cherny, and Pak C. Sham (2003). "Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits." *Bioinformatics*, 19(1), 149-150.

Schumacher, Johannes, Per Hoffman, Christine Schmäl, Gerd Schulte-Körne, and Markus Nöthen (2007). "Genetics of dyslexia: The evolving landscape." *Journal of Medical Genetics*, 44, 289-297.

Sklar, Purcell, et al (2009). "Common polygenic variation contributes to risk of schizophrenia and bipolar disorder." *Nature*, 460(6), August, 748-752.

von Hinke Kessler Scholder, Stephanie, George Davey Smith, Debbie A. Lawlor, Carol Propper, and Frank Windmeijer (2010). "Genetic Markers as Instrumental Variables: An Application to Child Fat Mass and Academic Achievement." University of Bristol Working Paper.

Weedon, Michael N., and Timothy M. Frayling (2008). "Reaching new heights: insights into the genetics of human stature." *Trends in Genetics*, 24(12), 595-603.