Why Is It Hard to Find Genes that are Associated with Social Science Traits? Theoretical and Empirical Considerations

Christopher F. Chabris[1]†*
James J. Lee[2]†
Daniel J. Benjamin[3]
Jonathan P. Beauchamp[4]
Edward L. Glaeser[4]
Gregoire Borst[2]
Steven Pinker[2]
David I. Laibson[4]

1. Department of Psychology, Union College, Schenectady, NY, USA
2. Department of Psychology, Harvard University, Cambridge, MA, USA
3. Department of Economics, Cornell University, Ithaca, NY, USA
4. Department of Economics, Harvard University, Cambridge, MA, USA

† These authors contributed equally to the work.

* Address correspondence to:

Christopher F. Chabris
Department of Psychology
Union College
807 Union Street
Schenectady, NY 12308 USA
chabris@gmail.com

Abstract

Though most behavioral traits are moderately to highly heritable, the genes that influence them are elusive: many published genetic associations fail to replicate. With physical traits like eye color and skin pigmentation, in contrast, several genes with large effects have been discovered and replicated. We draw on R.A. Fisher's geometric model of adaptation to explain why traits of interest to behavioral scientists may have a genetic architecture featuring hundreds or thousands of alleles with tiny individual effects, rather than a few with large effects, and why such an architecture makes it difficult to find robust associations between traits and genes. In the absence of strong directional selection on a trait, alleles with large effect sizes will probably remain rare, and such a lack of strong directional selection is likely to characterize most traits currently of interest in social science. We evaluate these predictions via a genome-wide association study (GWAS) that carefully measured over 100 physical and behavioral traits with a sample size typical of candidate gene studies. While we replicated several known genetic associations with physical traits, we found only two associations with behavioral traits that met the nominal genome-wide significance threshold. We use the theory and findings to discuss (1) the challenges for social science genomics, particularly the likelihood that genes are connected to behavioral variation by lengthy, nonlinear, interactive causal chains; (2) the prospects for dealing with these challenges; and (3) the inherent tradeoff between two ways of meeting these challenges: increasing sample size and improving phenotype measurement.

Introduction

People differ in their intelligence, personality, and behavior, and a century of research in behavioral genetics leaves little doubt that some of this variation is caused by differences in their genomes.[1,2,3] Nonzero (and sometimes substantial) heritability of psychological traits has been consistently established in twin, adoption, and family studies that have often been massive in size. But beyond establishing that genes matter, such studies say little about the detailed genetic architecture of psychological traits, i.e., how many genetic polymorphisms affect a trait, how the polymorphisms interact, what they are, and what they do.

The recent advent of affordable genome-wide association studies (GWAS) offers the exciting opportunity to understand the genetic factors that influence psychological trait variation with far greater precision. It has the potential to uncover some of a given trait's genetic architecture, including the number, genomic locations, average effects, and allele frequencies of the DNA variants that affect the trait. Even an incomplete understanding of a trait's genetic architecture could prove a boon to social scientists for at least four reasons.[4]

First, the presence of genetic variants can be detected with high reliability. Thus, they may constitute direct measures of constructs that were previously regarded as only latent. For example, there is some evidence that a person's genotype for the SNP in FTO associated with body mass index (BMI) may indicate a preference for certain kinds of high-calorie foods,[5] and one might speculate that other genes may affect how much body weight is produced from a person's caloric intake. These genetic variants could then be used as variables of interest, or as controls, in other models, testing models of the causation of obesity that formerly could only appeal loosely to "genetic factors."

Second, the discovery of genetic associations may identify or clarify the actual biological mechanisms that underlie social and health behaviors. For example, a mechanistic role for the hormone oxytocin in trust-related behavior has been suggested by findings that variants in the oxytocin receptor gene (OXTR) are associated with differences in performance in a behavioral-economic trust game (albeit with mixed results so far).[6,7] And just as in medicine, where genetic discoveries have suggested new pathways for understanding and treating disease (e.g., Crohn's disease[8]), genetic discoveries may help social scientists decompose crude concepts like "risk aversion" and "time preference," both of which play roles in health behaviors, into biologically meaningful subcomponents.

Third, under very special circumstances, genetic variants could be used as instrumental variables that would identify causal relationships from non-experimental data. For such analysis to be valid, the allele must *reliably* and *exclusively* affect a specific biological trait (and no other biological traits). If these strong conditions are met, then one can use the random assignment (during meiosis) of each person's genotype at that allele as a natural experiment to test the hypothesis that the biological trait, in turn, causes variation in some behavioral phenotype. For example, Chen and colleagues showed that SNPs in ALDH2 that are known to increase alcohol metabolism are associated with decreased blood pressure.[9] This provides evidence that alcohol consumption in fact causes an increase in blood pressure—under the crucial, and perhaps implausible, condition that those SNPs are assumed not to also affect blood pressure through

some other channel. Other studies of this type have been published,[10] but it seems likely that the circumstances in which the instrumental variable approach can work are rare.

Fourth, knowledge of individuals' genotypes could help in targeting social-science interventions to those who stand to benefit from them the most—an application of concepts from personalized medicine to public health and policy. Such a benefit is particularly likely to help children, since their abilities and preferences are less developed and harder to measure. For example, children with genotypes that confer a susceptibility to dyslexia might be offered personalized educational resources from a very early age.

The leap in precision from GWAS, compared to twin studies, promises to help not just working social and behavioral scientists but anyone interested in the evolutionary history and adaptive pressures that shaped the human species and its variation. Not only does an individual's genome provide a partial recipe for the development of his or her unique phenome (set of phenotypes) forward in time, but our species' array of genomic data provides a trace of our collective evolutionary history backward in time. For example, once it was discovered that mutations in the gene FOXP2 could cause a severe developmental deficit in speech and language, comparative genomic analyses showed that this gene's sequence had changed at least twice since the separation of humans and chimpanzees from their common ancestor, that it has been a target of natural selection rather than a legacy of random drift, and that it is shared with Neanderthals—all relevant to venerable and hitherto nearly unresolvable debates on the evolution of language.

Despite the extraordinary promise of extending genetic research to behavioral traits, results so far of studies that have searched for genetic variants associated with these traits have been disappointing: No strong, replicable associations have been discovered. Most of the claims of genetic associations with such traits have turned out to be false positives, or at best vast overestimates of true effect sizes. Chabris et al. found that across three independent samples, only one of twelve published associations of particular genes with general intelligence replicated, and this one replicated in only one sample out of three. [11] Worse, the new samples were considerably larger than the originals, which suggests that all of these reports were probably false positives. Similarly, Benjamin, Cesarini, Chabris et al. found a SNP associated with educational attainment and cognitive function, but could not replicate it in three independent samples.[4] Benjamin et al. likewise found no significant associations with any of a set of traits involving economic and political behavior.[4] Finally, Beauchamp et al. conducted a GWAS of educational attainment (i.e., years of education completed) and found no hits that met conventional genome-wide significance levels; those that approached significance did not replicate in a second sample. [12]

Difficulty in finding specific genes that correlate with traits that are known to be heritable is not unique to the social sciences. It is also a problem in GWAS of medical traits such as psychiatric diagnoses and susceptibility to common diseases, and even with certain physical traits, such as height. Table 1 summarizes the heritabilities estimated from twin studies of medical, physical, and social science traits, based on three review articles and some recent publications in behavioral economics; it shows that the heritabilities of physical and psychological traits are similar and substantial.

The discrepancy between the high heritability of both physical and psychological traits, and the rarity of replicable discoveries of particular genes for those traits, has been dubbed the problem of "missing heritability".[13] One possible resolution of this paradox is that each of the genes associated with a trait explains only a minuscule fraction of the total genetic variance—and hence these genes are difficult to identify statistically—but there is a huge number of such genes, and the heritability estimate reflects their aggregated effects.

Zuk et al. suggested that the discrepancy between heritability estimates from traditional biometrical studies of families and GWAS results thus far comes from the fact that biometrical studies will overestimate heritability if genes interact non-additively.[14] If this suggestion is correct, then it may be that GWAS approaches that do not grapple with the combinatorial explosion posed by the search for gene-gene interactions will fail to produce interesting results. This criticism of biometrical studies, however, only applies when such studies focus on only one type of kinship (e.g., twins reared together). Many human traits, including height and IQ, have been studied biometrically using many different kinds of kinships (twins reared together and apart, parents and offspring living together and apart, adoptive relatives who live together but are biologically unrelated). When these results are considered collectively, they converge on relatively large heritability values.

The evidence base for claims about heritability has been strengthened by a recently developed way to estimate heritability that examines genetic variation directly. The genomic-relatedness-matrix restricted maximum likelihood (GREML) technique[15] uses all of the genotypic data from SNP arrays to estimate, for each pair of participants in a dataset, their degree of genetic similarity (relatedness), and then correlates genetic relatedness with phenotypic similarity across all of the pairs. Note that this technique does not require the participants to be related in the conventional genealogical sense of being siblings or cousins. It exploits the fact that all individuals within a population are distantly related, and that the level of relatedness varies considerably among pairs of people. For example, Davies et al. reported a GREML analysis with about 550,000 common SNPs and 3000 subjects in which about 45% of the variance in general cognitive ability could be directly explained by the SNP variation;[16] Chabris et al. replicated this finding with a smaller sample.[11] In the original application of GREML, Yang et al. showed that 45% of the variance in height across 4000 subjects could be explained by ~300,000 common SNPs.[15] These estimates leave room for unmeasured genetic variation (e.g., uncommon SNPs, other non-SNP polymorphisms) to explain additional heritability.

In this context, a "common" variant is a polymorphic site where the minor allele shows a frequency exceeding a certain threshold (say .05), whereas a "rare" variant is a site where the frequency of the minor allele falls below this threshold. The GREML results finding substantial heritability owed to common variants tend to discredit the hypothesis that missing heritability arises because common variants typically studied in GWAS are merely surrogates for rare variants of powerful effect that, if only they could be discovered, would account for much more heritability.[17] Furthermore, Wray et al. provide a thorough analysis of the available GWAS results and show that a model relying exclusively on rare causal variants cannot account for the data.[18] It is important to note that, under any reasonable evolutionary model, *most* genetic variants affecting a given phenotype may be rare. All else being equal, however, common variants contribute more variability than rare variants, and thus it is not at all inconsistent to

expect that common variants will be responsible for a substantial portion of heritability.

Though medical and behavioral geneticists are becoming increasingly sympathetic to the many-common-genes-of-small-effect answer to the missing heritability question, it is still not known *why* such a diffuse polygenic architecture should be typical of quantitative traits. Nor is it known what might account for the exceptions that have been found.

A simple possible explanation invokes the length of the causal chain from genetic to phenotypic variation. For example, variation in pigmentation (e.g., of eyes, skin, and hair) arises from the number of melanosomes produced, as well as their size and shape, and the type of melanin synthesized.[19] These biochemical differences follow directly from changes in the composition or regulation of gene products, which in turn are strongly influenced by differences in DNA sequence. Indeed, a single SNP in HERC2 is largely responsible for blue eye color.[20]

In contrast, changes at the molecular and cellular level must be remote from their ultimate effects in most behavioral phenotypes, and even from many physical phenotypes such as body mass index (BMI). Consider that BMI may depend on what a person likes to eat, how often he eats, how much he exercises, details of his metabolism, and a host of other complex behaviors and physiological processes. Similarly, given that the physical basis of psychological attributes such as cognitive ability, conscientiousness, impulsivity, and risk aversion resides in intricate patterns of neural circuitry and interlocking biochemical feedback loops, we should perhaps expect any single genetic variant affecting such an attribute to contribute only a small fraction of the total variation in the phenotype.

Here we offer a second explanation (which is *not* mutually exclusive with the first), which invokes the differential action of natural selection. More than 70 years ago, R.A. Fisher proposed a geometric model of adaptation[21] that may be summarized in Figure 1, which depicts two quantitative traits as the vertical and horizontal dimensions on a two-dimensional space (representing a slice of the vast multidimensional space of possible phenotypes). Point **A** represents the current mean phenotype of the species (in this example, a low value of trait 1 and an intermediate value of trait 2). Point **O** represents the optimum favored by natural selection. Suppose that **A** was once optimal, because selection had pushed the population to its optimum value, but that **O** no longer coincides with **A** because an abrupt environmental change occurred that demands a different (in this case higher) value of trait 1.

What would have to happen for selection to adapt the organism to the new optimum? One possibility is a new mutation arising in a single individual and, if beneficial, reaching fixation (100% allele frequency) in the population. In the model the fixation of a mutation corresponds to adding a vector of random direction to the population's current trait-space position at **A**. This feature of the model captures two key observations: (1) mutations have no inherent tendency to increase the fitness of their bearers, and (2) any single mutation may affect several distinct traits, and therefore this mutation could change the population's mean values of both traits 1 and 2. The subset of new phenotypes that would result in an increased level of adaptation is depicted in Figure 1 as the interior of the circle centered on **O**, representing all the combinations of trait values that are closer to the optimum (using the Euclidean distance metric).

The diagram also helps one understand the fates of mutations with different effect sizes. Note that any mutation whose effect on the traits exceeds the diameter of the circle would not be fixed, because it leaves the population farther from the optimum than when it started (point **A**). Selection would simply favor the status quo. In general, the smaller the mutation, the more likely it is to be beneficial, because there are many small moves that can be made from **A** that stay within the circle, but few very large ones—most large moves will overshoot the circle or move away from it. The fact that a smaller move is more likely to take the population into the circle should already be evident from Figure 1. As the number of traits/dimensions increases to larger values (which cannot be depicted in a two-dimensional figure), the greater ease with which smaller moves take the population into the "hypersphere" becomes quite dramatic.[*]

Fisher argued that mutations of large effect are relatively unimportant in evolution, since they will rarely move a population closer to **O**. And the closer to the optimum the organism already is, the less likely large mutations are to be beneficial. Fisher draws an analogy to the process of focusing a microscope. When a microscope is already close to the correct focusing point, a small random perturbation of the knob is likelier than a larger perturbation to bring it closer to exact focus.

We will now expand Fisher's argument to explain the puzzling contrast between some physical phenotypes like skin or eye color on the one hand and social science and medical phenotypes on the other. Suppose that trait 1 was previously under strong stabilizing selection and thus has negligible genetic variation at the time of the environmental shift that makes **O** the new optimum (this state of affairs would correspond to a tight clustering of trait 1 values around point **A**). Since the rate of the approach to the optimum via existing genetic variation (i.e., variation that does not result from de novo mutations) is bounded above by trait 1's heritability (per the breeder's equation),[22] a population with negligible genetic variability in that trait is unlikely to adapt quickly towards **O** unless a mutation of large effect arises and reaches fixation—e.g., a mutation that took the population to **A'**, where its new value for trait 1 is much closer to the optimum.

Alternatively, suppose that stabilizing selection on trait 1 had been much weaker, permitting the buildup of substantial genetic variation (leading to a wide scatter of trait 1 values around **A**). In this case a mutation of large effect is far less likely to become common through positive selection. At the same time that this mutation is struggling to increase its frequency, the existing genetic variation is enabling the population to adapt toward **O.** If **O** lies within the current range of genetic variation (which is true for trait 1 under the assumption of the more variable population in Figure 1) and selection is even moderately strong, then the population mean shifts from **A** to **O** even without the arrival of a mutation of large effect. As the population evolves, the diameter of the circle bounding all points of higher adaptation continuously shrinks. Once the magnitude of the mutation that would have taken the population to **A'** exceeds the diameter of the circle, the mutation is disfavored and is very likely to be eliminated from the population.[†]

---

[*] If the number of traits (n) is large, then the probability that a random mutation of length r takes the population into a hypersphere of radius z is $1 - \Phi(x)$, where Φ is the cumulative distribution function of the standard normal distribution and $x = r\sqrt{n}/(2z)$ [ref. 21].

[†] A numerical example may help to illustrate our argument. Suppose that the fixation of a new

To complete our explanation, we need to assume that the polymorphic sites contributing to existing genetic variation tend to be small in effect. Even under weak stabilizing selection, variants of large effect experience greater selection pressure and consequently are more likely to be found at a low minor allele frequency.[25–27] This implies that any common (i.e., high-frequency) variants contributing to standing genetic variation will typically be small in effect. Thus, we might expect many loci of small effect to explain most of the heritable variation underlying a quantitative trait—unless there was recent selection for the trait that was strong relative to the initial variability. If a trait turns out to be associated with many genetic loci of small effect and few or no loci with large effects, then we would have evidence that this trait has not experienced such selection.

In the remainder of this paper, we will show how this evolutionary analysis can help epidemiologists and social scientists make sense of the genetics of behavior in the era of rapidly expanding genome scans. We report the results of our own GWAS of more than a hundred human phenotypes, both physical phenotypes such as body size and pigmentation, and behavioral phenotypes of great interest to social scientists, such as general intelligence, memory ability, verbal fluency, impulsivity, risk aversion, fairness, and utilitarianism. We measured a wide variety of cognitive, personality, and behavioral-economic traits so that we could generalize across types of traits and compare the behavioral to the physical phenotypes. In other words, without sampling freckles, eye color, and height in a single study, we could not make general claims about physical traits; without measuring religiosity, memory, and impulsiveness in a single study, we could not make general claims about behavioral traits; and without measuring both categories we could not compare them. To our knowledge this study is the first to examine associations between a genome-wide panel of single-nucleotide polymorphisms (SNPs) and such a broad spectrum of phenotypes; almost all previous association studies of behavioral traits have examined only one or a few candidate genes and phenotypes. In addition to including both physical and behavioral traits, the study examined traits that are expected to be both monogenic and polygenic. An additional innovation is that the behavioral phenotyping was intensive, relying

---

mutation is the only means for the population to increase its level of adaptation—that is, there is initially no genetic variation along the selected direction. Then if the selective advantage of the new mutation is 5%, it will take about 500 generations to increase from a frequency of .001 to .999 [ref. 23]. A selective advantage of roughly this magnitude seems reasonable for many of the mutations affecting pigmentation. Now suppose that the population contains substantial genetic variability in the trait. In particular, suppose that the trait has heritability 100% and follows a standard normal distribution. If we stipulate that the old and new optima are separated by 2 phenotypic units (and that each unit continues to correspond to a 5% change in relative fitness; i.e., a 5% gain in offspring per generation), then standard quantitative-genetic results[23,24] imply that the population will reach the new optimum in 40 generations. If the preexisting variants of small effect have pleiotropic effects, the adaptation time may be somewhat longer. Nevertheless, in a race between the fixation of a major mutation and polygenic adaptation, the latter will often have a profound advantage. Once polygenic adaptation has brought the population close to the new optimum, the major mutation will become disfavored while still at a low frequency.

not just on standardized paper-and-pencil tests but on individual computerized tasks, sometimes administering hundreds of trials to quantify a single trait. This step is essential because crude measurement of behavioral traits could lead to false negatives and thus would not help explain the puzzling failure to find associated genes. Thus, each of our 419 participants was tested individually in a laboratory session lasting an average of 3.5 hours.

To preview the results: Despite an adequate sample size for detecting large effects and despite high-precision measurements, we found few associations between SNPs and traits at an appropriately stringent significance threshold. Since many of our measured phenotypes (including our behavioral phenotypes) are known to be heritable,[28] the absence of strong associations in our data indicates that—aside from pigmentation—both physical and behavioral traits are mainly affected by numerous genes with small effects.

After presenting the results, we discuss their implications for future genetic association studies of behavioral traits, which are likely to become ever more common as the cost of genotyping and sequencing declines. In addition to our analysis of the evolutionary genetics of heritable variation, we introduce two other key issues in designing and interpreting such studies: the effects of selection bias for participant inclusion in such studies, and the tradeoffs between measurement error and statistical power in selecting simple, fast, inexpensive assessments of a traits versus the sort of complex, time-consuming, and potentially expensive assessments that we conducted.

Methods

Participants were recruited, and data and samples were collected, at two sites: Harvard University in Cambridge, MA, and Union College in Schenectady, NY. Efforts were made to recruit from the surrounding communities a more representative sample than the typical college student population: Paper fliers were posted at various public locations, advertisements were placed in free newspapers and on Craigslist, and the study was made available to the Psychology Department Study Pool at Harvard.

Participants first completed an online screening questionnaire that included items regarding age, medical history, and grandparental ethnicity. Participants who were younger than 18 or older than 45, or who reported a history of bipolar disorder, schizophrenia, or severe head trauma were excluded. To help control for ancestral confounding of genotypes and trait levels,[29] we recruited a sample of predominantly Western European ancestry, which was ascertained at the screening process by asking potential participants to list the country of origin or ancestry for each of their biological grandparents. A total of 419 participants provided complete, usable genetic and phenotypic data.

Eligible participants were invited to either the Harvard or Union lab for a data collection session lasting typically from three to four hours. Participants gave informed consent after the nature of the procedure had been fully explained to them. A diverse set of cognitive, personality, economic, attitude, demographic, and physical phenotypes were collected via computerized tasks, paper-and-pencil surveys, and face-to-face interaction. DNA was collected via two

mouthwash samples in the lab, and then extracted and genotyped elsewhere. Population stratification was investigated and controlled for in all genetic analyses reported here. We used the program PLINK for genotypic data cleaning and analyses.[30] (See Supporting Online Material for a complete list of phenotypes, descriptions of select phenotypes, and details of DNA collection, extraction, genotyping, and analysis of population stratification.)

Linear regression was performed to test for purely additive association between SNPs and all polytomous and continuous traits. Logistic regression was performed for dichotomous traits. We chose the standard genome-wide significance threshold of 5×10^{-8} for declaring a SNP-trait association to be statistically significant.[31] Under a frequentist approach aiming to minimize the chance that even a single declared "hit" is a false positive, the large number of examined traits would require an even more stringent threshold. However, we follow the suggestion of the Wellcome Trust Case-Control Consortium,[32] who adopt a quasi-Bayesian justification for retaining the standard genome-wide significance threshold; it maintains a constant ratio of true to false positives as the number of markers and traits increases (so long as statistical power and prior probabilities for any given association do not change). Moreover, since our primary goal is to compare results across phenotypes, what is most important is to have a common threshold across phenotypes, and adopting the standard threshold maximizes comparability of our results with other published results.

For any SNP showing an association with a trait at the significance threshold 5×10^{-8}, we re-ran PLINK with our cognitive ability composite and NEO Openness, Neuroticism, and Agreeableness factor scores as additional covariates in an effort to control for selection bias.[33] Selection bias may be an underappreciated contaminant in gene-trait association studies.[34] To understand the bias, consider this analogy: Suppose that a driveway will be wet in the morning as the consequence of two possible causal mechanisms: whether it rained last night, and whether a sprinkler was activated (Figure 2A). Suppose also that the two causal variables are independent; that is, taking all days into account, there is no correlation between whether it rains and whether the sprinkler turns on. If we only consider mornings on which the pavement is wet, however, we will spuriously conclude that the two causes are negatively correlated. For instance, if we see that the pavement is wet and we know that it did not rain last night, we can be confident that the sprinkler was activated. We only see the true non-correlation when we consider all days. Suppose that the probability of rain and the probability of sprinkler activity are both 0.5 and are independent. If one checked the driveway every morning, wet or dry, then one would observe rain and no sprinkler a quarter of the mornings, sprinkler and no rain a quarter of the mornings, both a quarter of the mornings, and neither a quarter of the mornings—the lack of association is apparent. Now suppose one checked only the mornings with wet driveways. On a majority of the mornings (two-thirds), one would discover either rain with no sprinkler or a sprinkler with no rain. In other words, one would find a negative correlation, but only because those mornings that would have diluted the correlation to zero were excluded. The basic principle emerging from this example is that if one inadvertently conditions an observation on the common effect (is the driveway wet?) of multiple causes (rain or sprinkler), one can counterfactually create the illusion of a non-zero correlation among the causes.[‡]

---

‡ An example using continuous variables may also help to illustrate the concept of selection bias, and its generality. Suppose that intelligence and athletic ability (both continuous traits) are

This same principle applies in GWAS. Suppose that high levels of either trait 1 or 2 are independent causes of a person ending up as a participant in our study, either because the trait affects whether the person decides to volunteer or it affects whether we decide to include his or her data (Figure 2B). Then we will spuriously find any gene that affects trait 2 to be associated with trait 1, even if trait 1 is not at all affected by genetic variation. That is because among people who participate in the study, traits 1 and 2 will appear to be (negatively) correlated, and therefore a cause of scoring high on trait 2 will appear to also be a cause of scoring low on trait 1. Controlling for the other traits affecting participation may not fully solve the problem (even if we know what these traits are), because the trait of interest may itself be connected to the other participation-related traits in a complex causal graph, and therefore the decision to condition linearly on the other traits could in principle introduce further bias. In practice, however, conditioning on traits that may affect study participation is likely to be a conservative procedure. For example, if one trait mediates the genetic effect of another, then controlling for the mediating trait will suppress the genuine effect of the genetic variant on the downstream trait of interest, and therefore is unlikely to generate additional false positives.

We performed a numerical simulation to illustrate the extent to which selection bias may distort GWAS results. We stipulated two initially independent traits affecting participation in the study; the sum of an individual's *z*-scores on these traits needed to exceed 3 in order for the individual to be in the pool of participants. This corresponds to slightly less than two percent of the general population being available to participate. We believe that this simulated situation is not so farfetched as a model of some ongoing projects (e.g., the Personal Genome Project;[36] 23andMe[37]). We stipulated that each trait has a heritability of 0.50 and is affected by loci all with allele frequency 0.50 and average effect (regression coefficient) 0.05; each causal locus thus account for 0.25% of the variance in its trait. The results were striking: The estimated effects of the true causal variants with respect to a given trait were centered at 0.03—off by 40%. Similarly, the "effects" of the variants on the wrong trait (of the two traits, the one that the variants did not affect) were centered at –0.02. In a situation where it is important to distinguish miniscule effects from zero, a spurious effect of 0.02 cannot be considered trivial. Although more thorough numerical and analytical investigations are certainly worthwhile, this example illustrates that researchers performing GWAS of behavioral traits should be aware of the consequences of selection bias.

---

uncorrelated in the population at large. However, if we limit our observations to the students attending a university that uses both of these attributes as admissions criteria, then we will find that intelligence and athleticism are negatively correlated. If we encounter a student at this university with low intelligence, then it becomes more probable that the student is a good athlete. Otherwise the student would likely not have been admitted. This negative correlation between intelligence and athleticism among admitted students holds even if admission is not a deterministic function of these two attributes; other attributes (e.g., musical talent) and "random noise" may play a role. Verma and Pearl provide a rigorous mathematical proof that conditioning on a common effect induces dependence among the causes.[35] The apparent dependence does not have to be a negative correlation as in these examples; an apparent positive correlation would result if, say, students high in *both* athleticism and intelligence were especially likely to be admitted.

Table 2 includes the sample statistics for the Multidimensional Attribute Battery (MAB) and NEO personality inventory, two instruments used in our study that have detailed population norms. Compared to the norming samples for the MAB, our participants show much higher means and smaller standard deviations, suggesting that cognitively able individuals were more likely to participate in the study. The relationship between the NEO personality traits and study participation is more complex. Our study participants show conspicuously higher levels of Openness than the norming samples. The trait of Openness is defined by a willingness to examine new ideas and try new activities, and thus it is plausible that higher levels of this trait may be a cause of volunteering for scientific research. Our study participants also show consistently lower levels of Neuroticism and higher levels of Agreeableness. (Interestingly, our study participants are more variable than the norming samples, perhaps because people with higher cognitive ability are more variable in their responses to personality questionnaires.[38]) Furthermore, the fact that students were overrepresented among our participants indicates that the selection bias may have already operated extensively at an earlier point. That is, even if we could have taken a random sample of all students attending the top 200 colleges (say), the process of college admissions would still have exerted considerable selection bias distinguishing this special population from its larger age cohort. As a reasonable attempt to control for selection bias, then, we will use general cognitive ability, Openness, Neuroticism, and Agreeableness as additional covariates whenever a novel SNP-trait association shows a significant *p*-value. Without doing this, we might spuriously find, for example, that a gene associated with greater Openness was also negatively associated with all the traits that are correlated with Openness, such as political liberalism (see below).

Results

As can be seen in Table 3, we found at least a marginal signal for all SNPs previously found to be associated with eye color, hair color, freckling, and skin color [19,37,39-42] (with the exception of one study that digitally quantified eye color[43]) and that were either present in our cleaned set of genotyped SNPs or represented by a proxy SNP with an $r^2 > .60$. Note that despite our relatively small sample size, the effects of the intronic SNP rs12913832 in HERC2 on eye and hair color were statistically significant at the stringent, standard GWAS threshold.

A meta-analysis has identified over 180 genomic regions containing a variant affecting height.[44] Due to the weak effect of each individual variant, however, we did not replicate any of these loci with genome-wide significance. However, of the 94 loci either present in our set of SNPs or represented by a proxy, 65 loci had estimated effects with the correct sign and 29 did not (binomial test $p < .0001$). There is also an enrichment of low *p*-values; whereas only nine or ten *p*-values less than .10 were expected under the null distribution, we observed 16 (significantly more, according to a binomial test, $p < .05$). These trends are consistent with most of these loci being true positives despite our inability to extract a strong signal from them. A selection of the height variants showing marginal significance in our data is shown in Table 4A, along with the nonsynonymous SNP rs1815739 in ACTN3 that has been found to affect athletic performance.[45]

Another recent meta-analysis has identified 32 genomic regions containing a variant affecting

body mass index (BMI).[46] BMI, even more than height, seems to be affected by many loci of small effect. Consistent with this view, 11 of the 17 known BMI loci represented in our data had estimated effect sizes of the correct sign; however, the wrong-signed loci were the most statistically significant.

Table 4B shows our results for a selection of SNPs previously reported to be associated with general cognitive ability,[47–51] personality,[52,53] working memory,[54] and episodic memory,[55] all of which we measured extensively. We observed little evidence for these associations in our own data. In concordance with a previous study,[56] we failed to replicate a reported association between a common SNP in the gene KIBRA and episodic memory, despite a putative functional validation in the original study both by an analysis of gene expression and by fMRI.[55] This suggests that most of the SNPs reported in earlier association studies of behavioral traits may either have been false positives or have overestimated effect sizes. Applying a threshold of 5×10^{-8}, we did not observe any loci significantly associated with the traits in Table 4B.

We did find a significant association between political conservatism and rs10952668 (Table 5). This SNP lies in LOC642355, a pseudogene on chromosome 7. Not surprisingly, the SNP also showed an association with the highly correlated trait of Democrat vs. Republican ($\beta = .260$, $p < .02$). We also observed a significant association between rs1402494, which lies in a gene desert on chromosome 4, and gambling frequency. These are the only two novel associations that reached genome-wide significance, and besides these, only eye color and hair color also produced significant associations.

Interestingly, the SNP associated with political conservatism, rs10952668, also showed marginal evidence for association with the personality traits Openness ($\beta = .142$, $p < .06$) and Agreeableness ($\beta = .130$, $p < .08$), which are correlated positively with political liberalism.[57] Since the correlation is positive, contrary to findings from political psychology that conservatives tend to be less Open and Agreeable (in the sense of compassionate[58]), these results raise the possibility that the association between rs10952668 and conservatism may be attributable to selection bias rather than the gene causing the personality traits typical of conservatives. (Since to our knowledge this potential selection artifact has not been discussed in the genetic epidemiology literature—although it has parallels in the effects of natural selection on linkage disequilibrium—we explore it at some length in the Discussion below.) After we added general cognitive ability, Openness, Neuroticism, and Agreeableness as covariates in an attempt to control for selection bias, the association of rs10952668 and conservatism diminished and fell short of significance. The association of rs1402494 and gambling frequency appears robust against our attempts to control for selection bias. We conclude that both of these associations must be replicated in much larger samples before they are accepted as true positives.

Discussion

The contrast between pigmentation and the other phenotypes examined in this study is striking (Tables 3–4). Given a significance threshold of 5×10^{-8}, our study had statistical power approaching 0.80 to detect any locus accounting for more than 10% of the variance in any trait. We retained some power (0.12) for loci accounting for as little as 5% of the variance. The fact

that we measured so many phenotypes implies that we should have obtained several hits if a large proportion of the phenotypes were indeed affected by such loci. Because we only obtained at most two new hits, however, loci with effects of this magnitude on the non-pigmentation traits we studied (see Table S1) must be uncommon. In agreement with previous studies,[11,53,59,60] we conclude that cognitive ability, personality dimensions, social attitudes, and most other traits of interest to behavioral scientists are affected by numerous loci of small effect. In this respect the behavioral traits we studied resemble height and BMI rather than pigmentation.

How can we explain the differences in genetic architecture between the pigmentation traits and the other physical and behavioral traits? One possibility is that the architecture hinges on the length of the causal chain between gene and phenotype. Pigments, after all, are molecules, and you can change a molecule, thereby giving a person a different eye color, by changing a single gene. It's not as easy to make a person more intelligent, utilitarian, altruistic, or impulsive by changing one gene, owing to the greater complexity in the mechanisms that lead a person to be intelligent or altruistic in the first place. With gross physical traits like BMI and height, the problem may be that there are *too many* ways that genes can directly affect the phenotype; indeed, it may be hard for a genetic change *not* to affect them, just as most changes to the features of (say) a car or laptop computer have consequences for its size and weight, which engineers have to trade off with many minute compensations.

The other explanation invokes the evolutionary model of the causes of genetic architectures we outlined earlier, which relates the effect size of genetic polymorphisms to the magnitude and recency of changes in the adaptively optimal level of the trait. After the loss of body hair in our lineage, pigmentation probably came under strong stabilizing selection in our ancestors, who needed protection from the African sun. More recently, the out-of-Africa migrants ancestral to Europeans and East Asians experienced a sudden and drastic shift in the optimal level of pigmentation—perhaps because of the need to sustain cutaneous synthesis of vitamin D in northern climates,[61] although others have implicated sexual selection or as-yet unidentified evolutionary pressures.[62–64] In any event the result was that several depigmenting mutations of large effect increased rapidly in frequency.[65–67] Table 4A lists those mutations that have not yet reached fixation and are thus still polymorphic in Europeans.

No such recent environmental change—one with clear consequences for the direction and magnitude of the optimum—is apparent for other phenotypes such as height, BMI, and the behavioral traits we examined. Though differences in climate and food availability may select for different optima in body size and shapes, they fluctuate rapidly across space and time and may not show the consistent selection pressure that changes in latitude, altitude, and cloud cover apparently exerted on pigmentation. Intelligence is a highly general and universally adaptive trait, which can translate into fitness benefits (via successful problem-solving) in any environment. If human populations have long been at the optimum, then existing mutations are likely to be small in effect. Such variants are likely to be small in effect even if the optimum has changed over time—as may have happened in the cases of intelligence[68,69] and religiosity[70]—so long as the change occurred very gradually. In particular, intelligence may be a highly general and universally adaptive trait, responding more to coevolutionary pressures exerted by language and sociality than to any sudden change in the physical environment. Personality traits, too, are far less predictably correlated with physical environments than are pigmentation traits.

Evolutionary game theory has established theoretical rationales for the persistence of multiple behavioral phenotypes (e.g., hawk and dove strategies) in the same population.[71,72] Analogously, the selective environment for personality may consist of the local distributions of the personalities of other people,[72] and the mixture is unlikely to have changed in a systematic way with recent shifts in the human population.

Even if selection has acted on these traits since the dispersal of homo sapiens from Africa, the new optima could have been quickly reached by small shifts in allele frequency at many minor loci, leaving any major mutants at the low frequencies determined by the interaction of mutation, drift, and stabilizing selection.[73] As discussed above, the result of such dynamics would be the observed absence of common variants with large effects.

Our two proposals for explaining the pattern in Tables 3–5 lead to the following suggestions for future GWAS of behavioral traits. First, to understand the causal chain between genetic and phenotypic variation, we should try to narrow the chasm from both sides. Doing so requires seeking and validating endophenotypes that lie closer on the causal chain to genetic variation than the coarse and easily measured phenotypes we are used to. Second, researchers seeking variants of large effect should ideally study populations where directional selection may have recently produced a phenotypic change that is large relative to the initial standing variation. Recent studies of altitude adaptation in Tibetans exemplify both of these suggestions.[74–76] The genes successfully associated with red blood cell count and hemoglobin concentration in these studies would have been more difficult to identify if the phenotype had been characterized at a level as abstract as "altitude tolerance." Moreover, the recent and rapid divergence of Han Chinese and Tibetans in altitude tolerance after the latter began to occupy a highland environment was plausibly driven by a selection differential large enough to pull variants of large effect away from the boundary of frequency zero. It is, however, an open question how many social-science traits can be studied by looking for recent directional selection.

As for traits with more typical evolutionary histories, the expectation of small effect sizes requires that much larger samples be ascertained than are common in social-science genetics research. We see two promising approaches. One is for researchers to take advantage of the potential for large sample sizes by allying with the burgeoning field of personal genomics, in which a large base of volunteers or consumers provide genotype and phenotype information.[37,77,78] It is crucial, though, to check these samples for selection biases, because many phenotypes of interest are likely to be causes of participation in personal genomics itself. For example, an individual with a liability to a particular disease may be strongly motivated to participate in a personal genomics study by self-interest or altruism; participants also must be wealthy enough to afford the service. We conjecture that our findings of elevated cognitive ability and intellectual openness among research volunteers will generalize to future studies. If so, it is prudent to collect reliable measurements of these traits in all GWAS that are not based on population samples and to note any unusual sample distributions on these traits when reporting SNP-trait associations.

The other approach is the traditional epidemiological study, which attempts to minimize the impact of personal characteristics on study participation by recruiting a population-based sample. This will remain an important complement to volunteer- and consumer-driven approaches.

Recall that Chabris et al. consulted three population-based studies and found that only one out of 12 published genetic associations with general intelligence could be replicated within them, and that one only one out of three times.[11] One explanation is that the original associations came from small convenience samples similar to the one we studied here.

There is, however, a tradeoff inherent in using large population-based studies for gene discovery. Most of these projects are directed towards medical outcomes rather than social-science traits (with some notable exceptions, such as the Health and Retirement Study, the Wisconsin Longitudinal Study, and the English Longitudinal Study of Aging; the first of these now has GWAS data available, and the others may soon). Data collection in these surveys, although often face-to-face and longitudinal, distributes time and effort across many phenotypes that are measured with short questionnaires (or even single questions). The disadvantage of such studies is that whenever the underlying trait of interest is continuous, quick or brief measures are inherently less reliable (i.e., are subject to more measurement error) than are more detailed ones.

Genetic association studies, then, present researchers with a tradeoff between using high-quality or high-technology (e.g., neuroimaging) measures of each phenotype, which are often only feasible for small samples, and having a large sample in which the phenotype is measured poorly. In social science research, this dilemma is commonly resolved in favor of smaller samples with higher quality measures—and perhaps for this reason, that is the strategy in most of the social-science genetic association studies conducted to date, including the one we reported here. But because the genetic architecture of behavioral traits is likely to feature very weak genetic associations, our intuitions regarding the appropriate research strategy may not be correct when carried over from non-genetic social science research, where effect sizes are typically much larger. There is as yet no straightforward way to calculate an expected effect size for genetic associations in social science, so the best we can do is to assume that effects will be similar to those found for other complex (polygenic) traits—tiny.

Figure 3 displays the results of a set of power calculations that quantify the tradeoff. The phenotype is assumed to be normally distributed. The y-axis shows effect sizes in terms of R^2, the fraction of variance in the phenotype explained by variation in a single genotype, ranging from 0 to 0.01 (one percent) in increments of 0.001 (one-tenth of one percent). The x-axis is the sample size. Each curve graphs the locus of effect-size/sample-size pairs that gives 50% power to detect the association at $p = 5\times10^{-8}$ for a given phenotype reliability. The phenotype reliability is measured in terms of the test-retest correlation, i.e., the correlation between two independent measurements of the phenotype. We consider the cases where reliability is equal to 1.0, 0.8, 0.6, 0.4, and 0.2.

For the very small effect sizes that can be expected for behavioral traits, Figure 3 indicates that it will generally be better to sacrifice phenotype quality in favor of larger sample sizes. For example, consider an effect size of $R^2 = 0.001$ (0.1% of the variance). This is the size of the association found in a meta-analysis of the association between cognitive ability and variation in the COMT gene in 67 independent samples, and it is likely to be biased upward because the meta-analysis found evidence of publication bias.[79] Since cognitive ability is among the most reliably measured social science traits and since the meta-analysis found evidence of publication bias, such an effect size is likely to be representative of the largest associations we can expect for

a behavioral trait. Given $R^2 = 0.001$, for a perfectly-measured phenotype (reliability = 1.0), 50% power requires a sample size of 30,000 individuals. This is far too large a sample to obtain high-quality measures of behavioral traits, which generally requires bringing the research subjects into a laboratory and conducting repeated tests spanning many minutes or hours. In contrast, for a phenotype with test-retest reliability of 0.6—which is typical of behavioral phenotypes measured by brief questionnaires—50% power requires a sample size of 50,000 individuals. Samples at least this large have recently become feasible. Medical datasets that have already collected GWAS data could much more easily add brief behavioral questionnaires to their ongoing data collections than onerous laboratory sessions. Since such medical datasets in aggregate comprise hundreds of thousands of participants, such a research strategy should be possible.[4]

Conclusions

We conducted a Genome-Wide Association Study on more than 100 carefully measured phenotypes in more than 400 subjects, but found very few loci of large effect associated with any trait other than the pigmentation of eyes and skin. This includes a substantial proportion of the traits that have been of theoretical interest to behavioral scientists in recent decades. Four points emerge from our analysis:

1. The genetic architecture of trait variation cannot be taken as constant across traits, particularly the expectation that a single gene or a small number of genes will have a noticeable effect on the trait. First, the shortness of the causal chain between the DNA and the trait matters a great deal, with single-gene effects being more likely for traits generated by a single protein or regulatory shift. Second, the genetic architecture of a trait is intimately intertwined with its evolutionary history. The implications flow in both directions: the discovery of an association between a gene and a trait can illuminate the evolution of our species, and the evolutionary process determines which associations we can most readily discovered. In particular, stabilizing selection of moderate strength, which permits a substantial background of weak or rare variants, supplies the fuel for polygenic adaptation and may obviate the need for mutations of large effect to arise after a sudden environmental change.

2. Many psychological traits of interest to researchers are themselves plausible causes of participation in scientific research, which raises the potential of spurious associations. Measuring such traits (e.g., cognitive ability and personality) and incorporating them into analyses is one strategy for dealing with this issue.

3. If there are two ways to measure a trait—a high-reliability measure that can be performed only on a small sample because of the required time, effort, and resources, versus a lower-reliability brief measure that can be administered to a large sample—power analyses suggest that using the lower-reliability measure with the larger sample size is likely to be the best strategy. Researchers interested in the genetic architecture of behavioral traits should therefore consider working with large-scale survey datasets such as the HRS, WLS, and ELSA, as well as medical-genetic studies that are willing to conduct social science surveys among their participants.

4. Genetic associations with behavioral traits have proven notoriously difficult to replicate. This

is not because the relevant traits are not heritable or the original studies were poorly designed or knowingly underpowered; researchers at the time lacked the resources for conducting more genotyping and assembling larger samples, and they were hoping to find common alleles with large effect size. Our discussion of Fisher's model, and the empirical experience accumulated in the first fifteen years of social science genetics, suggest that individual gene effect sizes for traits not under strong selection are likely to be extremely small, and therefore require extremely large datasets to be detected.

The fact that faster, cheaper, and more powerful methods of genotyping have led to fewer, smaller, and less reliable findings on the connection between genes and behavior, despite the near-certainty that such connections exist, stands as one of the disappointments of 21st century science. To make progress, we should shift away from the traditional model of epidemiology via statistical significance testing, in which large significant correlations are the standards of success and worthy of newspaper headlines, while negative results are considered a failure and destined for the file drawer. It has become increasingly clear that this practice has led to mischief both in epidemiology and in social science,[80,81] and it may also be preventing the discovery of important scientific insights. If we have learned that behavioral genetic variation is caused by many genes with effects that are too small to currently measure, then we have also learned something important about the physiology and evolutionary history of such traits. With nature as with people, the Yiddish expression may apply: No answer is also an answer.

Acknowledgments

We thank Stephen M. Kosslyn for his support. Brian Atwood, Chris Eur, Kathleen Huber, Sara Igoe, Minji K. Lee, Melissa Liebert, Jaclyn Mandart, Ben Orlin, Mike Puempel, Esther Snyder, Martha Widger, Linda Yao, and Chelsea Zhang provided research assistance. Funding was provided by research funds of the authors, as well as NIA grant T32-AG00186 to the National Bureau of Economic Research. The study reported here was approved by the Institutional Review Boards of Harvard University and Union College.

Author Contributions

Conceived and designed the experiments: CFC, JJL, DJB, DIL, GB. Wrote the paper: CFC, JJL, DJB, SP. Analyzed the data: JJL, JPB, GB. Principal Investigators of the study: CFC and DIL. Critically reviewed and edited the manuscript: CFC, JJL, DJB, ELG, SP, DIL.

References

1. Plomin R, DeFries JC, McClearn GE, McGuffin P. *Behavioral genetics.* 5th ed. New York, NY: Worth Publishers; 2008.
2. Turkheimer E. Three laws of behavior genetics and what they mean. *Current Directions in Psychological Science*. 2000; 9: 160–164.
3. Pinker S. *The Blank Slate: The Modern Denial of Human Nature*. New York, NY: Viking; 2002.
4. Benjamin DJ, Cesarini D, Chabris CF, Glaeser EL, Laibson DI, Guðnason V, et al. The Promises and Pitfalls of Genoeconomics. *Annual Review of Economics*. Preprint. September 2012.
5. Cecil JE, Tavendale R, Watt P, Hetherington MM, Palmer CAN. 2008. An obesity-associated FTO gene variant and increased energy intake in children. *New England Journal of Medicine, 359,* 2558–2566.
6. Israel S, Lerer E, Shalev I, Uzefovsky F, Riebold M, Laiba E, et al. The oxytocin receptor (OXTR) contributes to prosocial fund allocations in the dictator game and the social value orientations task. *PLoS One*. 2009; 4: e5535
7. Apicella CL, Cesarini D, Johannesson M, Dawes CT, Lichtenstein P, B. Wallace J, et al. No Association between Oxytocin Receptor (OXTR) gene polymorphisms experimentally elicited social preferences. *PLoS ONE*. 2010; 5: e11153.
8. Hirschhorn JN. Genomewide association studies—Illuminating biologic pathways. *New England Journal of Medicine*. 2009; 360: 1699–1701.
9. Chen L, Davey Smith G, Harbord R, Lewis S. Alcohol intake and blood pressure: a systematic review implementing Mendelian Randomization approach. *PLoS Medicine.* 2008; 5: 461-471.
10. Fletcher J, Lehrer S. 2009. The effects of adolescent health on educational outcomes: causal evidence using genetic lotteries between siblings. *Forum for Health Economics & Policy*. 2009; 12: Health and Education, Article 8.
11. Chabris, Christopher F, Hebert BM, Benjamin DJ, Beauchamp JP, Cesarini D, et al. Most published genetic associations with general intelligence are probably false positives. *Psychological Science*. Preprint.
12. Beauchamp JP, Cesarini D, Johannesson M, der Loos M, Koellinger P, Groenen PJF, et al. Molecular genetics and economics. *Journal of Economic Perspectives*. 2011; 25: 1-27.
13. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. Finding the missing heritability of complex diseases. *Nature.* 2009; 461: 747–753.
14. Zuk, O., Hechter, E., Sunyaev, S. R., & Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *PNAS*, *109*, 1193-1198.
15. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*. 2010; 42: 565-569.
16. Davies G, Tenesa A, Payton A, Yang J, Harris SE, Liewald D et al. Genome-wide association studies establish that human intelligence is highly heritable and polygenic. *Molecular Psychiatry*. 2011; 16: 996-1005.
17. Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H. H., & Goldstein, D. B. (2010). Rare variants create synthetic genome-wide associations. *PLoS Biology*, *8,* e100294.
18. Wray, N. R., Purcell, S. M., & Visscher, P. M. (2011). Synthetic associations created by

rare variants do not explain most GWAS results. *PLoS Biology*, *9*, e1000579.
19. Sturm RA. Molecular genetics of human pigmentation diversity. *Human Molecular Genetics.* 2009; 18: R9.
20. Sturm RA, Duffy DL, Zhao ZZ, Leite FP, Stark MS, Hayward NK, Martin NG, Montgomery GW. A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color. *American Journal of Human Genetics.* 2008; 82(2): 424–31.
21. Fisher, RA. *The Genetical Theory of Natural Selection: A Complete Variorum Edition.* Oxford, UK: Oxford University Press; 1999.
22. Crow, J. F. (1986). *Basic concepts in population, quantitative, and evolutionary genetics*. New York: Freeman.
23. Lande R. Quantitative genetic analysis of multivariate evolution, applied to brain: Body size allometry. *Evolution.* 1979; 33: 402–416.
24. Lynch, M. & Walsh, B. (1998). *Genetics and the analysis of quantitative traits*. Sunderland, MA: Sinauer.
25. Wright S. The distribution of gene frequencies under irreversible mutation. *Proceedings of the National Academy of Sciences of the United States of America.* 1938; 24: 253–259.
26. Hastings A. Second-order approximations for selection coefficients at polygenic loci. *Journal of Mathematical Biology.* 1990; 28: 475–483.
27. Eyre-Walker A. Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proceedings of the National Academy of Sciences of the United States of America.* 2010; 107: 1752–1756.
28. Cesarini D, Dawes CT, Johannesson M, Lichtenstein P, Wallace B. Experimental game theory and behavior genetics. *Annals of the New York Academy of Sciences.* 2009; 1167: 66–75.
29. Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, et al. Demonstrating stratification in a European American population. *Nature Genetics.* 2005; 37: 868–872.
30. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*. 2007; 81: 559–575.
31. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, et al. Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nature Reviews Genetics.* 2008; 9: 356–369.
32. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature.* 2007; 447: 661–683.
33. Pearl J. *Causality: Models, reasoning, and inference*. 2nd ed. New York, NY: Cambridge University Press; 2009.
34. Lee JJ. 2010. Review of "Intelligence and how to get it: Why schools and cultures count" by R.E. Nisbett. *Personality and Individual Differences, 48,* 247-255.
35. Verma and Pearl (1988). Causal networks: Semantics and expressiveness. In *Proceedings of the Fourth Workshop on Uncertainty in Artificial Intelligence* (pp. 352-259). Mountain View, CA.
36. Ball, M. P., Thakuria, J. V., Zaranek, A. W., Clegg, T., Rosenbaum, A. M., Wu, Xiaodi, … Church, G. M. (2012). A public resource facilitating clinical use of genomes. *Proceedings of the National Academy of Sciences USA*, *109*, 11920-11927.
37. Eriksson N, Macpherson JM, Tung JY, Hon LS, Naughton B, et al. Web-based,

participant- driven studies yield novel genetic associations for common traits. *PLoS Genetics.* 2010; 6: e1000993.
38. Aitken Harris J, Vernon PA, Jang KL. Testing the differentiation of personality by intelligence hypothesis. *Personality and Individual Differences.* 2005; 38: 277–286.
39. Stokowski RP, Pant PV, Dadd T, Fereday A, Hinds DA, et al. A genomewide association study of skin pigmentation in a South Asian population. *American Journal of Human Genetics*. 2007; 81: 1119–1132.
40. Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, et al. Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nature Genetics.* 2007; 39: 1443–1452.
41. Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Ranar T, et al. Two newly identified genetic determinants of pigmentation in Europeans. *Nature Genetics.* 2008; 40: 835–837.
42. Han J, Kraft P, Nan H, Guo Q, Chen C, et al. A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. *PLoS Genetics*. 2008; 4: e1000074.
43. Liu F, Wollstein A, Hysi PG, Ankra-Badu GA, Spector TD, et al. Digital quantification of human eye color highlights genetic association of three new loci. *PLoS Genetics*. 2010; 6: e1000934.
44. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MW, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature.* 2010; 467: 832–838.
45. MacArthur DG, Seto JT, Raftery JM, Quinlan KG, Huttley GA, et al. Loss of ACTN3 gene function alters mouse muscle metabolism and shows evidence of positive selection in humans. *Nature Genetics.* 2007; 39; 1261–1265.
46. Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics.* 2010; 42: 937–948.
47. Payton A, Holland F, Diggle P, Rabbitt P, Horan M, et al. Cathepsin D exon 2 polymorphism associated with general intelligence in a healthy older population. *Molecular Psychiatry.* 2003; 8: 14–18.
48. Plomin R, Turic TM, Hill L, Turic DE, Stephens M, et al. A functional polymorphism in the succinate-semialdehyde dehydrogenase (aldehyde dehydrogenase 5 family, member A1) gene is associated with cognitive ability. *Molecular Psychiatry.* 2004; 9: 582–586.
49. Gosso MF, van Belzen M, de Geus EJC, Polderman JC, Heutink P, et al. Association between the CHRM2 gene and intelligence in a sample of 304 Dutch families. *Genes, Brain and Behavior.* 2006; 5: 577–584.
50. Gosso MF, de Geus EJC, van Belzen MJ, Polderman TJC, Heutink P, et al. The SNAP-25 gene is associated with cognitive ability: Evidence from a family-based study in two independent Dutch cohorts. *Molecular Psychiatry.* 2006; 11: 878–886.
51. Zinkstock JR, de Wilde O, van Amelsvoort TAMJ, Tanck MW, Baas F, et al. Association between the DTNBP1 gene and intelligence: A case-control study in young patients with schizophrenia and related disorders and unaffected siblings. *Behavioral and Brain Function.* 2007; 3: 19.
52. van den Oord EJCG, Kuo PH, Hartmann AM, Webb BT, Moller HJ, et al. Genomewide association analysis followed by a replication study implicates a novel candidate gene for neuroticism. *Archives of General Psychiatry.* 2008; 65: 1062–1071.
53. de Moor MHM, Costa PT, Terracciano A, Krueger RF, de Geus EJC, et al. Meta-analysis

of genome-wide association studies for personality. *Molecular Psychiatry*. Preprint.
54. Egan MF, Goldberg TE, Kolachana BS, Callicott JH, Mazzanti CM, et al. Effect of COMT Val108/158 Met genotype on frontal lobe function and risk for schizophrenia. *Proceedings of the National Academy of Sciences of the United States of America*. 2001; 98: 6917–6922.
55. Papassotiropoulos A, Stephan DA, Huentelman MJ, Hoerndli FJ, Craig DW, et al. (2006). Common Kibra alleles are associated with human memory performance. *Science,* 314: 475–478.
56. Need AC, Attix DK, McEvoy JM, Cirulli ET, Linney KN, et al. (2008). Failure to replicate effect of Kibra on human memory in two large cohorts of European origin. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics.* 2008; 147: 667–668.
57. Jost JT, Glaser J, Kruglanski AW, Sulloway FJ. Political conservatism as motivated social cognition. Psychological Bulletin. 2003; 129(3): 339–375.
58. Hirsh JB, DeYoung CG, Xu X, Peterson JB. Compassionate liberals and polite conservatives: Associations of agreeableness with political ideology and moral values. *Personality and Social Psychology Bulletin.* 2010; 36: 655–664.
59. Davis OSP, Butcher LM, Docherty SJ, Meaburn EL, Curtis CJC, et al. A three-stage genome-wide association study of general cognitive ability: Hunting the small effects. *Behavioral Genetics.* 2010; 40: 31–45.
60. Verweij KJH, Zietsch BP, Medland SE, Gordon SD, Benyamin B, Dale R, et al. A genome-wide association study of Cloninger's temperament scales: Implications for the evolutionary genetics of personality. *Biological Psychology.* 2010; 85: 306–317
61. Jablonski NG, Chaplin G. Human skin pigmentation as an adaptation to UV radiation. *Proceedings of the National Academy of Sciences of the United States of America*. 2010; 107: 8962–8968.
62. Cavalli-Sforza LL, Menozzi P, Piazza A. *The history and geography of human genes*. Princeton, NJ: Princeton University Press; 1994.
63. Frost P. European hair and eye color: A case of frequency-dependent sexual selection? *Evolution and Human Behavior*. 2006; 27: 85–103.
64. Cochran G, Harpending H. *The 10,000-year Explosion: How Civilization Accelerated Human Evolution*. New York, NY: Basic Books; 2009.
65. Rogers AR, Iltis D, Wooding S. Genetic variation at the MC1R locus and the time since loss of human body hair. *Current Anthropology*. 2004; 45: 105–108.
66. Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, et al. Localizing recent adaptive evolution in the human genome. *PLoS Genetics*. 2007; 3: e90.
67. Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, et al. Signals of recent positive selection in a worldwide sample of human populations. *Genome Research*. 2009; 19: 826–837.
68. Bajema, C. (1963). Estimation of the direction and intensity of natural selection in relation to human intelligence by means of the intrinsic rate of natural increase. *Biodemography and Social Biology*, *10*, 175-187.
69. Van Valen, L. (1974). Brain size and intelligence in man. *American Journal of Physical Anthropology*, *40*, 417-423.
70. Blume, M. (2009). The reproductive benefits of religious affiliation. In Voland, E. & Schiefenhovel, W. (Eds.), *The Biological Evolution of Religious Mind and Behavior* (pp.

117–126). Berlin, Germany: Springer.
71. Maynard Smith J, Price GR. The logic of animal conflict. *Nature.* 1973; 246: 15–18.
72. Penke L, Denissen JJA, Miller GF. The evolutionary genetics of personality. *European Journal of Personality.* 2007; 21: 549–587.
73. Kimura M. *The Neutral Theory of Molecular Evolution*. Cambridge, UK: Cambridge University Press; 1983.
74. Simonson TS, Yang Y, Huff CD, Yun H, Qin G, et al. Genetic evidence for high-altitude adaptation in Tibet. *Science.* 2010; 329: 72–74.
75. Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, et al. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science*. 2010; 329: 75–77.
76. Beall CM, Cavalleri GL, Deng L, Elston RC, Gao Y, et al. Natural selection on EPAS1 (HIF2) associated with low hemoglobin concentration in Tibetan highlanders. *Proceedings of the National Academy of Sciences of the United States of America*. 2010; 107: 11459–11464.
77. Dolgin E. Personalized investigation. *Nature Medicine*. 2010; 16: 953–955.
78. Lunshof JE, Bobe J, Aach J, Angrist M, Thakuria JV, et al. Personal genomes in progress: From the Human Genome Project to the Personal Genome Project. *Dialogues in Clinical Neurosciences*. 2010; 12: 47–60.
79. Barnett JH, Scoriels L, Munafò MR. Meta-analysis of the cognitive effects of the catechol-O-methyltransferase gene Val158/108Met polymorphism. *Biological Psychiatry*. 2008; 64: 137-144.
80. Ioannidis JPA. Non-replication and inconsistency in the genome-wide association setting. *Human Heredity.* 2007; 64: 203–13.
81. Carpenter S. Psychology's bold initiative. *Science.* 2012; 335: 1558–1561.
82. Boomsma D, Busjahn A., Peltonen L. Classical twin studies and beyond. *Nature Reviews Genetics.* 2002; 3: 872–882.
83. Visscher PM. Sizing up human height variation. *Nature Genetics.* 2008; 40:489-490.
84. Bouchard, TJ. Genetic influence on human psychological traits. *Current Directions in Psychological Science.* 2004; 13(4): 148–151.
85. Cesarini D, Johannesson M, Magnusson P, Wallace B. The behavioral genetics of behavioral anomalies. *Management Science*. 2012; 58: 21-34.
86. Cesarini D, Johannesson M, Lichtenstein P, Sandewall O, Wallace B. Genetic variation in financial decision-making. *Journal of Finance*. 2010; 65: 1725-1754.
87. Cesarini D, Dawes CT, Fowler J, Johannesson M, Lichtenstein P, Wallace B. Heritability of cooperative behavior in the trust game. *Proceedings of the National Academy of Sciences*. 2008; 105: 3271-3276.
88. Taubman P. The determinants of earnings: Genetics, family, and other environments: A study of white male twins. *American Economic Review*. 1976; 66: 858-870.
89. Plomin R, Owen MJ, McGuffin P. The genetic basis of complex human behaviors. *Science*. 1994; 264: 1733–1734.
90. Barnea A, Cronqvist H, Siegel S. Nature or Nurture: What Determines Investor Behavior? *Journal of Financial Economics.* 2010; 98: 583–604.
91. Jackson DN. *Multidimensional Aptitude Battery II*. 2nd ed. Port Huron, MI: Sigma Assessment Systems; 1998.

92. Costa PT, McCrae RR. *NEO Personality Inventory–Revised (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources; 1992.

Figure 1. R.A. Fisher's geometric model of adaptation.[21] **A** is the current mean phenotype of the population, **A'** is the mean phenotype that would result if the mutation denoted by the arrow were to be instantly fixed, and **O** is the new optimum favored by natural selection. The narrow distribution of trait 1 values around **A** is the situation that would prevail under strong stabilizing selection, while the broad distribution would prevail under weak stabilizing selection.

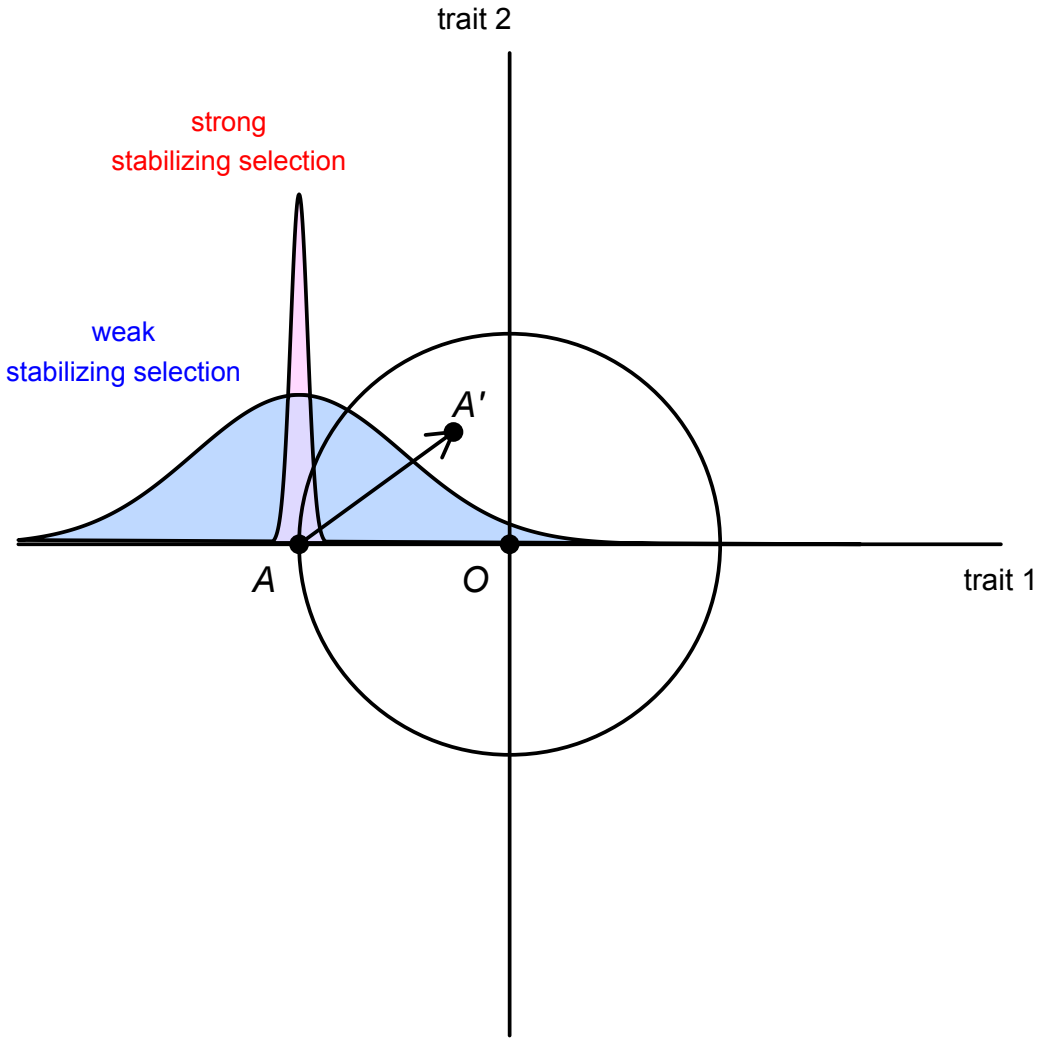

Figure 2. Examples of directed acyclic graphs containing a "collider" (the common effect of two or more causes).[33] Conditioning on observing a collider alters the apparent covariation among the causes; for example, two independent causes that are uncorrelated when all observations are considered can appear to be negatively correlated when only observations containing the collider are considered.

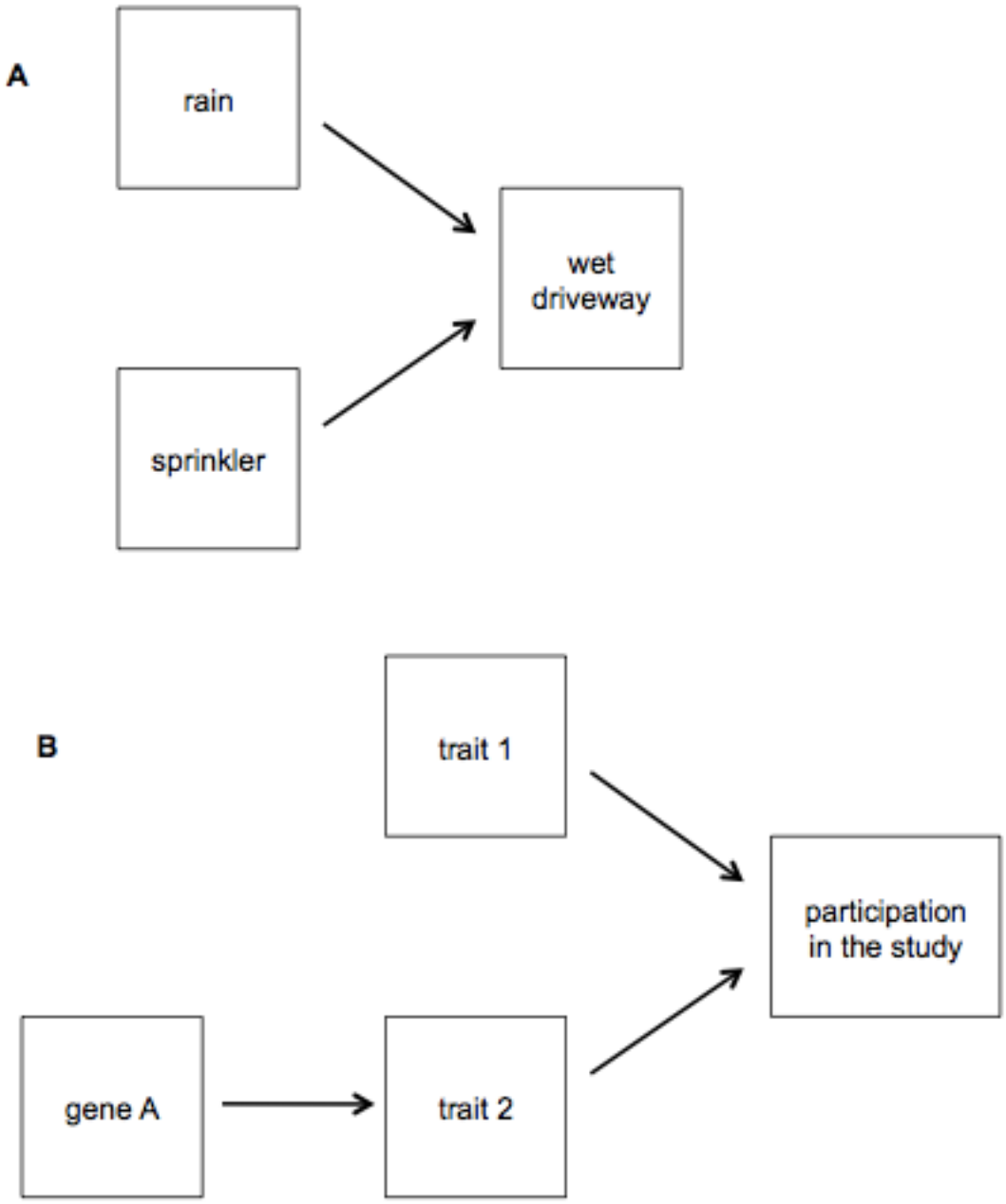

Figure 3. An illustration of how the reliability (measurement error, denoted here as Rho) of a phenotype affects the relationship between effect size of a genetic association and the sample size required to achieve 50% statistical power to detect the effect at the genome-wide significance threshold of 5×10^{-8}. For example, if one expects a genotype to explain 0.4% of the variance in a trait ($R^2 = .004$), then a sample of about 10,000 subjects is required to achieve 50% power when reliability is 0.80, but a sample of 20,000 subjects is required if reliability is 0.40. That is, with a sample of 20,000 instead of 10,000, instruments that are only one-quarter as reliable provide the same power to detect the effect.

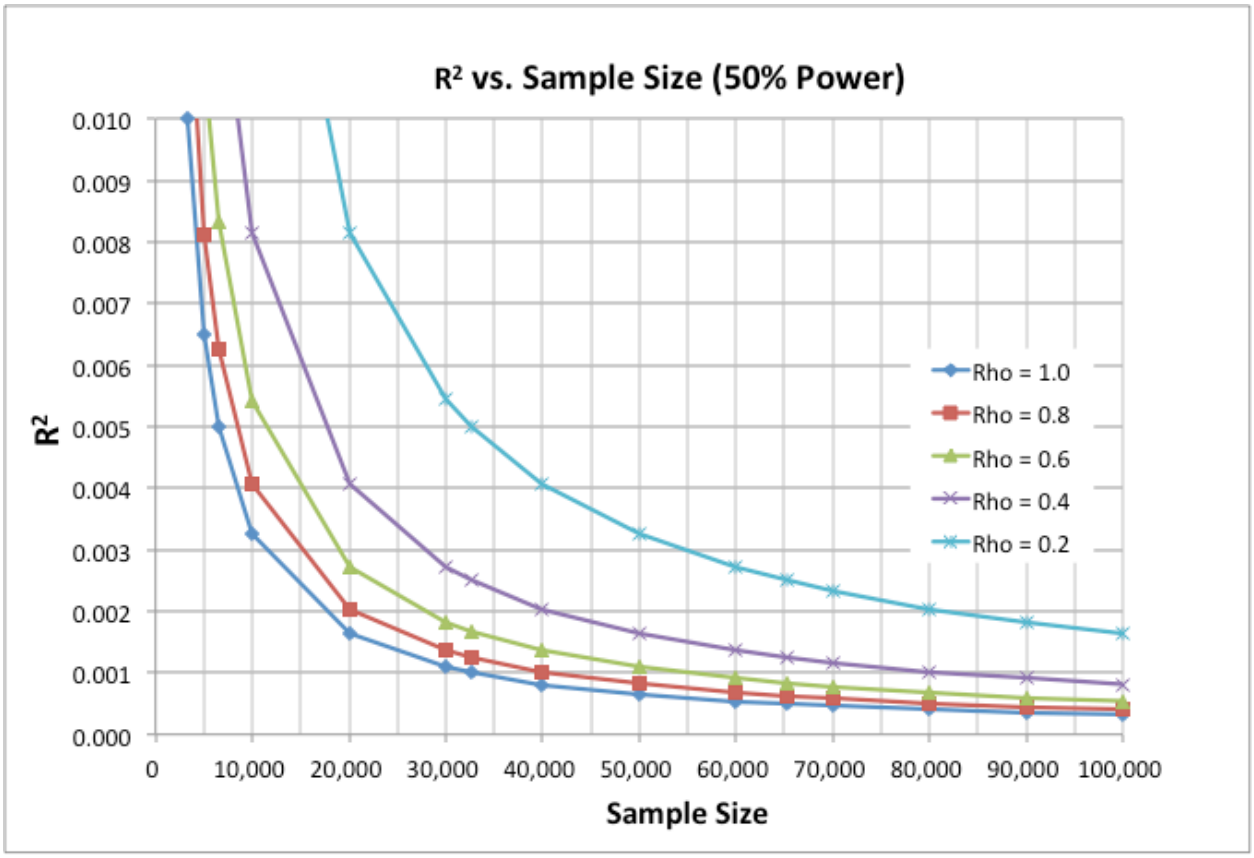

Table 1. Heritabilities of Selected Medical, Physical, and Behavioral Traits

Phenotype	Heritability	Source
Medical and Physical Traits:		
Lipoprotein A level (age 17)	95%	Boomsma et al.[82]
LDL Cholesterol level (age 44)	69%	Boomsma et al.[82]
HDL Cholesterol level (age 44)	67%	Boomsma et al.[82]
Heart rate (age 17)	44%	Boomsma et al.[82]
Respiration rate (age 44)	61%	Boomsma et al.[82]
Testosterone level (age 17)		Boomsma et al.[82]
Males	66%	
Females	41%	
Birth weight	10%	Boomsma et al.[82]
Height (ages 16–adult)	80%*	Visscher et al.[83]
Behavioral Traits:		
Problem behavior (age 3)		Boomsma et al.[82]
Externalizing		
Males	49%	
Females	73%	
Internalizing		
Males	61%	
Females	66%	
Personality Traits (adults)		Bouchard[84]
Neuroticism	48%	
Extraversion	54%	
Openness to Experience	57%	
Agreeableness	42%	
Conscientiousness	49%	
General Cognitive Ability (age 18)	81%	Boomsma et al.[82]
Boredom susceptibility (age 18)	50%	Boomsma et al.[82]
Anxiety (age 18)	54%	Boomsma et al.[82]
Depression (age 18)		Boomsma et al.[82]
Males	39%	
Females	53%	
Smoking (yes/no, at age 18)		Boomsma et al.[82]
Males	66%	
Females	32%	
Alcohol Use (yes/no, at age 18)		Boomsma et al.[82]
Males	48%	
Females	75%	
Sports Participation (age 18)	47%	Boomsma et al.[82]
Religiosity (adults)	38%	Bouchard[84]

Specific religion practiced (age 18)	0%	Boomsma et al.,[82] Bouchard[84]
Conservatism (adults)	55%	Bouchard[84]
Risk Attitudes		Cesarini et al.[85]
General willingness to take risk	21%	
Willingness to take financial risk	26%	
Risk aversion	34%	
Portfolio volatility	25%	Cesarini et al.[86]
Cooperation		Cesarini et al.[87]
Trust	15%	
Trustworthiness	18%	
Income (single year)	38%	Taubman[88]
Income (single year)		Benjamin et al.[4]
Men	37%	
Women	28%	
Education (years)	28%	Taubman[88]

Behavioral Traits – Estimates Corrected for Measurement Error:

Risk Attitudes		Cesarini et al.[85]
General willingness to take risk	35%	
Willingness to take financial risk	37%	
Risk aversion	54%	
Income (20-year average)		Benjamin et al.[4]
Men	58%	
Women	46%	

* Estimated from genome-wide SNP data from twin and sibling pairs in Australia.

Notes: Estimates are averages of male and female heritabilities except when heritabilities are provided separately for both sexes (these are cases in which heritability differs by a large amount between males and females). Except in the third section, heritability estimates are not adjusted for differences in measurement error, longitudinal stability, or test-retest reliability of the phenotypes. Heritabilities may also vary with age; e.g., general cognitive ability becomes more heritable with age. Summaries of heritabilities of these and other phenotypes may be found in Plomin et al.,[89] Boomsma et al.,[82] Bouchard,[84] and Barnea et al.[90]

Table 2. Characteristics of the sample in age, sex, general cognitive ability (MAB scales), and personality traits (NEO Five-Factor Inventory scales).

Trait	mean	SD
age (years)	25.2	6.44
sex	67.6% female	
MAB Arithmetic	.797 (0)	.836 (1)
MAB Similarities	1.054 (0)	.601 (1)
MAB Vocabulary	1.386 (0)	.891 (1)
NEO Neuroticism (college, female)	21.90 (25.83)	8.38 (7.59)
NEO Neuroticism (adult, female)	18.71 (20.54)	9.13 (7.61)
NEO Neuroticism (college, male)	18.53 (22.49)	10.04 (7.92)
NEO Neuroticism (adult, male)	18.84 (17.60)	10.46 (8.61)
NEO Extraversion (college, female)	30.10 (31.27)	6.89 (5.64)
NEO Extraversion (adult, female)	29.19 (28.16)	7.55 (5.82)
NEO Extraversion (college, male)	29.08 (29.22)	6.10 (5.97)
NEO Extraversion (adult, male)	29.70 (27.22)	8.64 (5.85)
NEO Openness (college, female)	34.02 (27.94)	6.57 (5.72)
NEO Openness (adult, female)	34.42 (26.98)	5.57 (5.87)
NEO Openness (college, male)	31.79 (27.62)	6.57 (6.08)
NEO Openness (adult, male)	31.36 (27.09)	7.04 (5.82)
NEO Agreeableness (college, female)	33.80 (31.00)	5.51 (5.33)
NEO Agreeableness (adult, female)	34.42 (33.76)	4.71 (4.74)
NEO Agreeableness (college, male)	31.46 (28.76)	6.05 (5.24)
NEO Agreeableness (adult, male)	32.00 (31.93)	5.70 (5.03)
NEO Conscientiousness (college, female)	33.64 (31.02)	7.40 (6.53)
NEO Conscientiousness (adult, female)	32.29 (35.04)	7.15 (5.78)
NEO Conscientiousness (college, male)	30.17 (30.21)	6.54 (7.19)
NEO Conscientiousness (adult, male)	33.33 (34.10)	8.04 (5.95)

The summary statistics reported in the respective manuals are given in parentheses next to the corresponding sample statistics. The MAB scores were scaled as standard normal using the tables in the MAB manual.[91] The NEO summary statistics were calculated for participants between the ages of 18 and 22 for purposes of comparison with the college norms in the NEO manual[92] and for participants age 30 and over for comparison with the adult norms.

Table 3. Association results for pigmentation phenotypes.

trait	reported SNP	proxy SNP	r^2	minor allele	sample MAF	HapMap MAF	effect size	*p*-value	gene
eye darkness	rs12913832			A	.222	.208	.998	$2x10^{-68}$	*HERC2*
eye darkness	rs12896399	rs1075830	.615	A	.460	.308	.167	.003	*SLC24A*
eye darkness	rs1393350			A	.266	.192	–.154	.02	*TYR*
eye darkness	rs1408799			T	.313	.300	.095	.11	*TYRP1*
hair darkness	rs12913832			A	.223	.208	.840	$1x10^{-13}$	*HERC2*
hair darkness	rs12896399	rs1075830	.640	A	.460	.308	.372	$9x10^{-5}$	*SLC24A4*
hair darkness	rs12821256			C	.095	.142	–.352	.03	*KITLG*
red hair	rs1805007			T	.076	.147	7.44	$2x10^{-6}$	*MC1R*
red hair	rs1015362			T	.278	.233	.507	.09	*ASIP*
freckling	rs1805007			T	.076	.147	.613	$6x10^{-6}$	*MC1R*
freckling	rs1042602			A	.346	.417	–.223	.005	*TY R*
freckling	rs2153271	rs1416742	.949	G	.384	.373	–.139	.07	*BNC2*
freckling	rs619865			A	.098	.108	.178	.15	*ASIP*
skin darkness	rs1805007			T	.076	.147	–.267	.005	*MC1R*
skin darkness	rs1042602			A	.346	.417	–.118	.03	*TY R*
skin darkness	rs619865			A	.098	.108	–.156	.07	*ASIP*

Eye darkness was reported on a 3-point scale. Hair darkness was recorded on 9-point scale. Red hair was recorded as a dichotomous trait, and its effect size is reported as an odds ratio. Freckling and skin darkness were recorded on 5-point scales. All effect sizes for non-dichotomous traits are reported as the expected change in trait value per each additional copy of the minor allele. All alleles are coded according to NCBI build 36 coordinates on the forward strand.

Table 4A. Association results for physical phenotypes.

trait	reported SNP	proxy SNP	r^2	minor allele	sample MAF	HapMap MAF	effect size	*p*-value	gene
standing height	rs7460090			C	.134	.117	- .188	.07	*SDR16C5*
standing height	rs237743			A	.231	.308	.175	.04	*ZNFX1*
standing height	rs6439167			T	.201	.183	- .191	.03	*C3orf47*
standing height	rs889014			T	.347	.375	- .124	.10	*BOD1*
standing height	rs7274811	rs3213183	.692	A	.304	.267	- .140	.07	*ZNF341*
standing height	rs7759938	rs369065	1	C	.332	.364	.172	.02	*LIN28B*
standing height	rs3764419	rs9890032	.982	G	.401	.375	- .188	.009	*ATAD5/RNF135*
standing height	rs3791675			T	.228	.275	- .305	4×10^{-4}	*EFEMP1*
standing height	rs724016			G	.428	.483	.121	.10	*ZBTB38*
standing height	rs1351394	rs7968682	.983	T	.499	.517	- .120	.10	*HMGA2*
strength	rs1815739	rs540874	1	A	.428	.458	.252	.006	*ACTN3*

Effect sizes for height are reported in standard deviation units. Note that these effect sizes tend to be inflated because of the "winner's curse." Strength was reported on a 5-point scale.

ble 4B. Association results for the behavioral phenotypes with previously reported SNPs.

ait	reported SNP	proxy SNP	r^2	minor allele	sample MAF	HapMap MAF	effect size	*p*-value	gene
eneral cognitive ability	rs2760118	rs7775073	.982	G	.316	.317	.062*	.42	*ALDH5A1*
eneral cognitive ability	rs324650			T	.464	.467	.026	.72	*CHRM2*
eneral cognitive ability	rs363050			G	.444	.475	- .027	.72	*SNAP-25*
eneral cognitive ability	rs17571	rs17834326	.781	A	.083	.083	- .051*	.70	*CTSD*
eneral cognitive ability	rs760761	rs2619545	1	C	.196	.192	- .033*	.72	*DTNBP1*
onscientiousness	rs2576037	rs7233515	.879	A	.400	.408	- .038	.60	*KATNAL2*
euroticism	rs12883384			A	.410	.317	- .014*	.85	*MAMDC1*
aired-associate recognition	rs17070145			T	.338	.267	.065	.37	*KIBRA*
-back accuracy	rs4680			A	.449	.517	.027	.72	*COMT*

Effect sizes are reported in sample standard deviation units. An asterisk indicates that the estimated effect in our study had a sign opposite to what had been previously reported.

Table 5. Novel association results for behavioral phenotypes.

trait	reported SNP	minor allele	sample MAF	HapMap MAF	effect size	*p*-value
liberal vs conservative	rs10952668	T	.458	.392	.552 (.478)	2×10^{8} (1×10^{-6})
gambling frequency	rs1402494	G	.206	.241	.278 (.276)	3×10^{8} (6×10^{-8})

Liberal vs conservative was reported on a 7-point scale. Gambling frequency was reported on a 5-point scale. Effect sizes and *p*-values after adjustment for general cognitive ability, Openness, Neuroticism, and Agreeableness are given parenthetically. Note that effect estimates may be inflated as a result of the winner's curse.

SUPPORTING ONLINE MATERIAL

Measures

Table S1 lists all of the phenotypes measured in this study. Any variable taking more than ten values was regarded as quantitative rather than polytomous (ordered categorical). A parenthetical N in Table S1 indicates that we were able to remove sex differences in mean and variance from a quantitative variable and then use a quantile transformation to render the resulting scores normally distributed. These transformations should increase statistical power to detect genetic associations for traits showing sex differences. Below are details on some of the behavioral phenotypes whose labels in Table S1 are not self-explanatory. (Except as noted, all economic games were played with real monetary incentives.)

3-back. Participants viewed a succession of words, each new word appearing every three seconds. Participants were instructed to indicate as quickly and accurately as possible whether each word matched the word seen three items previously. This task has often been employed as an indicator of working memory capacity.[1]

Barratt Impulsiveness Scale (BIS). This self-report has been found to measure three distinct factors (inattention, motor impulsiveness, and lack of planning).[2] We used the sum of these three factor scores as a measure of this self-report's general factor.

Cambridge Face Memory Test (CFMT). Participants studied three photos of each of six target human faces and were then tested with a series of forced-choice items, each consisting of three faces, one of which was a target. This test has been shown to be a sensitive measure of prosopagnosia (a specific deficit in recognizing other people by their facial features) and also normal variability in the ability to recognize faces.[3, 4]

Dictator game. Each participant was asked to imagine being randomly and anonymously paired with another participant. The participant was then asked to allocate ten dollars between the members of the pair. How much of the ten dollars each participant is willing to give away to the other person in this task has been used as a measure of the participant's heritable altruistic tendencies.[5,6] Because the distribution of allocation was almost bimodal, nearly all participants giving away either zero or five dollars, we treated this phenotype as dichotomous; all participants who gave anything at all were given the higher score.

Discounting the future. Participants were presented a set of choices between smaller prompt rewards and larger delayed rewards. Temporal discount rates inferred in this way, have been found to be associated with substance abuse and other outcomes.[7]

General cognitive ability. We combined the following indicators into a standardized cognitive ability composite: (1) a short form of Raven's Advanced Progressive Matrices[8], a measure of abstract reasoning ability; (2) the Arithmetic, Similarities, and Vocabulary subtests of the Multidimensional Aptitude Battery (MAB), which measure verbal ability; and (3) accuracy on a forced-choice version of the Shepard-Metzler Mental Rotation task (SMMR) a measure of spatial

ability.[9]

Inattentional blindness. Participants watched a video of two teams of three players, one team wearing white shirts and the other wearing black shirts, who moved around erratically in an elevator lobby. The passes were either bounce passes or aerial passes; players would also dribble the ball, wave their arms, and make other movements. After about 45 seconds, a person wearing a gorilla costume walked through the action. The relatively high proportion of participants who report not seeing the gorilla at all is generally regarded as surprising.[10] The causes of individual differences in this task are unknown. This finding has achieved wide publicity, so we treated any participant who reported having seen or heard of it as a missing data point; others were classified as either noticing or missing the gorilla.

Loss aversion. Participants were presented with a set of choices between (1) receiving nothing or (2) a 50% chance of gaining an amount **x** and a 50% chance of losing an amount **y**. This is a standard measure of aversion to suffering financial losses.[11] The main loss aversion measure involved real money stakes; a separate measure was made with fictitious higher stakes.

NEO Five-Factor Inventory. A 60-item self-report instrument with 12 items measuring each of the following five personality factors, which constitute the most widely accepted factorization of personality: Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness.

Paired-associate recognition. After studying a series of 25 word pairs, participants were given a recognition test in which they were given the first word in a pair and had to choose the second from among four presented alternatives. The words in the pairs were abstract and unrelated, and the distractor words were other words from the experiment, making this task difficult.[12]

Religiosity. We administered a standard scale to measure religiousness.[13]

Risk aversion. Participants were presented with a set of choices between (1) a 100% chance of receiving an amount **x** or (2) a 50% chance of receiving an amount **y > x** and a 50% chance of receiving nothing. Risk-averse choices involved turning down a larger expected value prospect (e.g., 50% chance of receiving $10) in favor of a smaller guaranteed amount (e.g., 100% chance of receiving $4). This is a standard measure.[11]

Shape memory. In a study phase, participants were presented a series of irregular shapes, one at a time. In a test phase, participants then had to press one key if the shape they were viewing had already been presented in the study phase, another key if it was new.

Social attitudes. Items asking for attitudes toward abortion, alcohol consumption, and other social issues were taken from an existing scale.[14] Because the factor model postulated by the scale's authors did not fit our data well, we analyzed each item separately.

Spatial memory. In a study phase, participants viewed a circular array of gray dots. Several of the dots briefly turned black, one at a time. The display continued in a test phase, where participants indicated whether each black dot had also turned black during the study phase.

Serial Reaction Time Task (SRTT). Participants viewed a line of four squares. During each of 384 trials a black diamond briefly appeared in one of the squares, and in response participants had to press one of four corresponding keys, using four fingers of their preferred hand. Unbeknownst to the participants, a fixed subsequence of the stimuli appeared repeatedly throughout the task, alternating with runs of stimuli chosen at random. Response time (RT) tends to decrease with each successive presentation of the repeating subsequence, although most participants do not consciously notice the repetition. The mean difference in RT between the repeating stimuli and the random stimuli was taken as a measure of implicit skill learning.

Utilitarianism. Participants were presented with a set of moral dilemmas in which participants rated on a 1–5 scale the appropriateness of a "utilitarian" response to the situation.[15] A typical item: "You are at the wheel of a runaway trolley quickly approaching a fork in the tracks. On the tracks extending to the left is a group of five railway workmen. On the tracks extending to the right is a single railway workman. If you do nothing the trolley will proceed to the left, causing the deaths of the five workmen. The only way to avoid the deaths of these workmen is to hit a switch on your dashboard that will cause the trolley to proceed to the right, causing the death of the single workman. Is it appropriate for you to hit the switch in order to avoid the deaths of the five workmen?"

Verbal fluency. Participants were given one minute to utter as many distinct words as possible beginning with a certain letter. Person names, places, and numbers were not counted. The letters F, A, and S were used. The counts of the uttered words beginning with these letters appeared to be equal indicators of a common factor after standardization.

Vividness of Visual Imagery Questionnaire (VVIQ). Participants were told to visualize certain scenes or persons and rate the vividness of distinct aspects of the mental image.[16]

To make the choices in the economic tasks (intertemporal choice, fairness, loss aversion, risk aversion) meaningful to the participants, we told them at the outset that their choices in these tasks might be implemented with real money. The payment policy worked as follows: At the end of the session, each participant rolled a six-sided die. If he or she rolled a six, then a further random draw was conducted to choose one item from the behavioral-economic tasks, which was then fulfilled for that participant. For example, suppose that the participant rolled a six, and the second draw selected an item from the discounting task. If the participant expressed a preference for **x** dollars 30 days from now over **y** dollars 60 days from now, then the participant was written a check for **x** dollars dated 30 days from the date of the phenotyping session (or given a debit card for the same amount that would be activated on the same date). Any losses suffered in the loss aversion task came out of $5 cash given to each participant at the beginning of the session. This $5 was given in addition to the advertised $50 compensation.

The MAB subtests were scored according to the instructions in the test manual.[17] Factor analyses of the BIS, the NEO, religiosity, utilitarianism, verbal fluency, and the VVIQ resulted in solutions with nonzero uniquenesses. For these phenotypes we estimated factor scores by Bartlett's method, which is equivalent to maximum likelihood (ML) if the uniquenesses are

normally distributed. A few participants were missing some data as a result of omits, photocopying errors, computer failures, and other administrative issues. Participants' factor scores were treated as missing if they responded to fewer than half of a scale's indicators. We used the OpenMx package in R to perform all factor analyses.[18] All coding and scoring of phenotypic measures was performed blind to participant genotypes.

Parameters describing the responses of each participant during the behavioral-economic tasks were estimated by ML, assuming choice error drawn from an extreme-value distribution. For example, an "interest rate" for discounting utility flows over time was estimated for each person and used as the phenotype for the discounting task.

DNA Collection, Extraction, and Genotyping

At two points during the phenotyping session, participants provided DNA samples by washing their mouths with 10 ml of Scope mouthwash, which dislodges loose cells, and then releasing the mouthwash into a Nalgene bottle. Samples were stored either in a freezer at –20° C or in packed dry ice until DNA extraction. Genomic DNA was extracted using a QIAamp DNA Blood Mini Kit according to the manufacturer's recommended protocol.

Genomic DNA samples normalized to 50 ng/μl were genotyped at either Stanford Genome Technology Center (SGTC) or Expression Analysis (EA) in Durham, North Carolina, in four batches, using the Affymetrix Genome-Wide Human SNP Array 6.0. SNP genotypes were called using the Birdseed v2 algorithm applied to each batch individually. The median call rate before application of quality-control criteria was 99.64%. Between-batch reproducibility was assessed by genotyping both samples provided by each of two participants. Average genotype concordance between replicates was 99.7%.

Our quality-control criteria at this stage excluded all participants missing more than 7% of their genotypic data, all SNPs with minor allele frequency (MAF) less than .05, all SNPs deviating from Hardy-Weinberg equilibrium at a significance threshold of 5×10^{-8}, and all SNPs missing more than 5% of their calls. We then computed the principal components of the resulting genotype matrix with the program EIGENSTRAT.[19] To guard against population stratification, all participants who were more than six standard-deviation units from the origin on any of the top 10 PCs were iteratively excluded (a total of 14 participants).

After application of all quality control measures, the final cleaned dataset included 401 individuals and 661,107 SNPs. Nine statistically significant principal components at a significance threshold of .05 were found. The components corresponding to the fourth and fifth largest eigenvalues weakly distinguished the two genotyping laboratories, despite the application of our quality-control steps. The first, second, third, and sixth components were significantly correlated with the geographical distance of grandparental origin from England. The seventh component tended to spread out individuals reporting non-British grandparents, whereas the eighth component tended to separate those reporting two or more British grandparents from those reporting one or none. The ninth component tended to spread out individuals reporting

British grandparents, perhaps reflecting structure within Britain. To control for remaining stratification, we included all nine significant principal components as covariates in the tests for SNP-trait association.

Table S1. All phenotypes measured. Any phenotype measured in paper mode was administered as a traditional paper-and-pencil test. Self-report refers to questionnaire data recorded either on paper forms or a SurveyMonkey questionnaire. Phenotypes measured in computer mode were implemented as PsyScope tasks requiring participants to provide keyboard input. Physical traits were directly measured by an experimenter using either a measuring tape or a standard bathroom scale. Audio refers to sound-recorded data that was later transcribed and coded.

Phenotype	Mode	Scale
3-back accuracy	computer	quantitative (N)
3-back RT	computer	quantitative (N)
acne severity as adolescent	self-report	polytomous
acne severity as adult	self-report	polytomous
acne severity overall	self-report	polytomous
alcohol consumption frequency (last 12 months)	self-report	polytomous
alcohol drinks per drinking occasion	self-report	quantitative
alcohol total drinks in last year	self-report	quantitative (N)
allergic to animals	self-report	dichotomous
allergic to drugs	self-report	dichotomous
allergic to food	self-report	dichotomous
allergies (any)	self-report	dichotomous
anticipated remaining life expectancy	self-report	quantitative (N)
asthma as adult	self-report	dichotomous
asthma as child	self-report	dichotomous
athleticism	self-report	polytomous
attitude toward abortion on demand	self-report	polytomous
attitude toward alcohol	self-report	polytomous
attitude toward attention-drawing clothes	self-report	polytomous
attitude toward being the center of attention	self-report	polytomous
attitude toward being the leader of groups	self-report	polytomous
attitude toward big parties	self-report	polytomous
attitude toward capitalism	self-report	polytomous
attitude toward castration as sex crime punishment	self-report	polytomous
attitude toward death penalty for murder	self-report	polytomous
attitude toward doing athletic activities	self-report	polytomous
attitude toward dressing well at all times	self-report	polytomous
attitude toward education	self-report	polytomous
attitude toward exercising	self-report	polytomous
attitude toward getting along well with others	self-report	polytomous
attitude toward illegal drugs	self-report	polytomous
attitude toward legalized gambling	self-report	polytomous
attitude toward loud music	self-report	polytomous
attitude toward making racial discrimination illegal	self-report	polytomous
attitude toward open-door immigration	self-report	polytomous
attitude toward organized religion	self-report	polytomous
attitude toward playing chess	self-report	polytomous

attitude toward playing organized sports	self-report	polytomous
attitude toward public speaking	self-report	polytomous
attitude toward reading books	self-report	polytomous
attitude toward rollercoaster rides	self-report	polytomous
attitude toward smoking	self-report	polytomous
attitude toward voluntary euthanasia	self-report	polytomous
back pain	self-report	dichotomous
BIS inattention	self-report	quantitative (N)
BIS general	self-report	quantitative (N)
BIS motor	self-report	quantitative (N)
BIS nonplanning	self-report	quantitative (N)
body mass index	measured	quantitative (N)
body type (scrawny to obese)	self-report	polytomous
ca_eine mg per day	self-report	quantitative
CFMT	computer	quantitative (N)
cigarette packs per day	self-report	polytomous
cleft chin	self-report	dichotomous
co_ee cups per day	self-report	polytomous
corrective lenses needed currently	self-report	dichotomous
corrective lenses needed at any time	self-report	dichotomous
curl tongue	self-report	dichotomous
Democrat vs. Republican	self-report	polytomous
dental braces worn (ever)	self-report	dichotomous
dental braces worn or needed (ever)	self-report	dichotomous
dictator game	self-report	dichotomous
dimples	self-report	dichotomous
discounting the future	self-report	quantitative (N)
drink alcohol (ever)	self-report	dichotomous
earlobes free (vs. hanging)	self-report	dichotomous
evening person	self-report	dichotomous
exercise amount per week	self-report	polytomous
exercise intensity	self-report	polytomous
exercise regularly	self-report	dichotomous
eye color	self-report	polytomous
facial hair color	self-report	polytomous
facial hair color (red vs. not red)	self-report	dichotomous
farsighted	self-report	dichotomous
first toe longer than second toe	self-report	dichotomous
floss teeth regularly	self-report	dichotomous
freckles on face	self-report	polytomous
gambling frequency	self-report	polytomous
general cognitive ability	multiple	quantitative (N)
hair color	self-report	polytomous
hair color (red vs. not red)	self-report	dichotomous
hair curliness	self-report	polytomous
hair on middle segment of any finger	self-report	dichotomous

happiness sumscore	self-report	quantitative (N)
hay fever	self-report	dichotomous
heterosexual	self-report	dichotomous
hitchhiker's thumb	self-report	dichotomous
hours of sleep average	self-report	quantitative
hours of sleep last night	self-report	quantitative
illegal drug use	self-report	polytomous
inattentional blindness	computer	dichotomous
in-person contact with family or very close friends	self-report	dichotomous
last doctor's appointment for checkup	self-report	polytomous
liberal vs conservative	self-report	polytomous
loss aversion	self-report	quantitative
MAB Arithmetic	paper	quantitative (N)
MAB Similarities	paper	quantitative (N)
MAB Vocabulary	paper	quantitative (N)
memory problems	self-report	dichotomous
migraines at any time	self-report	dichotomous
migraine frequency	self-report	polytomous
migraine within last 12 months	self-report	dichotomous
morning person	self-report	dichotomous
multivitamin supplement	self-report	dichotomous
nearsighted	self-report	dichotomous
NEO Agreeableness	self-report	quantitative (N)
NEO Conscientiousness	self-report	quantitative (N)
NEO Extraversion	self-report	quantitative (N)
NEO Neuroticism	self-report	quantitative (N)
NEO Openness	self-report	quantitative (N)
paired-associate recognition	computer	quantitative (N)
percentage of income saved over last 3 years	self-report	quantitative
physical attractiveness	self-report	polytomous
quality of sleep	self-report	polytomous
RAPM	computer	quantitative (N)
religiosity	self-report	quantitative
right-handed	self-report	dichotomous
risk aversion	self-report	quantitative (N)
seat belt use	self-report	polytomous
shape memory accuracy	computer	quantitative (N)
shape memory response time	computer	quantitative (N)
sitting height	measured	quantitative (N)
skin color and sun exposure response	self-report	polytomous
SMMR accuracy	computer	quantitative (N)
SMMR response time	computer	quantitative (N)
smoked cigarette (ever)	self-report	dichotomous
soda cups per day	self-report	polytomous
spatial memory accuracy	computer	quantitative (N)
spatial span response time	computer	quantitative (N)

SRTT accuracy	computer	quantitative
SRTT overall RT	computer	quantitative (N)
SRTT improvement in RT	computer	quantitative (N)
standing height	measured	quantitative (N)
strength	self-report	polytomous
stress level within last 12 months	self-report	polytomous
sunscreen or protective clothing use	self-report	polytomous
tea cups per day	self-report	polytomous
time woke up this morning	self-report	quantitative (N)
tobacco use frequency (current)	self-report	polytomous
tobacco user (current)	self-report	dichotomous
tobacco user (ever)	self-report	dichotomous
unprotected sex	self-report	polytomous
utilitarianism	self-report	quantitative (N)
verbal fluency	audio	quantitative (N)
vision quality (uncorrected)	self-report	polytomous
VVIQ	self-report	quantitative (N)
weight	measured	quantitative (N)
weight (maximum)	self-report	quantitative (N)
widow's peak	self-report	dichotomous

References

1. Gray JR, Chabris CF, Braver TS. Neural mechanisms of general fluid intelligence. *Nature Neuroscience* 2003; 6: 316–322.
2. Patton JH, Stanford MS, Barratt ES. Factor structure of the Barratt impulsiveness scale. *Journal of Clinical Psychology.* 1995; 51: 768–774.
3. Duchaine B, Nakayama K. The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia. 2006;* 44: 576–585.
4. Wilmer JB, Germine L, Chabris CF, Chatterjee G, Williams M, et al. Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of Sciences of the United States of America.* 2010; 107: 5238–5241.
5. Knafo A, Israel S, Darvasi A, Bachner-Melman R, Uzefovsky F, et al. Individual differences in allocation of funds in the dictator game associated with length of the arginine vasopressin 1a receptor RS3 promoter region and correlation between RS3 length and hippocampal mRNA. *Genes, Brain and Behavior*. 2008; 7: 266–275.
6. Cesarini D, Dawes CT, Johannesson M, Lichtenstein P, Wallace B. Experimental game theory and behavior genetics. *Annals of the New York Academy of Sciences.* 2009; 1167: 66–75.
7. Chabris CF, Laibson DI, Schuldt, JP. Intertemporal choice. In: Durlauf SN, Blume LE, eds. *The new Palgrave Dictionary of Economics*. 2nd ed. London, UK: Palgrave Macmillan; 2008: 536-542.
8. Bors, DA, Stokes, TL. Raven's advanced progressive matrices: Norms for first-yearuniversity students and the development of a short form. *Educational and Psychological Measurement.* 1998; 58: 382–398.
9. Shepard RN, Metzler J. Mental rotation of three-dimensional objects. *Science. 1971;* 191: 952–954.
10. Simons DJ, Chabris CF. Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception.* 1999; 28: 1059–1074.
11. Beauchamp JP, Benjamin DJ, Chabris CF, Laibson DI. How Malleable are Risk Preferences and Loss Aversion? *Harvard University Mimeo*. March 2012.
12. Wilmer JB, Chatterjee G, Chabris CF, Gerbasi ME, Germine L, Nakayama K, Duchaine, B. A brief test battery for developmental prosopagnosia. In press, *Cognitive Neuropsychology.*
13. Koenig LB, McGue M, Krueger RF, Bouchard TJ. Genetic and environmental influences on religiousness: Findings for retrospective and current religiousness ratings. *Journal of Personality.* 2005; 73: 471–488.
14. Olson JM, Vernon PA, Harris JA, Jang KL. The heritability of attitudes: A study of twins. *Journal of Personality and Social Psychology.* 2001; 80: 845–860.
15. Greene JD, Sommerville RB, Nystrom LE, Darley JM, Cohen JD. An fMRI investigation of emotional engagement in moral judgment. *Science.* 2001; 293: 2105–2108.
16. Marks DF. Visual imagery differences in the recall of pictures. *British Journal of Psychology.* 1973; 64: 17–24.
17. Jackson DN. *Multidimensional Aptitude Battery II*. 2nd ed. Port Huron, MI: Sigma Assessment Systems; 1998.
18. R Development Core Team. *R: A Language and Environment for Statistical Computing*.

Vienna, Austria: R Foundation for Statistical Computing; 2010.
19. Price AL, Patterson N, Plenge RM, Weinblatt ME, Shadick NA, et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics.* 2006; 38: 904–909.