White Paper

Integrating genomic data and social science
Challenges and opportunities

Jeremy Freese
Department of Sociology
Northwestern University
1810 Chicago Avenue
Evanston, IL 60208
jfreese@northwestern.edu

Why should social scientists be interested in using molecular genetic data? Here are five reasons:[1]

1. Given evidence from twin- and other family-based designed studies of the causal importance of genetic differences on a wide range of outcomes of social scientific interest, integrating genetic causes into social science theories is a necessary task toward understanding and explaining variation in these outcomes.
2. Abundant evidence points to the potential for genetic causation confounding estimates of social or other environmental causes on outcomes, and thus failure to account for confounding by genetic differences can lead to large biases throughout social science studies of individual-level outcomes.
3. The *strict intragenerational exogeneity of the DNA sequence*—that the DNA sequence does not change as a result of external events or internal development—means that DNA sequence information is intrinsically prospective, no matter when in the life course it is collected. This, in turn, suggests the possibility of genetic data being leveraged using "natural experiments" methods to better estimate effects of particular environmental causes on outcomes in situations that might otherwise appear intractable because of pervasive endogeneity.[2] (This strict exogeneity should not be confused with parallel interest in "epigenetics," which broadly speaking encompasses various mechanisms by which environments modify DNA expression.)
4. Genetic measurement provides an entirely new and more powerful set of tools for studying migration and mating patterns.
5. Given the usual failure of conventional social science models of individual outcomes to explain much of the existing variation in those outcomes, gene-environment interactions might be an important reason why individuals with similar social backgrounds and similar measured experiences often still have very different outcomes.

To date, of course, research using molecular genetic data has been dominated by the pursuit of medical knowledge. As social scientists become interested in using genetics in the study of a broader range of individual outcomes, an important question is whether they can use lessons from the history of medical genetics research to minimize the extent to which social scientists repeat the same problems. Most prominent here is that, for studying the association between genetic variants and outcomes: *"discovery" is the easy part; separating true discoveries from false ones is much harder.*

The medical genetics literature has a large number of published associations that has subsequently failed to be replicable, including some that have received considerable media attention.[3] The "candidate gene" approach of genetic research looks, from afar, like the proper method for science, with articulated hypotheses applied to data. In a candidate gene study, a specific genetic variant is hypothesized as influencing a specific

doi: 10.2990/30_2_88

outcome, with its effect possibly moderated by a specific environment (a gene-environment interaction, often abbreviated to G × E) or another specific gene.

Candidate gene studies are now regarded with much suspicion in many areas of medical genetics because *post hoc* explanations are relatively easy to recast as *a priori* hypotheses and because even findings from purely *a priori* hypotheses can result in a distorted literature due to publication biases of investigators, reviewers, and editors. Indeed, candidate gene studies that have been published so far in major sociology, demography, and political science journals have various features that, taken together, could be read almost as a catalog of indicators of results that are unlikely to replicate reliably in new samples: *ad hoc* model specification, *ad hoc* subgroup restrictions, *ad hoc* genetic models, and *ad hoc* selections of environmental variables for analyses of gene-environment interactions. Worth emphasizing here is that this work uses neither logic nor methods that differ from standard practice in social science; rather, the problem is that the experience of medical genetics strongly suggests that applying these practices to genetic data leads to abundant false positives.

Analyses have tended to eschew power analyses and have reported effect sizes far larger than any reasonable expectation about the possible effect size. To take one example, a study of educational attainment published in the *American Journal of Sociology* reported a main effect size for a genetic variant (*Taq*1a) on going to college that is as large as the total difference in college attendance rates between black and white respondents in the sample.[4]

Indeed, nothing as yet exists to contradict the gloomy hypothesis that the aforementioned social sciences have yet to publish a single genetic main effect or gene-environment interaction that is "real" in the sense of an established, replicable causal relationship that still appears reasonable. In other words, it is quite reasonable to suppose that *none* of the dozens of studies that have been published to date will withstand future empirical study on independent data.

The primary source of this problem is plain enough: *presently there are enough data to discover associations, but not enough data to discern true associations from false ones*. Most of the candidate gene studies in social science have so far relied on genetic data available from a single data source, the National Longitudinal Study of Adolescent Health (Add Health). Add Health deserves enormous credit for being pioneering in obtaining molecular genetic assays and in making its data securely available to a broad number of investigators without onerous co-authorship agreements. But, medical genetics makes plain that single-dataset discoveries of gene-outcome associations (and, worse, gene × environment outcome associations) are prone to very high rates of replication failure.

Accordingly, any literature for which "discoveries" of gene or gene × environment associations from single datasets and with *ad hoc* specifications are publishable is a literature that will be replete with false positive findings. Opinions vary about the pragmatic virtues of weeding out false positive findings before or after publication, but, one way or another, their weeding is an absolute necessity to establish any firm basis toward realizing any of the potential benefits of genetic data to social science presented above. This can only be done with more data that are available to more investigators. Following an emerging standard in medical genetics, the journal *Behavior Genetics* now has declared as a matter of policy that studies must have a replication in a different sample prior to publication.[5]

Happily, much more data is on the way, including assays in other large population datasets with long-established track records, like the Wisconsin Longitudinal Study (WLS) and the Health and Retirement Study (HRS). Funding for these initiatives is driven almost entirely by the prospective contributions of these datasets to health-related research. Fortunately, in the aforementioned cases a broader substantive range of outcomes of interest to social scientists happen to be included, along with a range of psychosocial measures that can be used toward possibly identifying mediating psychological mechanisms of genes and social science phenotypes. Add Health, WLS, and HRS all have significant cognitive assessments, for instance. Given the complexity and cost of assaying—even as the cost rapidly declines—social science funding sources may receive a better return from attempting to extend social science measures for which assays are available or underway, rather than supporting assaying of respondents for other datasets unless these offer particular advantages and wide availability. Importantly, a condition of investing in social science measures needs to be that these measures will be available to the broad community of investigators for analysis.

https://doi.org/10.2990/30_2_88 Published online by Cambridge University Press

In medical science there has been increasing movement towards consortia that allow for inference from combined datasets. Regardless of how cheap genotyping becomes, such consortia seem necessary for various types of social science studies with genetic data to be conducted with any appreciable statistical power, barring some substantial revision in the variance accounted for by individual genetic variants.

Consortia in medicine are logistically complicated in ways that are consistent with the high expense and broad distribution of specialized methodological expertise in health research. Even under bullish scenarios, the integration of genetics into social science will be carried forward by fewer people doing projects for less money, and consortia need to be organized in ways that are nimble, efficient in terms of staffing required to access data, and set up to share expertise as well as data. For this, *outreach projects focused on improving data accessibility and methodological training may be particularly valuable for social science*, especially insofar as they help toward building ties across institutions to compensate for the more diffuse affiliations of investigators.

Much of the social science interest in genetics has focused on possible gene × environment interplay, and often in terms of "contextual" environmental variables. An example would be recent work on how heritability of smoking varies as cigarette taxes vary.[6]

Consortia seem essential to the extension of work to molecular genetic data, given that power considerations are even more acute for estimating systematic moderation of causal effects than for estimating average causal effects. Beyond this, however, such studies may benefit particularly from ongoing work that is attempting to extend inference for sparse data by combining information across datasets (an example of this would be using census data to strengthen state-level inferences in a public opinion poll that would be otherwise too sparsely distributed). In other words, as statistical power appears likely to be a vital issue for any applications of molecular genetic data in social science, *the continued development of methods to increase power of studies by combining data sources* will be among those most beneficial to the enterprise.

Because health research has so far provided the major resource for molecular genetic data collection and analysis, much of the methodological apparatus has developed with health outcomes foremost in mind. As interest in genetics has broadened to substantive domains that are unrelated (or not directly related) to health, the possibility increases of methodological problems due to disanalogies between prototypic health outcomes and other outcomes social scientists study. Many social attainments, for example, are of interest to social scientists in no small part because of their intergenerational reproduction—that is, socioeconomic attainment in one generation provides an advantage toward socioeconomic attainment in the next. Moreover, ancestry itself is a source of social categorization and action by others upon that categorization.

For these reasons, what genetic research calls the problem of "population stratification" glosses over a series of fairly foundational social dynamics, whose implications for the study of genetics and attainments are at present largely unknown. Population stratification encompasses many ways that the nonrandom assortment of genes in populations can result in environmental causes of outcomes being mistaken for genetic causes. This marks a key area for theoretical development, but it also has the direct consequence that population stratification likely provides a much more significant problem for estimation than many social scientists presently appreciate. This is especially so for approaches to population stratification other than the simplest and most assured: the analysis of sibling data. *When evaluating different candidate data sources for investment, the special value of data with siblings—and, even better, combinations of siblings and parents—needs to be emphasized.*

Large-scale assaying has yielded some applications of inferences based on deviations from 0.5 inheritance by descent among full siblings.[7] This has been suggested as a general technique by which sibling data could be used to make inferences similar to what twin data are used for presently, which would have the auspicious consequence of alleviating concerns about particularities of twinning and twin-based sampling. Such a technique may make sibling data even more valuable to genetics studies. At the same time, with some further assumptions, genome-wide data can be used to apply a similar technique to the analysis of unrelated individuals, estimating the degree to which chance genetic similarities between unrelated individuals are associated with greater similarity in outcomes.[8]

Another area in which seemingly strong methodological promise is tempered by practical ambiguity is

https://doi.org/10.2990/30_2_88 Published online by Cambridge University Press

the use of genetic variants to conduct instrumental variable estimation (sometimes called "Mendelian randomization" techniques).[9] A medical example is using a known genetic marker of variation in C-reactive protein to estimate the relationship between C-reactive protein levels and cardiovascular disease, an apparent relationship that might be spurious.[10] An attempted social application has been to use genetic variants associated with obesity to estimate the relationship between obesity and socioeconomic attainments. On the one hand, the techniques seem very promising for addressing problems that otherwise might be intractable because of pervasive reverse causality, given the natural intragenerational exogeneity of genes. On the other hand, instrumental variables estimators imply satisfying the exclusion restriction that the only way that the instrument (i.e., the genetic variant) influences the outcome (e.g., socioeconomic attainment) is through the independent variable (e.g., obesity), and not through some other cause (e.g., cognitive ability). Given that genes typically have very small and multiple effects, this exclusion criterion will almost certainly be often violated, although the consequences of this violation might be mitigated by the ability to use a number of different variants as instruments. Again, though, it seems like basic methodological work is going to be central to figuring out whether substantive breakthroughs for social science can be achieved using these techniques or whether the problems in satisfying the assumptions of the techniques are effectively insurmountable.

For genetics research to realize its promise in the social sciences, we need more data and more development of methods with specific challenges of social science analysis at the fore. Additionally, genetics work in the social sciences remains relatively underdeveloped in terms of the integration of actual social science theory. For instance, basic sociological or economics theory expects people to specialize in areas in which they evince aptitude and that specialization will lead to further gaps in proficiency between specialists and others. If aptitudes in various domains are ubiquitously influenced by genes, as behavioral genetics would lead us to expect, then gene-environment correlations should be a ubiquitous and essential feature of the social world. At least regarding social attainments, genetic predictors of skills and attainments should be pervasively positively correlated with conditions promoting those same skills and attainments.

As another instance, a staple of epidemiological sociology is that social differences, especially in education, influence the extent to which individuals can act upon knowledge to achieve better health outcomes.[11] We would therefore expect take-up of knowledge gained from genetics to differ by social groups in ways that are presently underexamined. More than this, differences in action on the indirect information about inheritance and disease that already exist as family history may be an important systematic moderator of the relationship between genes and health outcomes. In other words, if people whose family histories lead them to perceive greater risk of a heart attack are motivated by this to exercise more, that would affect the observed relationship between genetic susceptibility and actual heart disease. If this kind of preventive action is done more (or more effectively) by those of higher socioeconomic standing, that would affect observed gene-environment interactions.[12]

The strength of social science is its dynamic vision of actors with beliefs and preferences interacting with one another and with larger institutions. The implication of causally relevant genetic differences needs to be fully integrated into that vision.

In the long run, molecular genetics work will almost certainly transform our understanding of basic human behavior and the conduct of the social scientific study of individual-level outcomes. Evidence of the importance of genetic causes of social science outcomes is abundant, as is their potential for revision and elaboration of our understanding of social causes—and data that will allow increased understanding of these causal relationships is increasing rapidly and inexorably. At the same time, we are at the point where pitfalls of inferences from genetic data are apparent and a key part of investment is figuring out the most efficient way of navigating these pitfalls, with a minimum of accumulated distrust from premature claims. Achieving this efficiency is going to require basic work on data availability, the dissemination of expertise, the creation of collaborative relationships across institutions, the development of methods, and the improved conceptual integration of genetics with social science theory.

Note

Jeremy Freese is Professor in the Department of Sociology and a Fellow of the Institute for Policy Research at

https://doi.org/10.2990/30_2_88 Published online by Cambridge University Press

Northwestern University. His work encompasses a variety of topics on the integration of biological, psychological, and social levels of analysis.

References

1. See also Jeremy Freese, "Genetics and the social science explanation of individual outcomes," *American Journal of Sociology*, 2008, 114:S1–S35.

2. Jason M. Fletcher and Steven F. Lehrer, "Using genetic lotteries within families to examine the causal impact of poor health on academic achievement," NBER Working Paper Series, July 2009, http://ssrn.com/abstract=1434663

3. John P. A. Ioannidis, "Why most published research findings are false," *PLoS Medicine*, 2005, 2:e124.

4. Michael Shanahan, Stephen Vaisey, Lance D. Erickson, and Andrew Smolen, "Environmental contingencies and genetic propensities: Social capital, educational continuation, and a dopamine receptor polymorphism," *American Journal of Sociology*, 2008, 114:S260–S286.

5. John K. Hewitt, "Editorial policy on candidate gene association and candidate gene-by-environment interaction studies of complex traits," *Behavior Genetics*, 2012, 42:1–2.

6. Jason D. Boardman, "State-level moderation of genetic tendencies to smoke," *American Journal of Public Health*, 2009, 99:480–486.

7. Peter Visscher, Sarah E. Medland, Manuel A. R. Ferreira, Katherine Morley, Gu Zhu, Belinda K. Cornes, Grant W. Montgomery, and Nicholas G. Martin, "Assumption-free estimation of heritability from genome-wide identity by descent sharing between full siblings," *PLoS Genetics*, 2006, 2:e41.

8. Sang Hong Lee, Naomi R. Wray, Michael E. Goddard, and Peter M. Visscher, "Estimating missing heritability for disease from genome-wide association studies," *American Journal of Human Genetics*, 2011, 88: 294–305.

9. George Davey Smith, "Mendelian randomization for strengthening causal inference in observational studies: Application to gene-by-environment interactions," *Perspectives on Psychological Science*, 2010, 5:527–545.

10. Debbie A. Lawlor, Roger M. Harbord, Jonathan A. C. Sterne, Nic J. Timpson, and George Davey Smith, "Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology," *Statistics in Medicine*, 2008, 27:1133–1163.

11. Karen Lutfey and Jeremy Freese, "Toward some fundamentals of fundamental causality: Socioeconomic status and health in the routine clinic visit for diabetes," *American Journal of Sociology*, 2005, 110:1326–1372.

12. Tyler J. Vander Weele, "Genetic self-knowledge and the future of epidemiologic confounding," *American Journal of Human Genetics*, 2010, 87(2):168–172.

https://doi.org/10.2990/30_2_88 Published online by Cambridge University Press