Frequently Asked Questions (FAQ)

This description of the *Nature* paper "*Genome-wide association study identifies 74 loci associated with educational attainment*" includes the following information:

1. Background: authorship, goals, definition of "educational attainment," previous research
2. Study design and results: genes, variants, and biology linked to educational attainment
3. Social implications of the study: potential use in medical research and in policy
4. Appendices: quality-control measures, further reading and references

The document was prepared by Daniel J. Benjamin, David Cesarini, Christopher F. Chabris, Philipp D. Koellinger, David Laibson, Michelle N. Meyer, and Peter M. Visscher. It draws from and builds on the FAQs for earlier SSGAC papers. For clarifications or additional questions, please contact (daniel.benjamin@gmail.com).

1. Background

Who conducted this study? What was the group's overarching goal?

The authors are members of the Social Science Genetic Association Consortium (SSGAC), a multi-institutional research group that aims to draw statistically rigorous links between genetic variants—for instance, base-pairs of DNA that vary across people—and social science variables such as behavior, preferences, and personality. The SSGAC was formed in 2011 to overcome a specific set of scientific challenges. First, most traits and behaviors are influenced by hundreds or thousands of genetic variants, and almost all of these genetic variants have extremely weak effects on their own (though, when combined, their collective effects can be meaningful). Second, to rigorously identify such variants, scientists must study hundreds of thousands of people, and therefore a promising strategy is for many investigators to pool their data into one large study. This approach has borne considerable fruit in medical genomics; recent successful studies of the genetics of autism (Gaugler et al., 2014), schizophrenia (Ripke et al., 2014), and many other diseases and conditions would not have been possible without large consortia in which members shared their data. The SSGAC is an attempt to recapitulate this research model for understanding genetic associations with non-medical traits.

The SSGAC is organized as a working group of the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE), a successful medical consortium. It was founded by three social scientists—Daniel Benjamin (University of Southern California), David Cesarini (New York University), and Philipp Koellinger (Vrije Universiteit Amsterdam)—who believe that genetic data could have a substantial positive impact on research in the social sciences. 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), Tõnu Esko (Molecular Genetics, Broad Institute and Estonian Genome Center), Albert Hofman (Epidemiology, Harvard), Robert Krueger (Psychology, University of Minnesota), David Laibson (Economics, Harvard), Sarah Medland (Statistical

Genetics, QIMR Berghofer Medical Research Institute), Michelle Meyer (Bioethics, Clarkson University and Icahn School of Medicine at Mount Sinai), and Peter Visscher (Statistical Genetics, University of Queensland).

The SSGAC is committed to the principles of reproducibility and transparency. Prior to conducting genetic association studies, power calculations are carried out to determine the necessary sample size for the analysis (assuming realistically small effect sizes associated with individual genetic variants). These, together with an analysis plan, are posted on the Open Science Framework's preregistration website. Major publications are usually accompanied by a FAQ document (such as this one). The FAQ document is written to communicate to the public what was found and what can and cannot be concluded from the research findings.

The SSGAC's first major project was a genome-wide association study (GWAS) on educational attainment—similar to this study but using a much smaller sample—published in *Science* (Rietveld et al., 2013). The study is summarized in a FAQ posted on the SSGAC website: http://www.thessgac.org/#!faqs/e0udx. Subsequent papers have been published in *Proceedings of the National Academy of Sciences* and *Psychological Science,* among other journals.

The current study focuses on a variable called "educational attainment." What is educational attainment?

Educational attainment is the amount of formal education a person completes. People vary considerably in how much education they complete. Education is recognized throughout the social and medical sciences as an important predictor of many other life outcomes, such as income, occupation, and health (e.g., Ross and Wu, 1995; Cutler and Lleras-Muney, 2008).

What was already known about the genetics of educational attainment prior to this study?

Decades of twin and family studies have found that one reason people differ in educational attainment is that they differ genetically. For example, several studies have found that identical twins are more similar to each other in their educational attainment than fraternal twins are to each other (Taubman, 1976; Branigan, McCallum, and Freese, 2013). Nonetheless, educational attainment is *also* strongly influenced by social and other environmental factors.

Recent research using molecular genetic data—in other words, using data that measures each person's DNA and identifies variation at the molecular level across people—has similarly found that genetic factors play a role, accounting for at least 20% of variation in educational attainment (e.g., Rietveld et al., 2013).

These findings imply that there are genetic variants associated statistically with more educational attainment (people who carry these variants will tend on average to complete more formal education) and genetic variants associated statistically with less educational attainment (people who carry these variants

will tend on average to complete less formal education). It is important to emphasize that these associations represent average tendencies in the population—not pre-determined outcomes for each person. It is likely that many genetic variants matter more or less depending on environmental context (such as a country's school system and the quality of an individual's school).

In the SSGAC's first major publication (Rietveld et al., 2013), we conducted a genome-wide association study (GWAS) in a sample of roughly 100,000 individuals and identified three genetic variants statistically associated with educational attainment. In the same paper and subsequent work (Rietveld et al., 2014a), we verified that the associations with those variants were replicated in separate samples of individuals (25,000 and 35,000 people, respectively). There were two key takeaways from this work:

(1) A GWAS approach can identify specific genetic variants statistically associated with social-science variables if the study is conducted in large enough samples (at least one hundred thousand people).

(2) A specific genetic variant that is associated with a *social science* variable is likely to have much smaller predictive power for that trait than a specific genetic variant that is associated with a bio-medical outcome (Chabris et al., 2015). For example, the known genetic variant with the largest effect on height predicts 0.4% of the variation in height across individuals in the sample, whereas the three variants identified by Rietveld et al. (2013) each predict only 0.02% of the variation in educational attainment in the sample.

2. Study design and results

What did you do in this paper? How was the study designed?

The central contribution of the paper is a genome-wide association study (GWAS) of about 300,000 people (based on combined results from 64 separate analyses conducted in cohorts of participants from 15 different countries). This is by far the largest sample size ever studied for genetic associations with any social science outcome. We included only individuals of European descent to reduce statistical confounds that otherwise arise from studying ethnically diverse populations (see the discussion of population stratification in Appendix 1). For each person in our data, we analyzed approximately 9 million genetic variants called single nucleotide polymorphisms, or SNPs. (SNPs are the most common type of genetic variant (a way in which the genomes of people can differ), but they are not the only type of genetic variant.)

We subsequently gained access to a large, independent sample of roughly 110,000 individuals (data from the U.K. Biobank). We used this new dataset to replicate the genetic associations that we initially reported.

In the remainder of the paper, we used the findings from the GWAS for a range of additional analyses that explored (among other things) the biological pathways associated with genetic variants of interest, and the genetic overlap between educational attainment and other outcomes such as Alzheimer's disease.

What did you find in the GWAS?

In our "discovery sample" (of roughly 300,000 people), we found 74 SNPs associated with educational attainment. These include the 3 genetic variants identified in our earlier study. In our "replication sample" (of roughly 110,000 additional people from the U.K. Biobank), these findings held up extremely well. For example, in our replication sample, 72 of the 74 SNPs were associated with educational attainment with the same sign as in the discovery sample (i.e., those that were associated with higher educational attainment in the first sample also did so in the replication sample, and those that were associated with lower educational attainment in the first sample also did so in the replication sample).

As a group, the 74 SNPs explain 0.43% of the variation in educational attainment across individuals in the sample. Individually, each of the 74 SNPs had an extremely small influence on educational attainment. The variant with the strongest association explained only 0.035% of the variation in educational attainment. Put another way, the difference between people with 0 and 2 copies of this genetic variant predicts (on average) about 9 extra weeks of schooling.

How do we reconcile our finding that the predictive power of each individual SNP association is extremely small and the finding from much previous work that at least 20% of the overall variation in educational attainment is associated with genetic factors? The two findings taken together imply that the genetic associations with educational attainment result from the cumulative effects of at least thousands (probably millions) of different genetic variants, not just a few.

This is not a surprise: educational attainment is a complex phenomenon, and our study focuses on only a tiny piece of the puzzle. In this paper, we only examine one type of genetic variant (SNPs), we consider only one of many forms of genetic difference among individuals, and we conduct only preliminary and exploratory analyses of how the effects of genetic variants differ depending on environmental conditions. There are substantial additional sources of molecular genetic variation that remain to be discovered. These other genetic effects, environmental effects, and their interactions are important topics of active research, and of future work by the SSGAC.

Can you use the results in this paper to meaningfully predict a particular person's educational attainment?

No.

Each *individual* genetic variant has a very small effect. It is true that many genetic variants combined together into an index can explain much more of the variation across individuals. Such an index is called a "polygenic score." However, when we construct a polygenic score using all ~9 million SNPs in our data, we still find that on average the polygenic score explains only 3.2% of the variation across individuals. It will likely be possible to construct a polygenic score whose explanatory power is closer to 20% as the available sample sizes for GWAS get larger.

Our existing polygenic score is not an accurate predictor of any individual's educational attainment. Even a (currently non-existent) polygenic score that could account for 20% of variation in educational attainment would pale in comparison to other scientific predictors. For comparison, professional weather forecasts correctly predict about 95% of the variation in day-to-day temperatures. Weather forecasters are vastly more accurate forecasters than social science geneticists will ever be.

Yet a polygenic score, based on our study, that reflects 3.2% of the variation in educational attainment, is large enough to be useful in *social science studies*, which focus on average or aggregated behavior in the population (not individual outcomes). Indeed, with 80% statistical power (the conventional threshold for adequate power), the effect of our polygenic score can be detected in a study with 250 individuals (notably, many orders of magnitude smaller than the sample sizes we needed to be able to construct the score). Therefore, the polygenic score provided by our study can be useful in social science studies that have at least 250 participants and in which the participants' genomes have been measured.

Are the variants associated with higher educational attainment in your study also associated with other outcomes?

In the data we analyzed for this paper, we find that on average SNPs associated with higher educational attainment are also associated with increased cognitive performance and intracranial volume, increased risk of bipolar disorder, decreased risk of Alzheimer's disease, and lower neuroticism. These results highlight the potential relevance of our results for medical research, but future research is needed to shed light on the reasons why genetic variants are shared in common across these traits.

What do your results tell us about human biology and brain development?

We can draw inferences about biological pathways using computational methods that examine whether genes known to be involved in particular biological systems are especially likely to be associated with educational attainment.

We found that genes identified by our analyses tend to be strongly active in the brain, especially prenatally, and are especially likely to be involved in neural development. Moreover, the specific SNPs we identify tend to be in regions of the genome believed to be involved in regulation of gene activity in the fetal brain.

It is not surprising that genes may influence educational attainment in part because of their effects on brain development. Cognitive abilities and personality traits (such as conscientiousness and resilience) that matter for school performance may be partially reflected in how the brain is organized. It is perhaps more surprising that our study of educational attainment generates a biological picture of brain development that is clearer than those generated by previous GWAS that focused directly on brain structures. We believe that the relative clarity of the biological picture we observe is due to the large sample size of our study, which afforded us greater statistical power than previous GWAS. Since it will remain much easier to measure educational attainment than to conduct brain scans in large samples of

individuals, we believe that GWAS of educational attainment will continue to play a useful role in understanding the biology of brain development.

Did you find "the genes for" educational attainment?

No. We did not find "the genes for" educational attainment—or for anything else. Characterizing the results this way is misleading for many reasons. First, educational attainment is primarily determined by environmental factors, not genes. Second, the explanatory power of each individual genetic variant that we identify is extremely small. Our results show that genetic associations with educational attainment are comprised of thousands, or even millions, of genetic variants, each of which has a tiny effect size. Third, environmental factors are likely to increase or decrease the impact of specific genetic variants. Indeed, in the current paper we report exploratory analyses that provide suggestive evidence of such gene-environment interactions. Finally, genes do not affect educational attainment directly. Rather, genes that are associated with educational attainment might influence many different biological factors that in turn affect psychological characteristics that finally influence educational attainment.

Does this study show that an individual's level of educational attainment is determined at conception?

No. Even if it were true that genetic factors accounted for *all* of the differences among individuals in educational attainment (which they certainly do not), it would *still* not follow that an individual's number of years of formal schooling is "determined" at conception. There are at least three reasons for this:

First, some genetic effects may operate through environmental channels. As an illustrative example, suppose—hypothetically—that the genetic variants we identified help students to memorize and, as a result, to become better at taking tests that rely on memorization. In this example, changes to the intermediate environmental channels—the type of tests administered in schools—could have drastic effects on individuals' educational attainment, even though their genetic variants would not have changed. A genetic association with educational attainment might not be found at all if schools did not use tests that rely on memorization. More generally, the genetic associations that we found might not apply as strongly if the education system were organized differently than it is at present.

Second, even if the genetic associations with educational attainment operate entirely through non-environmental mechanisms that are difficult to modify (such as direct influences on the formation of neurons in the brain and the chemical interactions among them), there could still exist powerful environmental interventions that could change the genetic relationships. In a famous example suggested by the economist Arthur Goldberger, even if all variation in unaided eyesight were due to genes, there would 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 increase educational attainment in the entire population and/or reduce differences among individuals.

Third, even if the genetic effects on educational attainment were not influenced by changes in the

environment, those environmental changes themselves could still have a major impact on the educational attainment of the population as a whole. For example, if young children were given more nutritious diets, then everyone's school performance might improve, and college graduation rates might increase. By analogy, 80%-90% of the variation across individuals in height is due to genetic factors. Yet the current generation of people is much taller than past generations, entirely due to changes in the environment such as improved nutrition.

Can environmental factors modify the effects of the specific genetic variants you identified?

We believe the answer is yes, and we report some exploratory analyses of this question in the paper. We examined a sample of Swedish individuals born between 1929 and 1958. During the 1950s and 1960s, when many of these individuals were in school, Sweden (like many other European countries) introduced a comprehensive new schooling system that extended mandatory schooling from seven to nine years, eliminated the lower level in secondary school, and postponed ability tracking from around age 10 until age 16. Another set of reforms sought to increase equality of outcomes and opportunity by increasing the availability of high schools, colleges, and universities.

We find that the association between educational attainment and our polygenic score (an index of the genetic variants in our data) is only about half as large among Swedish individuals born in the late 1950s compared with those born in the early 1930s. This finding is consistent with the possibility that the Swedish reforms reduced the effects of genetic variants in generating differences in educational attainment. While the analyses we report are exploratory, we believe that one contribution of our paper is to pave the way for more in-depth studies of such gene-environment interactions.

3. Social implications of the study: potential use in medical research and in policy

What policy lessons or practical advice do you draw from this study?

None whatsoever. *Any* practical response—individual or policy-level—to this or similar research would be extremely premature. In this respect, our study is no different from genome-wide association studies (GWAS) of complex medical outcomes. In medical GWAS research, it is well understood that identifying genetic variants that affect disease risk is merely a first step toward understanding the underlying biology of that disease. It is not sufficient to assess risk for any specific individual. It is not appropriate to base policies and practices on such assessments.

Do your findings have implications for health? Could they be used to advance medical research?

There is a well-known relationship between educational attainment and health outcomes, and this connection has been one motivation for our research. Indeed, some of the genetic variants we identify may be associated with educational attainment because they affect the health of people who carry them (which, in turn, could impact the amount of education a person receives). Our analyses of genetic overlap

suggest that some of the same genes that matter for educational attainment also matter for Alzheimer's disease, bipolar disorder, and schizophrenia. In previous work (Rietveld et al., 2014b), we found that an index of genetic variants associated with educational attainment had some predictive power for dementia in older individuals, and several groups of medical researchers have used the genetic variants and polygenic score identified by our earlier study on educational attainment (Rietveld et al., 2013) to study other health conditions including dyslexia and psychiatric disorders. By making the results of our analyses publicly available at www.thessgac.org, we hope to facilitate such research.

Could this kind of research lead to discrimination against, or stigmatization of, people with the relevant genetic variants?

There is always a risk that research will be misinterpreted or misused. In the case of behavioral genetic research, one risk is that findings may be misinterpreted (whether willfully or not) and misused to stigmatize or discriminate. One response to this risk is to abstain from conducting behavioral genetics research. In this case, however, we do not think that the best response to the possibility that useful knowledge might be misused is to refrain from producing that knowledge. Indeed, there are at least two major ethical problems with abstention.

First, behavioral genetics research, including studies of the relationships between genes and a variety of social and cognitive traits, is already being conducted and will continue to be conducted. Not all of this work involves appropriate scientific methods or transparent communication of results. In this context, researchers who are committed to developing, implementing, and spreading best practices for conducting and communicating potentially controversial research, including behavioral genetics research, arguably have an ethical responsibility to participate in the development of this body of knowledge—rather than abstain from it and hope for the best. In essence, we believe that we have an ethical duty to set the record straight.

For instance, 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, fail to replicate, and are likely to be false-positive findings (Benjamin et al., 2012; Chabris et al., 2012; Chabris et al., 2015). This same point was made in an editorial in *Behavior Genetics* (the leading journal for the genetics of behavioral traits), which stated that "it now seems likely that many of the published [behavior genetics] findings of the last decade are wrong or misleading and have not contributed to real advances in knowledge" (Hewitt, 2012). 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.

In our view, responsible behavioral genetics research 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, this FAQ document).

Second, one should not assume that behavioral genetics research carries only the potential to increase stigmatization. One benefit of recent behavioral genetics research is that it has clarified the *limits* of deterministic views of complex traits by establishing upper bounds for the amount of variation among individuals attributable to common genetic variants—thus perhaps making discrimination and stigmatization *less* likely in the future. Pre-existing claims of genetic associations with complex social-science outcomes have reported widely varying effect sizes, many of them purporting to explain ten to one hundred times as much of the variation across individuals as did the genetic variants we have found in this study and in our other studies.

The bottom line is that individual genetic variants have very little explanatory power for educational attainment, and even composite indexes of millions of genetic variants have too little explanatory power to usefully predict any individual's educational attainment.

4. Appendices

Appendix 1: Quality control measures

There are many potential pitfalls that can lead to spurious results in genome-wide association studies (GWAS). We took many precautions to guard against these pitfalls.

One potential source of spurious results is incomplete "quality control (QC)" of the genetic data. To avoid this problem, we used state-of-the-art QC protocols from medical genetics research (Winkler et al., 2014). We supplemented these protocols by developing and applying additional, more stringent QC filters.

Another potential source of spurious results is a confound known as "population stratification" (e.g., Hamer and Sirota, 2000). To illustrate, suppose we were conducting a GWAS on height. People from Northern Europe are on average taller than people from Southern Europe, and there are also small differences in how often certain genetic variants occur in Northern and Southern Europe. If we combine samples of Northern and Southern Europeans and perform a GWAS that ignores the regions the individuals come from, then we would find genetic associations for these variants. However, those associations would simply reflect the fact that the variants are correlated with a population (Northern or Southern Europe) and may actually have nothing to do with height.

In our study we were extremely careful to avoid population stratification as much as possible. At the outset, we restricted the study to individuals of European descent, since population stratification problems are more severe when including European-descent and non-European-descent individuals in the same sample. As is standard in GWAS on medical outcomes, we controlled for "principal components" of the genetic data in the analysis; these principal components capture the small genetic differences across populations, so controlling for them largely removes the spurious associations arising solely from these small differences. To eliminate even weak population stratification effects, we controlled for a larger number of principal components than is standard in GWAS on medical outcomes (ten rather than four).

After taking these steps to minimize population stratification, we conducted a number of analyses to assess how much population stratification still remained in our data. The results of these tests indicate that there is some, but not much.

We conducted additional tests to confirm that our GWAS results are not driven by this remaining population stratification. To do so, we used a subset of the individuals in our data, 5,506 sibling pairs (from five of the datasets that contributed to our study). The key idea underlying our tests is to examine if *differences* in genetic variants across siblings are associated with *differences* in the siblings' educational attainment. If so, then these associations cannot be the result of population stratification. The reason is that full siblings (from the same genetic parents) share their ancestry entirely, and therefore differences in their genetic variants cannot be due to being from different population groups (genetic differences between siblings are random). Unfortunately, because our sample of siblings (~11,000 individuals) is much smaller than our overall GWAS sample (~300,000 individuals), our estimates of the effects of the genetic variants within the sibling pairs are much noisier than in the GWAS. However, we *can* test whether the GWAS results are entirely due to population stratification, because if they were, then the sibling estimates would not line up with the GWAS estimates. In fact, we find that the within-family estimates are more similar to the GWAS estimates in both sign and magnitude than would be expected by chance. These results imply that our GWAS results are not solely due to population stratification.

Appendix 2: Additional reading and references

Benjamin DJ et al. (2012). The promises and pitfalls of genoeconomics. *Annu Rev Econom* **4**: 627-662.

Branigan AR, McCallum KJ, Freese J (2013). Variation in the heritability of educational attainment: An international meta-analysis. *Northwestern University Institute for Policy Research Working Paper. 13-09*.

Chabris C et al. (2012). Most Reported Genetic Associations with General Intelligence Are Probably False Positives. *Psychol Sci* **23**(11): 1314-1323.

Chabris C et al. (2015). The Fourth Law of Behavior Genetics. *Curr Dir Psychol Sci* **24**(4): 304-312.

Cutler DM, Lleras-Muney A (2008). Education and Health: Evaluating Theories and Evidence. In House J, Schoeni R, Kaplan G, Pollack H (Eds.), *Making Americans Healthier: Social and Economic Policy as Health Policy* (Russell Sage Foundation, New York).

Editorial (2013). Dangerous Work. *Nature* **502**(7469): 5-6.

Gaugler, T et al. (2014) Most genetic risk for autism resides with common variation. *Nat Genet* **46**(8): 881-885.

Hewitt J (2012). Editorial Policy on Candidate Gene Association and Candidate Gene-by-Environment Interaction Studies of Complex Traits. *Behav Genet* **42**(1): 1-2.

Hamer DH, Sirota L (2000). Beware the chopsticks gene. *Mol Psychiatry* **5**(1): 11–13.

Nuffield Council on Bioethics (2002). *Genetics and Human Behavior: the ethical context* (Nuffield Council on Bioethics: London).

Parens E, Appelbaum PS (2015). An Introduction to Thinking about Trustworthy Research into the Genetics of Intelligence. *Hastings Center Report* **45**(5): S2-S8.

Rietveld CA et al. (2013). GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* **340**(6139): 1467–1471.

Rietveld CA et al. (2104a). Replicability and Robustness of GWAS for Behavioral Traits. *Psychol Sci* **25**(11): 1975-1986.

Rietveld CA et al. (2014b). Common genetic variants associated with cognitive performance identified using proxy-phenotype method. *Proc Natl Acad Sci USA* **111**(38): 13790–13794.

Ripke, S et al. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**(7510): 421-427.

Ross CE, Wu C (1995). The links between education and health. *Am Sociol Rev* **60**(5): 719-745.

Taubman P (1976). Earnings, education, genetics, and environment. *J Hum Resour* **11**(4): 447-461.

Winkler TW et al (2014). Quality control and conduct of genome-wide association meta-analyses. *Nat Protoc* **9**(5): 1192–212.