FAQs about “GWAS of 126,559 individuals identifies genetic variants associated with educational attainment”

The document was prepared by Daniel Benjamin, Mary Carmichael, David Cesarini, Christopher F. Chabris, Philipp D. Koellinger, David I. Laibson, Michelle N. Meyer, and Peter M. Visscher.

For clarifications or additional questions, please contact: Daniel Benjamin (daniel.benjamin@gmail.com), David Cesarini (dac12@nyu.edu), Philipp Koellinger (p.d.koellinger@vu.nl), or Peter Visscher (peter.visscher@uq.edu.au).

This study is the initial project of the Social Science Genetic Association Consortium (SSGAC). What is the SSGAC?

The SSGAC is a research infrastructure designed to stimulate dialogue and cooperation between medical researchers and social scientists. The SSGAC facilitates collaborative research that seeks to identify associations between specific genetic markers (segments of DNA) and behavioral traits, such as preferences, personality and social-science outcomes. One major impetus for the formation of the SSGAC was the growing recognition that most effects of individual genetic markers on behavioral traits are very small, and that, consequently, very large samples are required to accurately detect them. Several years ago medical researchers responded to a similar recognition—that most effects of individual genetic markers on complex diseases are very small—by forming research consortia in which groups collaborate by pooling results across many datasets. The SSGAC is an attempt to encourage analogous pooling among social-science geneticists and is organized under the auspices of the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE), a successful medical consortium. The SSGAC was founded by three social scientists (Daniel Benjamin, David Cesarini, and Philipp Koellinger) who are excited about the potentially transformative impact that genetic data could have on the social sciences, yet troubled by how current approaches are not bearing fruit. The Advisory Board for the SSGAC is composed of prominent researchers representing various disciplines: Dalton Conley (Sociology, New York University), George Davey-Smith (Epidemiology, University of Bristol), Albert Hofman (Epidemiology, Erasmus University), Robert Krueger (Psychology, University of Minnesota), David Laibson (Economics, Harvard), Sarah Medland (Statistical Genetics, Queensland Institute of Medical Research), Michelle Meyer (Bioethics, Harvard Law School), and Peter Visscher (Statistical Genetics, University of Queensland).

Why do you say current approaches are not bearing fruit?

An important theme in our earlier work has been to point out that most existing studies in social-science genetics that report genetic associations with behavioral traits have serious methodological limitations. The extent of the problems with existing study designs is increasingly recognized by the research community. Indeed, a leading journal for the genetics of behavioral traits recently issued an editorial statement that includes the following passage (Hewitt, 2012):

“The literature on candidate gene associations is full of reports that have not stood up to rigorous replication. This is the case both for straightforward main effects and for candidate gene-by-environment interactions (Duncan and Keller 2011). As a result, the psychiatric and behavior genetics literature has become confusing and it now seems likely that many of the published findings of the last decade are wrong or misleading and have not contributed to real advances in knowledge. The reasons for this are complex, but include the likelihood that effect sizes of individual polymorphisms are small, that studies have therefore been underpowered, and that multiple hypotheses and methods of analysis have been explored; these conditions will result in an unacceptably high proportion of false findings (Ioannidis 2005).”

Despite the growing awareness of these problems in the research community, studies in social-science genetics in which researchers report large genetic associations with behavioral traits continue to be published (and often receive uncritical media attention).

The evidence is now accumulating that many of these original studies fail to replicate (Benjamin et al. 2012; Chabris et al. 2012). In our view, one of the most important reasons why existing work has generated unreliable results is that their sample sizes were far too small, given that the true effects of individual genetic markers on behavioral traits are tiny.

Why is it important to know that effect sizes for behavioral traits are small?

The effect size, or strength of the relationship between an individual genetic variant and behavioral trait, determines which research strategies will succeed and which will fail. Virtually all existing studies in social-science genetics use sample sizes in the range 100-2,000. The tiny effect sizes for genetic variants identified in our study suggest that studies seeking to identify genetic influences on behavioral traits should include at least tens of thousands of research participants in order to generate accurate results.

If effect sizes are so small, why bother studying them?

Despite the tiny effect sizes, identifying genetic variants related to behavioral traits could be useful to social scientists for a number of reasons. Here we give two examples.

First, even if a genetic variant has a very small effect, identifying it may lead to insights regarding the underlying biological pathways. To take an example from medicine, genetic variants in the *LMTK2* (lemur tyrosine kinase 2) gene have small effects on an individual’s predisposition to prostate cancer. Nonetheless, knowing that this gene is involved can point scientists toward studying what the gene does, which may end up teaching us something critical about the pathology of prostate cancer.

Second, even if an *individual* genetic variant has a very small effect, many genetic variants taken together may have more predictive power. In one of our analyses, we estimate that when it becomes possible to analyze data from 1 million or more individuals—which is still several years away—many genetic variants taken together will be able to capture 15% of the variation across individuals in educational attainment. This amount of predictive power is still too low to be relevant for predicting any one person's educational attainment, but it would be useful for controlling for genetic factors when studying the effect of an education-promoting policy. When social scientists study expensive policy interventions, such as providing preschool to disadvantaged children, controlling for as many factors as possible can help generate more accurate estimates of the effectiveness of the policy.

What did you do in this particular study on educational attainment?

We conducted what is called a genome-wide association study (GWAS) by combining data from 54 cohorts with a total of about 125,000 individuals—a sample size about 10 times larger than any previous study of any social-scientific outcome. We study approximately 2 million genetic variants called single nucleotide polymorphisms, or SNPs. SNPs are the smallest and most common type of genetic variants (ways in which the genomes of people can differ), but they are not the only kind of genetic variation.

To create a harmonized measure of educational attainment across cohorts, we coded study-specific measures using the International Standard Classification of Education (ISCED) scale. This yielded two measurements of educational attainment: (1) a quantitative variable defined as an individual's years of schooling (EduYears); and (2) a binary variable for whether or not an individual had completed college (College). We sought to only include individuals in the study who were likely to have completed their formal schooling.

We conducted the study in two stages.

In the "discovery phase," we tested each of the ~2,000,000 SNPs for association with educational attainment using a sample of approximately 100,000 subjects from 42 cohorts.

In the "replication phase," we sought to replicate our findings using an additional (approximately) 25,000 subjects from 12 additional cohorts that became available after the discovery phase was completed.

In the "combined phase," we conducted the same GWAS analysis with the combined data from both the discovery and replication phases.

Finally, using the results of the combined meta-analyses of the discovery and replication cohorts, we conducted a number of additional analyses designed thoroughly investigate the genetic variants we identified as most strongly associated with educational attainment.

What did you find?

In the discovery phase, we found three genetic variants significantly associated with educational attainment—one associated with years of education, and two with college completion.

In the replication phase, we found that all three significant variants were also associated with educational attainment in the 12 independent cohorts (and that the strength of the relationship between these genetic variants and educational attainment was similar between the discovery and replication phases). This replication is important because it represents a separate test of whether these three variants were truly associated with education. If they were not—if they were just false positives, or chance findings due to statistical noise in the discovery phase—it is very unlikely that all three associations would replicate in an independent sample.

The observed effect sizes of these SNPs are small. For example, the size of the effect of the SNP identified in the analysis EduYears suggests that the difference between people with 0 and 2 copies of the variant predicting more education is about 2 months of education. One way to put the figure into perspective is to compare our findings to those in the medical and anthropometric literature. It is well-known that the SNP with the largest effect on human height explains about 0.4% of the variation (that is, extent to which individuals differ) in height. By contrast, the SNP we identified explains about 0.02% of the variation in years of education—only one-twentieth of the effect size found for height.

<u>In the combined phase</u>, we found seven additional significant SNPs (three for college completion and four for years of education). Of these seven, three are physically very close in the genome to the replicated SNPs and probably represent the same underlying genetic effect. The remaining four are in other places in the genome and warrant replication attempts in further research.

We next used *all* the genetic data, including the significant and non-significant SNPs, to create a "polygenic score" that represents our best prediction of each person's educational attainment based on their SNPs. We found that using all the genetic data in this way, we were able to explain a little over 2% of the variation between people in how many years of education they attained.

To explore one of the many possible channels through which the SNPs may be impacting educational attainment, we examined whether the same genetic risk score predicted cognitive function in a sample of Swedish military conscripts. (This analysis was done with just one of the cohorts, from Sweden, because that was the only one for which we had cognitive test data from all the participants.) We found that the score explained about 3% of the variance in cognitive function.

What does this study reveal about the biological pathways potentially affecting educational attainment?

Before explaining what we believe we have learned, it is important to understand three key limitations to inferring underlying biological influences from the current study.

First, educational attainment, like every complex behavior, is influenced by a large number of both genetic and environmental factors, and this study focuses on only a tiny piece of the puzzle. This study examines common SNPs, only one of many forms of genetic differences between individuals, and does not include specific measures of environmental factors. Therefore, it is likely that there are additional sources of genetic variation that remain to be discovered, as well as effects of SNPs that may differ based on environmental conditions (such as access to formal schooling). These other genetic effects, environmental effects, and interactions between genetic and environmental factors, are not discoverable from the current approach.

Second, the current study is limited to identifying regions of the genome—sometimes very large ones that include many different genes—that are statistically correlated with educational attainment; we cannot say for sure which specific variants or genes, if any, actually *cause* differences in educational attainment, or the mechanisms through which they may act. Even if we knew the true "causal" variants, their direct effects would likely be on health, cognitive function, or personality, which only have a downstream (i.e., eventual and indirect) impact on an individual's likelihood or ability to remain in a formal schooling environment.

Third, only the existing scientific literature and bioinformatics tools have been used to prioritize plausible biological candidates in the identified genomic regions. Thus, the accuracy of our conclusions are limited by current knowledge of genetic functions. As understanding of the underlying biological function of different genes develops over time, so will our ability to interpret findings arising from studies such as this one that finds statistical associations between genetic variants and human behavior.

For these reasons, strong conclusions about underlying biological mechanisms would be premature. We view our findings as providing a limited number of testable hypotheses for future research, providing a starting point for future studies to investigate in more detail a substantially narrowed field of likely genetic influences on educational attainment.

With those caveats in mind, we believe that our findings regarding biological pathways are strongly suggestive. Some identified genomic regions have previously been shown to affect cognitive functions or long-term memory in model organisms, or are predicted by bioinformatics analyses to influence neuron-related pathways (like cell differentiation, neurotransmitter signaling, and regulation of transmission of nerve impulse). In several cases, genes located in different genomic regions identified as associated with educational attainment appear to function as part of the same biological pathways. This convergence of evidence suggests that our findings are consistent with each other, which increases our confidence that the results may be accurate (rather than a spurious findings due to chance). Nonetheless, additional research, including experimental and molecular methods, will be required to investigate which, if any, of these candidate biological pathways have real, causal influences on behaviors related to educational outcomes.

What are the contributions of your study?

In addition to the most direct contribution—identifying several SNPs associated with educational attainment—we believe that our study also makes a number of other contributions. Here we list two.

First, it provides a methodological template that social-science genetics scholars could follow in future work. Genetic-association studies of behavioral traits have so far focused mostly on outcomes such as cognitive function and personality and have so far failed to document many associations that replicate consistently. The GWAS conducted to date have not found any genetic variants that are reliably associated with these phenotypes. One common view is that the appropriate response to the null findings in such studies is for researchers to gather better measures of the phenotypes and their facets in more environmentally homogenous samples—for example, by giving more in-depth personality tests to people who are all members of an isolated population. Our findings of replicable genetic associations for educational attainment demonstrate the feasibility of a complementary approach: identify an outcome variable that is measured with less precision, and therefore theoretically less directly connected to underlying genetic influences, but is available in a much larger GWAS sample and can be measured at a fraction of the cost. In our study, this corresponds to measuring educational attainment, which requires just a couple of survey questions, rather than giving a lengthy battery of cognitive function tests, which may require much more time or even expert interviewers to interact with each participant. The SNPs we identify could be used in follow-up research that directly studies precursors to educational attainment, such as cognitive function or personality traits such as perseverance.

Second, our study provides new evidence about what effect sizes can be expected for associations between individual genetic variants and complex behavioral traits. These effect-size estimates—which are roughly one-twentieth as large as those found for human height, a complex physical trait—will be useful for determining what sample sizes should be used in social-science genetics research. The estimates are also useful for assessing how much to trust existing reports of genetic associations from smaller samples.

These and other contributions—such as our exploration of potential biological pathways that might underlie the associations we observe—are likely to trigger follow-up research in various disciplines (such as the social sciences, epidemiology, and biology).

What policy lessons do you draw from this study?

None. *Any* practical response—genetic or environmental, individual or policy-level—to this or similar research would be extremely premature. In this respect, our study is no different from most GWAS of complex medical outcomes. In medical GWAS research, it is well understood that the known genetic variants are not yet predictive enough of complex diseases to be useful for assessing the risk to any given individual. Our paper shows that most genetic effects on behavioral are probably even smaller and more diffuse.

Did you find "the gene" for educational attainment?

No. We did not find "the gene" for educational attainment, cognitive function—or anything else. Educational attainment, like most complex behaviors and outcomes, is influenced by myriad genes, each with effects that are likely to be tiny (as well as a huge host of environmental factors).

Does this study show that an individual's level of education is determined at birth?

No, and this is probably one of the most common misconceptions about genetic research. Even if it were true—and it is certainly not—that genetic factors accounted for all of the variation across individuals in educational attainment, it would not follow that an individual's education is "determined" at birth. There are two distinct reasons for this.

First, some genetic effects may operate through environmental channels. For example, consider body mass index (BMI). Genetic factors may impact a person's BMI indirectly through genetic influences on food preferences, which in turn impact caloric intake and thus BMI. In this case, changes to the intermediate environmental channels can have drastic effects on the outcome. For example, lack of access to certain foods (or higher taxes on those foods) could cause substantial differences between an individual's genetically-predicted "propensity" and actual observed outcome in BMI. Similarly, genetic variants that are associated with educational attainment under current environmental conditions may no longer be associated if environmental conditions or policies change.

Second, even if the genetic effects operated entirely through non-environmental mechanisms that are difficult to modify, there could still exist powerful environmental interventions that do not contribute to differences across individuals in the current population. In a famous example attributable to the economist Arthur Goldberger, even if all the variation in eyesight were due to genes, there could still be enormous benefits from introducing eyeglasses. Similarly, policies such as a required minimum number of years of education and help for individuals with learning disabilities can have a drastic impact on educational outcomes for individuals who otherwise may be less likely to participate in formal schooling.

We found a handful of SNPs associated with educational attainment, each of which we estimate to have only a small effect on that outcome. Even if we had found that a single gene has a very large effect on educational attainment, that finding would be perfectly consistent with the possibility that environmental factors or interventions can modify or even cancel out this influence; traits that have a genetic component, even a large genetic component, are not necessarily immutable. For instance, the metabolic disorder phenylketonuria (PKU) is caused by gene mutations that prevent the carrier from metabolizing phenylalanine, an amino acid. Without environmental intervention, PKU leads to mental retardation and other serious medical problems. But through early genetic detection and an environmental intervention—namely, maintaining a special diet free of phenylalanine, monitoring of their protein levels and daily medication—individuals with PKU can lead lives with normal cognitive function and life expectancy.

How are your findings relevant for health?

There is a well-known relationship between educational attainment and health outcomes, and this connection was one important motivation for our study. Indeed, some of the genetic variants we identify may be associated with education because of their effects on health (which, in turn, could impact education). Some of the genes identified in our analyses have been previously implicated in studies of common diseases such as inflammatory bowel disease and rheumatoid arthritis. Our analyses also identified genomic regions that have been linked to brain and central nervous system development in non-human animals, including mice and zebrafish. This suggests that the regions we identify as associated with educational attainment are promising candidates for future exploration, as they may turn out to play a role in both health and cognition.

The "polygenic score" that we estimate—which aggregates the effects of many individual genetic variants—may also prove useful for health researchers. Indeed, several groups of medical researchers have already expressed interest in examining the predictive utility of the score for the health conditions that they study (including dyslexia and psychiatric disorders). By making the results of our analyses publicly available on www.ssgac.org, we hope that other researchers will make use of our findings to develop and test hypotheses for how genetic factors may jointly influence educational attainment and related health and cognitive outcomes.

Could this kind of research lead to discrimination against, or stigmatization of, people with the relevant genotypes, and doesn't that make the research ethically suspect?

It is always possible that research results may be used inappropriately by others, either willfully or due to misinterpretation. Those of us working in this area have an obligation to be vigilant about our use of language that may be susceptible to misinterpretation, and to communicate clearly not only potential benefits of the research that motivate us to do the work, but also its limitations. For a variety of reasons, in general we do not think that the best response to the possibility that useful knowledge might be misused is to refrain from producing the knowledge.

Behavioral genetics research, including studies of the relationships between genes and a variety of cognitive traits, is already being conducted and will continue to be conducted. Common misunderstandings of the results of genetic research are well documented, yet careless discussions of results and publication bias remain problems within the research community across many disciplines and fields. In this context, responsible researchers who are committed to developing, implementing, and spreading best practices for conducting and communicating potentially controversial research, including behavioral genetics research, should participate in the development of this body of knowledge, rather than abstaining from it and hoping for the best.

The results of behavioral genetics research can be used in positive, as well as negative, ways. One of the benefits of properly conducted behavioral genetics research is that it has made clear the limits of deterministic views of complex traits by establishing accurate upper bounds for effect sizes and prediction accuracy—thus perhaps making discrimination and stigmatization *less* likely in the future. Existing claims of genetic associations with complex social-science traits have reported widely varying effect sizes—most of them purporting to explain more than one hundred times as much variance as did the genetic variants we found in this study. Our evidence, from a much larger sample, suggests instead that individual SNPs associated with educational attainment have about one-twentieth as large effects as do SNPs for complex physical traits that are also influenced by many separate genes, such as height.

More ambitiously and longer-term—but quite plausibly—learning more about the genetic influences on educational attainment (or cognition) may lead to effective environmental interventions, as it did in the case of PKU (see "Does this study show that an individual's level of education is determined at birth?"). As noted above (see "If effect sizes are so small, why bother studying them?"), in order to study effective policies to reduce gaps in educational attainment or similar disparities, it is helpful to account for any genetic effects on those outcomes. In addition, identifying genetic variants that contribute to differences in educational attainment may lead to insights regarding the biological pathways underlying that outcome, and may provide a firm foundation for research on interactions between genetic and environmental influences.

In order to realize these benefits, however, behavioral genetics research must be carefully and responsibly conducted and communicated. Responsible behavioral genetics research, in our view, includes sound methodology and analysis of data; a commitment to publish all results, including any negative results; and transparent, complete reporting of methodology and findings in publications, presentations, and communications with the media and the public, including particular vigilance regarding what the results do—and do not—show (hence, these FAQs).

In summary, we agree with the Nuffield Council on Bioethics, which concluded in a 2002 report on behavioral genetics research, including research on cognitive function, that "research in behavioural genetics has the potential to advance our understanding of human behaviour and that the research can therefore be justified," and that "researchers and those who report research have a duty to communicate findings in a responsible manner."

References

Benjamin, Daniel J., David Cesarini, Christopher F. Chabris, Edward L. Glaeser, David I. Laibson, Vilmundur Guðnason, Tamara B. Harris, Lenore J. Launer, Shaun Purcell, Albert Vernon Smith, Magnus Johannesson, Patrik K.E. Magnusson, Jonathan P. Beauchamp, Nicholas A. Christakis, Craig S. Atwood, Benjamin Hebert, Jeremy Freese, Robert M. Hauser, Taissa S. Hauser, Alexander Grankvist, Christina M. Hultman, and Paul Lichtenstein (2012). "The Promises and Pitfalls of Genoeconomics." *Annual Review of Economics*, 4, 627-662.

Chabris, Christopher F., Benjamin M. Hebert, Daniel J. Benjamin, Jonathan P. Beauchamp, David Cesarini, Matthijs J.H.M. van der Loos, Magnus Johannesson, Patrik K.E. Magnusson, Paul Lichtenstein, Craig S. Atwood, Jeremy Freese, Taissa S. Hauser, Robert M. Hauser, Nicholas A. Christakis, and David Laibson (2012). "Most Published Genetic Associations with General Intelligence Are Probably False Positives." *Psychological Science*. doi:10.1177/0956797611435528