The Genetics of General Cognitive Ability

A Dissertation
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Robert Miles Kirkpatrick

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Advisers: Matt McGue, Niels G. Waller

May 2013

Acknowledgements

The reported research was made possible by NIH grants AA09367, AA11886, DA05147, DA13240, DA024417, and MH066140. The author, Robert M. Kirkpatrick, was supported by a Doctoral Dissertation Fellowship from the University of Minnesota Graduate School.

The author extends his special thanks to Matt McGue, William G. Iacono, Niels G. Waller, Saonli Basu, Michael B. Miller, Nathan Pankratz, Joshua D. Isen, William S. Oetting, and Scott I. Vrieze, all of whom provided helpful input while he was working on this dissertation. Indeed, most of you will be co-authors on the eventual publications, and he uses "we" instead of "I" throughout this document to reflect that team effort.

Abstract

General cognitive ability (GCA) is a highly heritable trait, with correlates in numerous other domains. This dissertation reports the results of three studies of the genetics of GCA, conducted with participants from the Minnesota Center for Twin & Family Research (MCTFR). Study #1 (N = 7,100) is a genome-wide association study plus other analyses that exploit genome-wide single-nucleotide polymorphism (SNP) data. Study #2 (N = 6,439) is an association study of a different class of genetic polymorphism, the copy-number variant (CNV). In this study, we detect CNVs from genome-wide SNP-allele-probe intensity data. We aggregate them into genome-wide mutational burden scores and also carry out genome-wide association scans for specific CNVs. Study #3 is a biometric moderation study in a sample of 2,494 pairs of twins, full siblings, and adoptive siblings. We compared models by their sample-size-corrected AIC, and based our parametric inference on model-averaged point estimates and standard errors. Taken as a whole, these three studies demonstrate that GCA is substantially heritable and massively polygenic, but it is also influenced by environmental factors, and its heritability can be moderated by contextual variables like age and family-of-origin socioeconomic status (SES).

Table of Contents

List of Tables

List of Figures

Introduction

"Verbal definitions of the intelligence concept have never been adequate or commanded consensus…Development of more sophisticated factor analytic methods than Spearman or Thurstone had makes it clear that there is a g factor, that it is manifested in either omnibus IQ tests or elementary cognitive tasks, that it is strongly hereditary, and that its influence permeates all areas of competence in human life…A century of research—more than that if we start with Galton—has resulted in a triumph of scientific psychology, the footdraggers being either uninformed, deficient in quantitative reasoning, or impaired by political correctness."
--Paul E. Meehl (1998/2006, p.435)

In 1904, Charles Spearman proposed his "two-factor" theory of intelligence, which asserted two sources of variation in any mentally demanding task: a general factor common to all such tests, *g*, and a specific factor particular to each, *s*. His theory was incomplete in that it neglected "group factors," sources of the covariation among similar tests that remains after accounting for *g* (which he acknowledged in his 1927 book). However, the essence of the theory has been confirmed: there is a positive manifold evident among all cognitive abilities, and any dataset of reasonably diverse mental measurements admits a single general factor of intelligence (Carroll, 1993; Jensen, 1998; Johnson & Bouchard, 2005; Johnson, te Nijenhuis, & Bouchard, 2007), often accounting for about half of the observed variance. Thus, Boring (1923) was correct, in a more profound sense than he perhaps intended, when he famously remarked that "intelligence is what the tests test": the general factor represents that which is common to all tests tapping into what one might intuitively term "intelligence."

Verbal definitions of "intelligence" invariably leave something to be desired. Perhaps the best definition of the term is from Wechsler (1939, p. 3): "Intelligence is the aggregate or global capacity of the individual to act purposefully, to think rationally and to deal effectively with his environment." Among the better definitions is that offered in a letter to the *Wall Street Journal,* scribed by Linda Gottfredson (1994/1997a) and signed by 52 psychologists:

> Intelligence is a very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience…it reflects a broader and deeper capability for comprehending our surroundings.

The theory of human mental abilities could easily become a confused hodge-podge of competing definitions of "intelligence," which underscores the import of the general factor. "Which of the many verbal definitions of 'intelligence' is correct for guiding research?" Gottfredson (2003, p. 293) inquires. "With *g* as the common yardstick, the question becomes moot."

At this point, the ontological status of *g* deserves some consideration. The *g* factor, of course, is a mathematical entity: it is a latent variable invoked by a mathematical model, factor analysis, to account for the positive intercorrelation of mentally demanding tasks. In contrast, "general cognitive ability" (GCA) is a theoretical construct: an idea formulated by psychologists that serves as a basic building block in the assembly of theories of human cognitive abilities. It is GCA with which the present work is concerned. Obviously, neither one is a concrete, tangible object or unitary physical phenomenon; in this particular regard, objections concerning the "reification" of intelligence (e.g., Gould, 1996) are valid as far as they go. More pertinent is the subtle and somewhat technical disclaimer that, strictly speaking, *g* and GCA are conceptually distinct. To ignore this distinction is to "take metaphor for math" and confuse the mathematical *g* with the substantive interpretation attached to it (Maraun, 1996). Despite these finer points, though, most of the time it does no harm to informally identify GCA with the *g* factor, because their shared property of *generality* is of fundamental importance in contemporary theory and research of human mental abilities.

Slightly complicating matters is the fact that "IQ" is also often used as a synonym of "general cognitive ability" or "*g*." Strictly speaking, even if we accept the approximate identification of general ability with *g*, IQ is distinguishable from both. IQ is neither a latent mathematical variable nor a substantive theoretical construct; it is an observed, manifest score on a test designed to assess an individual's cognitive functioning for the purpose of clinical evaluation or educational intervention. Nonetheless, the consequences of this conceptual blurring are rather benign. Full-scale IQ (FSIQ) scores from individually administered tests are *g*-loaded and are themselves weighted composites of scores from the test's various constituent subscales.

Ackerman and Lohman (2003, p. 278) explain that "[m]ost standardized IQ tests provide a reasonable estimate of *g*…However, not every test that purports to measure *g* is an IQ test." For instance, Raven's Progressive Matrices typically exhibits high *g*-loadings, but because it is nonverbal in content, it correlates only moderately with the verbally-saturated IQ score.

Thus, full-scale IQ is but one way to operationalize GCA. A single *g*-loaded test, such as the Raven, is another way, albeit a suboptimal one: GCA represents that which is common to all cognitively demanding tasks, not just one. Another reasonable operationalization is the first principal component extracted from a sizable battery spanning a wide variety of specific cognitive abilities (possibly including those from an IQ test), because it will have been computed from a representative sample of the universe of mental ability tests. *g* factors calculated separately from different test batteries (taken by the same participants) range from highly correlated (Johnson et al., 2004) to nearly collinear (Johnson et al., 2008), suggesting that the nature of a *g* factor extracted from a battery of tests does not depend upon the exact composition thereof. Thus, the *g* factor indeed appears to be *general* with respect to content and context, a property dubbed by Spearman (1927, p. 197) the "indifference of the indicator."

GCA is associated with other behavioral variables in a striking variety of domains. Reviews by Jensen (1980, 1998) and Herrnstein and Murray (1994) of GCA's many correlates are without peer in their scholarly breadth and thoroughness. First, it should come as no surprise that IQ is strongly associated with academic achievement and educational attainment, for these are the most well-replicated and uncontroversial correlates of intelligence tests (Jensen, 1998, 1980). The correlation of IQ with achievement test scores is highest at the primary school level (around 0.6 to 0.7) and lowest at the post-graduate level (0.3 to 0.4), due to the range-restriction of ability at higher levels of education (Jensen, 1980). IQ shows correlations ranging 0.60 to 0.75 with grades in primary and secondary school, and with years of formal education completed among adults. As long as the achievement criterion is not prone to ceiling or floor effects, the association between IQ and achievement appears linear in form across the IQ distribution, suggesting that IQ is not a “threshold variable” for achievement (Jensen, 1980). Scores on the *g*-loaded SAT, used for admission to college, correlate around 0.50 with grades while in college (Sackett, Borneman, & Connelly, 2008). The functional form of this association appears linear across the distribution of SAT scores (Cullen, Hardison, & Sackett, 2004). Scores on the various *g*-loaded tests used for post-graduate admissions correlate around 0.4 to 0.6 with first-year graduate GPA and around 0.35 to 0.45 with cumulative graduate GPA, and predict several other career outcomes reasonably well (Kuncel, Hezlett, & Ones, 2004; Kuncel & Hezlett, 2007). One of the most robust findings in industrial-organizational psychology is that, as a rule, GCA is the single best predictor of job performance, and few cognitive predictors show meaningful incremental validity over it (Ree, Earles, & Teachout, 1994; Schmidt & Hunter, 1998). However, GCA’s predictive utility is moderated by job complexity (Gottfredson, 1997b, 2003), so that general ability is a stronger predictor for more-complex occupations. More surprisingly, predicted job performance appears to be a strictly increasing function of *g*, even at the upper tail of the ability distribution (Coward & Sackett, 1990).

Given its correlations with educational and work-related outcomes, it comes as no surprise that general ability correlates with occupational status and income (Herrnstein & Murray, 1994; Strenze, 2007). GCA is a psychological variable that demands the attention of the sociologist as well as the epidemiologist: it is also correlated with longevity (Gottfredson & Deary, 2004) and the capability of meeting the demands of daily life (Gottfredson, 1997, 2003). GCA even has small, negative correlations with criminality (Herrnstein & Murray, 1994; Ellis & Walsh, 2003). At a more reductionist level of analysis, GCA is associated with performance on some elementary cognitive tasks, such as choice reaction time and visual inspection time (Jensen, 1998, chapter 8). Further, certain neurobiological variables are moderately correlated with GCA, such as brain size (Wickett, Vernon, & Lee, 2000; McDaniel 2005; Rushton & Ankney, 2009) and white-matter integrity (Gläscher et al., 2010).

How can individual variation in GCA be explained? There is abundant evidence, accumulated over several decades (or more than a century, if we go back to Galton, 1869) that a major portion of this variation is genetic in origin. As Meehl (1998/2006) states, GCA is a "strongly hereditary" trait: twin, family, and adoption studies typically estimate that 50% to 70% of its variance is heritable (Bouchard & McGue, 1981, 2003; Deary, Spinath, & Bates, 2006). More recently, as science has entered the genomic era, researchers have attempted to identify specific genetic polymorphisms that can account for phenotypic variation in GCA.

This dissertation reports three studies concerning the genetics of general cognitive ability. All three studies used participants in longitudinal family studies conducted by the Minnesota Center for Twin & Family Research who were assessed with an abbreviated form of age-appropriate Wechsler IQ test. The exact samples and measurements were not identical for all three studies, so details concerning same are provided in the appropriate section of each. The first two were molecular-genetic, each involving a different kind of polymorphism, whereas the third was quantitative-genetic, utilizing biometric analysis-of-variance. Study #1 investigated how well common single-nucleotide polymorphisms (SNPs) from across the genome can explain variation in GCA, by conducting a genome-wide association study (GWAS) along with related (and more informative) analyses. Study #2 involved a different class of polymorphism, copy-number variants (CNVs), which have been implicated in neurodevelopmental disease but have not been studied in connection with quantitative traits until relatively recently. Study #3 attempted to answer questions about how magnitudes of genetic and environmental influence on GCA can themselves be influenced by other variables, referred to as "biometric moderators." It focused primarily on two such biometric moderators. One of these, age, is well-documented in the existing literature. The other, family-of-origin socioeconomic status, is at once more interesting and more tentative. This dissertation closes with a brief review of the major conclusions of the three studies.

Study #1

Background

Linkage & Candidate-Gene Association

Decades of research from twin, family, and adoption studies have established that general cognitive ability (GCA) is a substantially heritable trait. Estimates of its heritability (h^2), the proportion of its variance that is attributable to genetic factors, typically range from 0.50 to 0.70 (Bouchard & McGue, 1981, 2003; Deary, Spinath, & Bates, 2006), and are sometimes as high as ~0.80 (Rijsdijk, Vernon, & Boomsma, 2002). In light of the empirical fact that genes influence cognitive ability, a natural subsequent question to ask is *which* genetic polymorphisms contribute to individual variation in the trait.

One strategy for hunting trait-relevant polymorphisms is linkage analysis. Linkage was first discovered in *Drosophila* by T.H. Morgan (1911) as an exception to Mendel's law of independent assortment. As originally conceptualized, the phenomenon of linkage is that certain Mendelian traits co-occur in a given pedigree, because the loci for those traits are in relatively close proximity on the same chromosome. Newton Morton (1955) developed the first statistical test for human-genetic linkage. Nowadays, contemporary methods generalize linkage analysis to complex quantitative traits in small pedigrees; thorough treatment of the underlying mathematics may be found in texts such as Ott (1999). Linkage analysis implicates a region of a chromosome, by providing evidence that some typed marker in the region co-segregates with a variant causal to the quantitative trait. This evidence arises when relatives' counts of identical-by-descent alleles for that marker correspond to their similarity on the trait, to a sufficiently greater degree than would be expected by chance.

The literature on linkage studies of cognitive abilities has been nicely summarized by Deary, Johnson, & Houlihan (2009). Three linkage studies of GCA—Posthuma et al. (2005), Luciano et al. (2006), and Dick et al. (2006)—provided converging evidence of linkage signals for broad regions on chromosomes 2q and 6p. Again, linkage analysis only implicates a chromosomal region; it is not specific enough to implicate a polymorphism or even a gene. Finer-grained analyses are needed to identify the genes and mutations within genes that underlie individual variation in GCA and other complex traits. Further, linkage analysis is severely limited by its low power to detect small effects. As Risch & Merikangas (1996, p. 1516) stated in an influential article,

> [T]he method that has been used successfully (linkage analysis) to find major genes has limited power to detect genes of modest effect, but…a different approach (association studies) that utilizes candidate genes has far greater power.

Association analysis is merely a test for whether the allelic state of a genetic polymorphism systematically covaries with the disease or quantitative trait of interest (typically via regression analysis). Unlike linkage, it can implicate a specific polymorphism, but the "causal" polymorphism must actually be typed, or alternately, it must lie in close chromosomal proximity—linkage disequilibrium[1] (LD) to a marker that is typed. Therefore, association analysis requires denser genotyping than linkage does. Consequently, for a number of years, association analysis saw use primarily in candidate-gene studies.

The rationale behind the candidate-gene study is simple. Allelic association requires denser genotyping of markers than linkage analysis does. So, typing markers within genes that are *a priori* plausibly related to the phenotype is a focused use of limited genotyping resources, which is (presumably) more likely to identify genetic variants that are truly associated with the phenotype. Unfortunately, the candidate-gene association literature has been plagued by apparent

[1] Linkage disequilibrium is the logical consequence of the mechanisms of linkage, applied over many generations to entire populations. The result is that loci very close to one another on a chromosome are least likely to be sundered by a recombination event, and therefore, polymorphisms within small "blocks" of DNA on a given chromosome tend to be transmitted together in the population. This essentially induces correlation between markers in tight proximity to one another on the same chromosome.

false positives that fail to replicate. This has occurred in human genetics at large (Hirschorn, Lohmueller, Byrne, & Hirschorn, 2002; Ioannidis, Ntzani, Trikalinos, & Contopoulos-Ioannidis, 2001), and has occurred in candidate-gene association research for GCA since its inception (Payton, 2006). In fact, one recent article concluded that "most reported genetic associations with general intelligence are probably false positives" (Chabris et al., 2012).

Rather presciently, Risch & Merikangas (1996) foreshadowed the advent of the genome-wide association (GWAS) study in their remark that an "approach (association studies) that utilizes candidate genes has far greater power, *even if one has to test every gene in the genome*" (p. 1516, emphasis supplied). The genome-wide association scan (GWAS) grew naturally out of researchers' (1) demand for denser and denser coverage of variation in more and more genes, and (2) growing dissatisfaction with replication failures in association studies of *a priori* biologically-hypothesized candidate genes. GWAS in the modern sense involves typing individuals on at least 300,000 SNPs throughout the genome (Balding, 2006); due to LD, SNPs that are typed can "speak on behalf" of non-genotyped SNPs and other polymorphisms that are nearby on the same chromosome. It is only within the past five years or so that biotechnology reached such sophistication that researchers can feasibly genotype a sample of participants on hundreds of thousands of SNPs, and engage in the atheoretical brute-force empiricism that is GWAS. Needless to say, there is an inherent multiple-testing problem in GWAS; the currently accepted standard for "genome-wide significance" is $p < 5 \times 10^{-8}$.

GWAS

In a sense, the IQ QTL Project (Plomin, 2003) carried out the first "genome-wide association study" of GCA (via DNA pooling; Daniels et al., 1998), with only 1,847 markers; it failed to uncover replicable association. A "low-density GWAS" for IQ has been reported by Pan, Wang, & Aragam (2011). Their discovery sample consisted of 1,335 Caucasian participants, in 292 nuclear families from the Collaborative Study on the Genetics of Alcoholism (COGA), who had been tested with the Wechsler Adult Intelligence Scale – Revised (WAIS-R). Their replication sample contained 614 children, in eight different countries, from the International Multi-Center ADHD Genetics Project (IMAGE). Pan et al. first conducted family-based association tests in their COGA discovery sample. No SNP reached genome-wide significance, but 22 SNPs had $p < 10^{-3}$, regarded as "suggestive association." These SNPs, and all other SNPs in their genic regions, were tested for association in the IMAGE sample. Again, there were no genome-wide significant hits, but suggestive signals were again observed in two different genes, *NTM* and *NR3C2*.

The first two "true" GWAS for GCA both used samples of children from the Twins Early Development Study (TEDS). Butcher, Davis, Craig, and Plomin (2008) reported the first, which used DNA pooling in a two-stage design. Subsequently, Davis et al. (2010) ran a similar study, which used DNA pooling in a three-stage design. The full samples of both studies comprised over 7,000 healthy 7-year-old Caucasian twins born in England and Wales, who took a battery of four ability tests via telephone interview, from which a composite *g* score was computed.

Prior to genotyping individual DNA samples, stage 1 of both studies entailed pooling DNA from small groups of participants at the upper and lower extreme of the ability distribution, and typing the pools on over 300,000 SNPs. Then, the SNPs most strongly differentiating the high- and low-ability pools were selected to be individually genotyped and tested for association in the rest of the full sample at a later stage. The two studies differed in that Davis et al. (2010) included a second stage of DNA pooling with extreme-score probands, as an attempt to replicate the top 3,000 SNPs from the first stage. Davis et al. (2010) then genotyped the top 28 SNPs from the second stage in 3,297 participants (one per family) from the full sample. On the other hand, Butcher et al. (2008) genotyped the top 37 SNPs from only one round of DNA pooling in 3,195 participants (one per family) from the full sample; 2 of these 37 SNPs overlapped with those that Davis et al. selected for individual genotyping. Butcher et al. reported that, at the uncorrected $\alpha = 0.05$, their stage-2 association analysis would have 100%, 98%, and 71% power to detect an additive SNP accounting for 1%, 0.5%, and 0.2%, respectively, of the phenotypic variance. Davis et al. reported that their stage-1 DNA-pooling analysis would have 80% power to detect an additive SNP accounting for 1.7% of the phenotypic variance, at a per-comparison $\alpha = 5 \times 10^{-7}$. With a Bonferroni correction for 28 hypothesis tests yielding a per-comparison $\alpha = 0.001786$, Davis et al.'s stage-3 association analysis would have 99.5% and 82% power to detect a SNP accounting for 1% and 0.5%, respectively, of the phenotypic variance.

Butcher et al. (2008) observed nominally significant association from 6 of 37 SNPs entered into the full-sample association analysis. After implementing Benjamini and Hochberg's (2005) step-up procedure to control false discovery rate, only one of these SNPs, rs1378810, was resolved as a discovery ($r^2 = 0.004$, corrected $p < 0.03$). Of Davis et al.'s (2010) 28 SNPs entered into the full-sample association analysis, 9 were nominally significant, but none survived Bonferroni correction or Benjamini and Hochberg's procedure.

The largest effect-size estimate that Davis et al. reported is $r^2 = 0.0024$. The largest effect-size estimate that Butcher et al. (2008) reported is $r^2 = 0.004$; the sum of effect sizes of their six nominally significant SNPs was only 1.2% of the variance. Butcher et al commented accordingly (p. 442, emphasis in original), and succinctly summarized the main lesson of GWAS for quantitative traits:

> One possible reason for not observing larger, common, single-locus SNP effects for *g* is that they do not exist…[I]t may be that for…quantitative traits…the main finding is the *exclusion* of SNPs of large effect size to the extent that coverage for common variants is virtually complete…[W]innowing the wheat from the chaff will be difficult, requiring extremely large samples, multiple-stage designs, and replication in independent samples.

As others have pointed out, the same lesson is apparent from GWAS for human height (see Visscher, 2008, and Turkheimer, 2011). Height is highly heritable, uncontroversial in definition, and easily measured, almost without error. And yet, the SNPs for height that initial GWAS have identified each accounted for around 0.3% or less of the phenotypic variance, and in total, 3% (though see Lango Allen et al., 2010). It would appear that variation in quantitative traits is attributable to a very large number of polymorphisms of very small effect.

Clearly, it is necessary to move beyond analyses of one SNP at a time. We refer to GWAS, combined with analyses that aggregate across multiple SNPs in some fashion, as "GWAS plus." We describe three such multi-SNP analyses: *VEGAS*, polygenic scoring, and *GCTA*.

GWAS Plus: Polygenic Scoring

Both TEDS GWAS (Butcher et al., 2008; Davis et al., 2010) illustrated a simple approach to combining the effect of multiple SNPs: for each participant, aggregate those alleles suggestively implicated in the GWAS into a "genetic score" for him/her. From the six nominally significant SNPs from the GWAS, Butcher et al. simply counted how many of the putative increaser alleles each participant carried. This score ranged from 1 to 11 in the subsample of 2,676 children in which it was calculated, and correlated $r \approx 0.10$ with general ability—a very significant result ($p < 3 \times 10^{-8}$). Similarly, Davis et al (2010) created a score from the nine nominally significant SNPs from the GWAS, which ranged from 6 to 16, and accounted for 1.2% of phenotypic variance. Davis et al. acknowledge that they conducted the genetic scoring analysis with the same participants in which they conducted the GWAS, so the analysis is almost certainly capitalizing on chance.

Shaun Purcell (with the International Schizophrenia Consortium, 2009) was perhaps the first to perform genetic scoring by weighting each selected SNP by its GWAS regression coefficient, and cross-validating in a separate sample. Not surprisingly, the genetic score's predictive performance upon cross-validation depended upon the GWAS p-value threshold set for a SNP to be included toward the score (International Schizophrenia Consortium, 2009, supplemental online material); at best, the genetic score could predict around 3% of the disease risk in the cross-validation sample.

Lango Allen et al. (on behalf of the GIANT Consortium, 2010) utilized genetic scoring subsequent to a GWAS for human height on a titanic scale: a combined sample of 133,653 participants, with called or imputed genotypes on over 2.8 million SNPs, and a replication sample of 50,074 participants. The GIANT Consortium ultimately identified 180 SNPs robustly associated with height. The genetic score from these loci predicted around 10% of the phenotypic variance in each cross-validation sample. When additional SNPs at varying significance thresholds were counted toward the score, it predicted as much as 16.8% of the variance in a cross-validation sample.

GWAS plus: *VEGAS*

VEGAS (Versatile Gene-based Association Study; Liu et al., 2010) is a program that tests each *gene* (specifically, all genotyped SNPs in each gene) for association with the phenotype, via parametric bootstrapping. A rather clever program, it takes GWAS results as its input, requiring only the rs numbers and GWAS p-values of each SNP. If an Internet connection is available, the program "knows" which of 17,787 autosomal gene(s), if any, contain each SNP. Within each gene, the program first converts each SNP p-value to the corresponding quantile from a central chi-square distribution on 1*df*, and sums them to produce an observed test statistic T_{obs} for that gene. The null hypothesis is that there is no association of any SNP in the gene with the phenotype. Under the null, and if there were zero LD among the gene's *m* SNPs, then $T \sim \text{chi}^2(m)$. Under the null, but at the other extreme of perfect LD among the *m* SNPs, then $T/m \sim \text{chi}^2(1)$.

But, *VEGAS* also “knows” the LD structure from reference datasets for three populations: HapMap CEU (Caucasians of European ancestry), CHB and JPT (Han Chinese and Japanese), and YRI (West Africans). The matrix of pairwise LD correlations for the user-specified population, $\mathbf{\Sigma}$, is employed in the random generation of test statistics under the null hypothesis. Specifically, in each iteration, an order-m vector is drawn from a multivariate normal distribution with zero mean and covariance matrix equal to $\mathbf{\Sigma}$. The elements of this vector are squared and summed, yielding the value of the test statistic for that iteration. The proportion of test statistics exceeding T_{obs} provides the p-value for the gene-based test of association. Liu et al. (2010) recommend a Bonferroni-corrected significance level of $p < 0.05$ / 17,787, or 2.8×10^{-6}, which is slightly conservative since genes’ boundaries overlap to some extent.

GWAS Plus: *GCTA*

GCTA (Genome-wide Complex Trait Analysis; Yang, Lee, Goddard, & Visscher, 2011) is a software package that implements what some (e.g., Benjamin et al., 2013) have referred to as GREML, for “genomic-relatedness restricted maximum-likelihood.” Instead of regressing a quantitative trait onto one marker at a time, *GCTA* instead assesses how much of the phenotypic variance is attributable to *all* the typed markers at once, which is accomplished by treating all the markers as random effects, and entering them into a mixed linear model fit by restricted maximum likelihood. *GCTA* thereby provides an unbiased estimate of the variance attributable to the typed SNPs, and a matrix of (roughly) genome-wide SNP correlations among participants—a genetic relationship matrix, obtainable from a genotyped sample of classically unrelated participants. Put simply, *GCTA* attempts to predict phenotypic similarity among individuals from their genetic similarity, and to predict phenotypic variance that would otherwise be treated as error. *GCTA* may be expected to outperform polygenic scoring, because it does not rely upon estimates of individual SNP effects, which are prone to sampling error (Visscher, Yang, &

Goddard, 2010).

For *n* participants typed on *m* SNPs, the *GCTA* model (Yang, Lee, et al., 2011) is expressed as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon} \tag{1.1}$$

where $\mathbf{y}$ is a random $n \times 1$ vector of scores on a quantitative trait, $\mathbf{X}$ is a model matrix of scores on covariates, $\boldsymbol{\beta}$ is a vector of the covariates' regression coefficients (fixed effects), and residual vector $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I}\sigma_\varepsilon^2)$. Further, $\mathbf{u}$ is an $m \times 1$ vector of random SNP effects, such that $\mathbf{u} \sim \mathcal{N}_m(\mathbf{0}, \mathbf{I}\sigma_u^2)$; $\mathbf{W}$ is an $n \times m$ matrix of participants' reference-allele counts, expressed as *z*-scores (i.e., columns are standardized).

We hereby condition upon the observed value of $\mathbf{X}$. Since the random effects have zero expectation, $E(\mathbf{y} \mid \mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$. Now define the phenotypic variance matrix, $\mathbf{V}$:

$$\mathbf{V} = var(\mathbf{y} \mid \mathbf{X}) = \mathbf{W}\mathbf{W}^T\sigma_u^2 + \mathbf{I}\sigma_\varepsilon^2 \tag{1.2}$$

Further define genetic relationship matrix $\mathbf{A} = \frac{1}{m}\mathbf{W}\mathbf{W}^T$. Matrix $\mathbf{A}$ is $n \times n$, and roughly, may be regarded as a matrix of correlations between different participants' genotypes. However, this is not strictly correct, since $\mathbf{W}$ is standardized by column (SNP) rather than by row (participant), and therefore the elements of $\mathbf{A}$ are not bounded between -1 and 1.

Let $\sigma_g^2 = m\sigma_u^2$, the variance attributable to all SNPs. The model may now be written:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{g} + \boldsymbol{\varepsilon} \tag{1.3}$$

where $\mathbf{g}$ is an $n \times 1$ vector of random genetic effects, distributed as $\mathcal{N}(\mathbf{0}, \mathbf{A}\sigma_g^2)$. Now,

$$\mathbf{V} = var(\mathbf{y} \mid \mathbf{X}) = \mathbf{A}\sigma_g^2 + \mathbf{I}\sigma_\varepsilon^2 \tag{1.4}$$

where σ_g^2 is the component of variance attributable to all typed SNPs and all untyped "causal" mutations in close LD with them. Herein, we refer to the ratio of σ_g^2 to phenotypic variance as h_{SNP}^2, for it is a lower-bound estimate of the additive heritability of the phenotype. Estimation is carried out via restricted maximum-likelihood; details of the algorithm may be found in Yang, Lee, et al. (2011).

Recent Developments

Davies et al. (2011) reported a "GWAS Plus" for cognitive abilities. The discovery sample contained 3,511 unrelated participants, combined from 5 cohorts of older adults in the United Kingdom, who were genotyped on 549,692 SNPs across the genome. The replication cohort comprised 670 Norwegian participants of a wide range of ages (18-79). For reasons not explained, Davies et al. extracted composite scores for both crystallized and fluid ability from the ability measures in each cohort, rather than a single principal component, and conducted separate analyses for fluid and crystallized ability. Possibly, they were concerned about the different developmental trajectories for fluid and crystallized ability in late adulthood (Schaie, 1994; Salthouse, 2004). Arguably, therefore, Davies et al. does not constitute a study of *general* cognitive ability.

Davies et al. combined association results from the 5 UK cohorts via meta-analytic techniques. No single SNP achieved genome-wide significance ($p < 5 \times 10^{-8}$). Gene-based tests in *VEGAS* implicated only one gene, *FNBP1L*, which was not confirmed in the replication cohort. Davies et al. performed polygenic scoring using the most lenient SNP inclusion threshold possible: *all* genotyped SNPs, irrespective of GWAS *p*-value. In the UK samples, this score predicted between 0.45% and 2.19% of the variance. Under cross-validation in the replication cohort, this score predicted less than 1% of the variance (statistically significant for both fluid and crystallized ability). Davies et al. emphasized that, when treating SNPs as single fixed effects, their individual effect sizes will be quite small, and estimated with considerable sampling error.

Instead, *GCTA*, though it is silent with regard to the individual contribution of each marker, treats all SNPs as random effects and estimates a single omnibus variance component. This seems to be one of its major advantages. In any event, the truly impressive results from Davies et al.

(2011) were from *GCTA*, which produced variance-component estimates equivalent to 40% of the variance in crystallized ability, and 51% of the variance in fluid ability. Davies et al. (p. 1) conclude that “human intelligence is highly heritable and polygenic.”

A recent study of GCA in children and adolescents reported by the Childhood Intelligence Consortium (CHIC; Benyamin et al., 2013) has borne out that same conclusion. The CHIC study represented a collaboration of six discovery cohorts (total N = 12,441) and three replication cohorts (N = 5,548). One of the replication cohorts was a sample of Caucasian adolescent participants from studies conducted at the Minnesota Center for Twin &Family Research (MCTFR, N = 3,367), which is a subset of the present study’s sample. The phenotype in all cohorts was either Full-Scale IQ score or a composite score derived from a battery of both verbal and non-verbal tests. GWAS SNP results were combined across discovery cohorts by meta-analysis. No SNP reached genome-wide significance. Among the top 100 SNPs from the discovery GWAS, none was significant after Bonferroni correction in any of the replication cohorts, though discovery sample’s estimated regression coefficients for these 100 SNPs were moderately positively correlated with those from two of the three replication cohorts, but not the MCTFR cohort.

Gene-based analysis with *VEGAS* in Benyamin et al.’s (2013) discovery sample suggested association with *FNBP1L* (*formin binding protein 1-like*, on chromosome 1; $p = 4 \times 10^{-5}$), which “is involved in a pathway that links cell surface signals to the actin cytoskeleton” (p. 3). This was also the most significantly associated gene in Davies et al.’s (2011) discovery cohort. However, one cohort was common to both studies—Davies et al. used adult IQ scores from the Lothian Birth Cohorts, whereas Benyamin et al. used their childhood IQ scores. When Benyamin et al. combined *VEGAS* results across all of their cohorts except the Lothian Birth Cohorts, the association with *FNBP1L* remained nominally significant (p = 0.0137), as did the top SNP in the gene ($p = 4.5 \times 10^{-5}$). Benyamin et al. regarded this as robust evidence of association between GCA and polymorphisms in *FNBP1L*.

Benyamin et al. (2013) also reported results of polygenic scoring analyses conducted in the replication cohorts. These analyses calculated polygenic scores from the SNP regression weights obtained in the meta-analytic GWAS results from the discovery sample. Eight such analyses were conducted in each replication cohort, with a different p-value cutoff for each. That is, polygenic score for each such analysis was computed from a set of SNPs the p-values of which exceeded some threshold in the discovery sample. The proportion of variance attributable to the polygenic score varied by p-value cutoff and by replication cohort, but was statistically significant for at least one analysis in each replication cohort. The best achieved in the MCTFR cohort was 0.5% of variance ($p = 5.52 \times 10^{-5}$). Finally, Benyamin et al. reported *GCTA* results for the three largest cohorts in the study, one of which was the MCTFR cohort. Estimates of h^2_{SNP} varied from 0.22 to 0.46, with the MCTFR estimate in between at 0.40; all three estimates were significantly different from zero. Based on all results, Benyamin et al. conclude that "[c]hildhood intelligence is heritable, highly polygenic and associated with *FNBP1L*" (p. 1).

In the present study, we report the detailed results of our "GWAS Plus" from our full sample of 7,100 Caucasian MCTFR participants, both adolescents and adults. We conducted our GWAS using over 2.6 million SNPs and a method appropriate for the complicated family structures in our dataset. We then conducted gene-based association tests in *VEGAS* with the SNP p-values calculated in our GWAS. We also carried out polygenic scoring analyses with five-fold cross-validation. Finally, we ran *GCTA* at different thresholds of biological relatedness to estimate how much of the phenotypic variance is attributable to all genotyped SNPs.

Methods

Sample

Participants.

Our participants came from two longitudinal family studies conducted by the MCTFR. The Minnesota Twin Family Study (MTFS; Iacono, Carlson, Taylor, Elkins, & McGue, 1999; Iacono & McGue, 2002; Keyes, Malone, Elkins, Legrand, McGue, & Iacono, 2009) is a longitudinal study of same-sex twins, born in the State of Minnesota between 1972 and 1994, and their parents. There are two age cohorts in this community-based sample, an 11-year-old cohort (10-13 years old at intake, mean age = 11.78) and a 17-year-old cohort (16-18 years old at intake, mean age = 17.48). Zygosity has been genomically confirmed for all twins included in the present study (Miller et al., 2012). The Sibling Interaction & Behavior Study (SIBS; McGue, Keyes, Sharma, Elkins, Legrand, et al., 2007) is a longitudinal adoption study of sibling pairs and their parents. This community-based sample includes families where both siblings are adopted, where both are biologically related to the parents, or where one is adopted and one is biologically related. As required by SIBS inclusion criteria, any sibling in the sample who was adopted into the family will not be biologically related to his or her co-sibling, which has been genomically verified for all SIBS participants in the present study (Miller et al., 2012). The age range at intake was 10-19 for the younger sibling, and 12-20 for the older. Written informed consent or assent was obtained from all participants, with parents providing written consent for their minor children. For the purposes of our analyses, the sample comprises six distinct family types:

1. Monozygotic- (MZ) twin families (*N* = 3,939 in 1143 families),
2. Digyzotic- (DZ) twin families (*N* = 2,114, in 638 families),
3. SIBS families with two adopted offspring (*N* = 291, in 224 families),
4. SIBS families with two biological offspring (*N* = 472, in 184 families),
5. "Mixed" SIBS families with 1 biological and 1 adopted offspring (*N* = 204, in 107 families),
6. Step-parents (*N* = 80).

As explained below, our method of analysis accounted for the clustering of individual participants within families. However, family-type #6, step-parents, do not fit neatly into a four-member family unit; we treated them as independent observations (in a sense, as one-person families) in our analysis. A total of $N = 7{,}100$ participants were included in our analyses. Descriptive characteristics of the sample are provided in Table 1-1.

Genotyping.

Participants who provided DNA samples were typed on a genome-wide set of markers with the Illumina Human660W-Quad array. Both DNA samples and markers were subject to thorough quality-control screens. 527,829 SNPs on the array were successfully called and passed all QC filters, which filters include call rate <99%, minor allele frequency <1%, and Hardy-Weinberg equilibrium p-value $< 10^{-7}$. After excluding DNA samples that failed quality-control screening, a GWAS sample of 8,405 participants was identified.

Population stratification occurs when one's sample of participants represents heterogeneous populations across which allele frequencies differ appreciably, and can produce spurious genetic association (or suppress genuine association). We therefore restricted our analyses only to participants who are Caucasian, of European ancestry ("White"), based upon both self-reported ancestry as well as principal components from EIGENSTRAT (Price et al., 2006). These principal components were extracted from an $n \times n$ covariance matrix of individuals' genotypes across SNPs (similar to matrix $\mathbf{A}$ described above). A White GWAS sample of 7,702 participants was identified. The sample for the present study is the 7,100 out of 7,702 White participants with available phenotype data. Details concerning genotyping, quality-control, and ancestry determination can be found in Miller et al. (2012).

Imputation.

Many known SNPs exist that are not on our Illumina array. But, by combining observed

SNP genotypes with what is known—*a priori,* from reference data—about haplotype frequencies in the population, the allelic state of common untyped SNPs can often be imputed with a high degree of accuracy. For SNP imputation, using HapMap2 (International HapMap Consortium, 2007) as the reference panel, we first phased our observed genotypes into expected haplotypes with *BEAGLE* (Browning & Browning, 2009), which takes information from genotyped relatives into account to improve phasing. We then input phased data into *Minimac*, a version of *MaCH* (Li, Willer, Ding, Scheet, & Abecasis, 2010), to impute SNP states for a total of 2,094,911 SNPs not on the Illumina array. We used the allelic dosages of these SNPs in our GWAS, which are individuals' posterior expected reference-allele counts on each imputed SNP. The quality of the imputation for an untyped SNP may be assessed by its imputation R^2 (Li et al., 2010), which is the ratio of the variance of its imputed dosages to its population variance (from reference data). Our GWAS only included dosages of imputed SNPs with imputation $R^2 > 0.5$, of which there were 2,018,818. Between these imputed SNPs and the 527,829 from the array, we analyzed a total of 2,546,647 SNPs in our GWAS.

Phenotypic measurement.

Measurement of GCA was included in the design of the intake assessment for most participants, by way of an abbreviated form of the Wechsler Intelligence Scale for Children-Revised (WISC-R) or Wechsler Adult Intelligence Scale-Revised (WAIS-R), as age-appropriate (that is, 16 or younger, and older than 16, respectively). The short forms consisted of two Performance subtests (Block Design and Picture Arrangement) and Verbal subtests (Information and Vocabulary), the scaled scores on which were prorated to determine Full-Scale IQ (FSIQ). FSIQ estimates from this short form have been shown to correlate 0.94 with FSIQ from the complete test (Sattler, 1974). Parents in the SIBS sample were an exception, in that they were not tested with this short form of WAIS-R until the first SIBS follow-up assessment. By design, only one parent per SIBS family returned for this follow-up, which was usually the mother. As a result, IQ data for SIBS fathers is very limited in its availability.

IQ-testing was also included in the design of the second follow-up for both age cohorts of MTFS twins, and for the fourth follow-up for the 11-year-old cohort. At these assessments, twins received a further abbreviated form of WAIS-R, consisting only of the Vocabulary and Block Design subtests, the scaled scores on which were again prorated to determine FSIQ. Of the 3,226 twins entered into our analysis, 903 were tested twice, and 337 were tested three times. Multiple testing occasions were spaced approximately seven years apart. To achieve a more reliable assessment of the phenotype, we simply averaged all available measures of FSIQ for each participant, and used these single within-person averages in analysis. FSIQ among participants entered into analysis ranged from 59 to 151 (also see Table 1-1). Twelve participants with FSIQ of 70 or below were included in analyses. Despite their low scores, these participants were not noticeably impaired and were capable of completing the multifaceted MTFS/SIBS assessment during their visit. They are therefore unlikely to meet diagnostic criteria for mental retardation (American Psychiatric Association, 1994), and instead, merely represent the low end of the normal-range distribution of GCA.

Analyses

Statistical power.

Because our participants are clustered within families, our effective sample size is less than 7,100. We conducted two sets of power calculations in *Quanto* (Gauderman & Morrison, 2006), one that assumed 7,000 independent participants (an aggressive estimate of our effective sample size) and one that assumed 2,000 independent participants (a conservative estimate of our sample size). Both assume a Type I error rate of $\alpha = 5 \times 10^{-8}$, i.e. genome-wide significance. With 7,000 independent participants, our GWAS would have at least 80% power to detect a SNP accounting for 0.6% of phenotypic variance. With 2,000 independent participants, our GWAS would have at least 80% power to detect a SNP accounting for 2% of phenotypic variance.

GWAS.

Our GWAS consisted of a large number of least-squares regressions of FSIQ onto the genotype (or imputed dosage) of each SNP, along with covariates, which were sex, birth year, and the first 10 principal components from Price et al.'s (2006) EIGENSTRAT, to control for any crypto-stratification (i.e., lurking population stratification in a sample of apparently homogeneous ancestry) within this White sample. One notable example of this kind of stratification was reported by Campbell et al. (2005), in which a SNP in the gene for lactase (*LCT*) was significantly, but spuriously, associated with height among European-Americans. Allele frequency for the SNP in question is known to vary among regions of Europe, and no association was observed when participants were matched on grandparental country-of-origin. Instead, the SNP appeared to mark participants' ancestral origins along a northwest-southeast axis running through the continent of Europe.

Because our participants are clustered within families, they were not sampled independently. To further complicate matters, the within-family covariance structure will depend upon the kind of family in question. We therefore employed a feasible generalized least-squares (FGLS)[2] method in our GWAS, via *RFGLS*, a package for the R statistical computing environment designed for FGLS regression in datasets with complicated family structures (Li, Basu, Miller, Iacono, & McGue, 2011).

RFGLS has a "rapid-FGLS" approximation, which we used to run the GWAS and which

[2] As is widely known (see Li, Basu, Miller, Iacono, & McGue, 2011), in multiple regression, when the residuals are uncorrelated and have mean zero and constant variance, the best linear unbiased estimate of the regression parameters is obtained as $\widehat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$; if the residuals are further normally distributed and stochastically independent, $\widehat{\boldsymbol{\beta}}_{OLS}$ is also the maximum-likelihood estimator. If the residuals are not uncorrelated, $\widehat{\boldsymbol{\beta}}_{OLS}$ will not be maximally efficient, and the degrees-of-freedom for its test statistics will be mis-specified. In practice, the (non-diagonal) residual covariance matrix must be estimated from data. If $\mathbf{V}$ is a consistent such estimator, then the feasible generalized least-squares estimator is obtained as $\widehat{\boldsymbol{\beta}}_{FGLS} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}$.

works as follows. First, an FGLS regression of the phenotype onto covariates only is run. Then, the residual covariance matrix from this single regression is saved to disk, so it can then be "plugged in" for use in all subsequent single-SNP regressions, with covariates. The approximation saves a considerable amount of computation time, since the residual covariance matrix is calculated only once. According to Li et al. (2011), it produces negligible bias in the resulting *p*-values, so long as no SNP accounts for more than 1% of phenotypic variance (which is a very reasonable assumption). Estimates of the fixed and random effects from the covariates-only FGLS regression are presented in Table 1-2. *P*-values from the GWAS are depicted in Figures 1-1 through 1-4.

GWAS Plus: *VEGAS*.

We conducted gene-based association tests in *VEGAS*, inputting the 515,385 autosomal SNPs on the Illumina array, and specifying HapMap CEU as the reference data for pairwise LD correlations. We also ran *VEGAS* inputting all 2,485,152 autosomal SNPs, both observed and imputed, in the GWAS. However, adding the imputed SNPs made no substantive difference, so we report results only for the 515,385 observed SNPs. The gene-based *p*-values are depicted in Figures 1-5 and 1-6.

GWAS Plus: polygenic scoring.

We conducted polygenic scoring with five-fold cross-validation. Since the family is the independent unit of observation in our dataset, we first randomly divided the sample into five subsamples of approximately equal numbers of families, and with each family type approximately equally represented in each[3]. Then, we ran a GWAS with the observed SNPs five times over, each time including four of the five subsamples—the calibration sample for that

[3] Special care had to be taken to ensure that at least one adoptive family with a father was in each subsample, without which RFGLS would be unable to estimate the residual adoptive father-offspring correlation it needs to form the residual covariance matrix. Such families were rare in this dataset partly because, as mentioned previously, only one parent per SIBS family was IQ-tested, and it was usually the mother. Also, most SIBS adoptees are of Asian ancestry, and were not included in this study to begin with.

iteration. Then, the left-out subsample served as that iteration's validation sample.

Each iteration, we used *PLINK* (Purcell, 2007) to produce polygenic scores for the participants in the validation sample based on the GWAS statistics from the calibration sample, at the same eight p-value cutoffs used by Benyamin et al. (2013): $p \leq 0.001$, $p \leq 0.005$, $p \leq 0.01$, $p \leq 0.05$, $p \leq 0.1$, $p \leq 0.25$, $p \leq 0.5$, and $p \leq 1$(i.e., all SNPs). We used two different weighting methods to calculate polygenic scores. The first simply used the GWAS regression coefficients from the calibration sample. The second weighted each SNP as either -1 or 1, depending on the sign of its coefficient. Thus, with eight p-value cutoffs and two weighting schemes, we produced 16 polygenic-score vectors in each validation sample.

To evaluate the performance of the polygenic scores under cross-validation in each iteration, we first ran a FGLS regression of the phenotype onto covariates only in the validation sample, and retained the residualized FSIQs. We then did another FGLS regression of the residuallized FSIQ onto polygenic score. We calculated the coefficient of determination as Buse's[4] (1973) R^2, which is the coefficient of determination from OLS regression, except that each sum is instead replaced by a quadratic or bilinear form in the vector of terms, with a weight-matrix coefficient (for our purposes, this weight matrix is the inverse of the residual covariance matrix obtained from regressing the residuallized phenotype onto the score). The Buse's R^2s are plotted by subsample, weighting scheme, and p-value cutoff in Figure 1-7.

GWAS Plus: *GCTA*.

We first computed the genetic relationship matrix **A** from all 7,702 White participants with genome-wide SNP data, which includes those for whom FSIQ scores were not available. We then ran *GCTA* with FSIQ as the phenotype and with birth-year and sex as fixed effects, 234 times over. Each *GCTA* run used a different ceiling for the maximum allowable degree of genetic

[4] We also calculated Nagelkerke's (1991) generalized R^2, and the squared Pearson correlation between polygenic score and residuallized phenotype. Nagelkerke's R^2 was typically very close to Buse's. The squared Pearson correlation was generally close to the other two, but tended to be higher, sometimes as much as twice as high as Buse's.

relatedness among participants. This ceiling ranged from 0.005 to 1.17, with increments of 0.005. As this ceiling increased, both the number of participants, and the degree to which participants in the analysis could be genetically related to one another, increased.

Results

GWAS

Figures 1-1 and 1-2 are "Manhattan plots" of the GWAS p-values from, respectively, the 2,546,647 observed and imputed SNPs, and the 527,829 observed SNPs only. The y-axis of a Manhattan plot is $-\log_{10}(p)$. The x-axis is divided into chromosomes, and within each chromosome, SNPs are ordered by base-pair position. Chromosomes above #22 refer to different parts of the sex chromosomes and to mitochondrial DNA (see figure captions). When only the observed SNPs are plotted, the only elevation above 6 occurs on chromosome 1. When observed and imputed SNPs are both included, more SNPs in that same region of chromosome 1 are elevated above 6, and there are elevations visible on chromosomes 16 and 21 as well. No SNPs reached genome-wide significance, which in this metric would be $-\log_{10}(p) > 7.30$. The two most extreme SNPs, rs3856228 and rs10922924, are on chromosome 1 but not located in a known gene.

Under the null hypothesis, p-values are uniformly distributed on (0, 1). Figures 1-3 and 1-4 are uniform quantile-quantile (QQ) plots of the GWAS p-values from, respectively, the 2,546,647 observed and imputed SNPs, and the 527,829 observed SNPs only. Under the null hypothesis, p-values from independent statistical tests are expected to follow the diagonal red line. Both QQ plots show some divergence from the null distribution, where the observed p-values tend to be more extreme than expected. To quantify this deviation, we can convert the p-values to quantiles from a chi-square distribution on 1 *df*, and compare their median to the null expectation of 0.455. When this is done with observed and imputed SNPs together, the median is 0.476; when done with observed SNPs only, the median is 0.470. This departure from the null may indicate massively polygenic inheritance of FSIQ, wherein few if any SNPs yield genome-wide significant association signals, but the overall distribution of test statistics reflects the presence of a large number of nonzero effects (Yang, Weedon, et al., 2011).

There are clearly some p-values that lie outside the confidence limits in Figure 1-3. However, because of LD among SNPs, the assumption of independent statistical tests is violated to begin with, and so one extreme result usually carries others with it. It stands to reason that this effect of LD would be more pronounced when imputed SNPs are included, since imputation methods rely on the LD (correlation) structure among SNPs to achieve denser coverage of the genome.

VEGAS

What *VEGAS* essentially does is test whether all SNP p-values in a gene significantly differ in distribution from the null. The resulting gene-based p-values from *VEGAS* are depicted in Figure 1-5, a Manhattan plot, and Figure 1-6, a QQ plot. Figure 1-6 suggests that *VEGAS* has a somewhat conservative bias in these data. This is most likely because the LD structure in HapMap CEU does not perfectly match the actual LD structure in our data. No gene in Figure 1-5 reaches the genome-wide significance level recommended for *VEGAS*, which in this metric would be $-\log_{10}(p) > 5.55$. When imputed as well as observed SNPs were input to *VEGAS* (results not shown), the only observable change was to push the signals on chromosomes 6 and 8 above 4. Our data do not support association of *FNBP1L* with GCA (p = 0.662).

Polygenic scoring

Figure 1-7 depicts cross-validation performance (Buse's R^2) of polygenic score, by subsample, p-value cutoff, and weighting scheme. One notable result here is that the polygenic score, when calculated from signed unit-weighted SNPs, performed about as well as when it was calculated from the actual single-SNP GWAS regression weights. We attempted the unit-weighting to strike a different balance between bias and variance. The GWAS regression weights, while unbiased, are estimated with considerable sampling error. On the other hand, unit weights are presumably biased, but possibly less variable over repeated sampling. In fact, unit weights can rival optimal least-squares weights in terms of predictive accuracy, especially when the overall amount of predictive error is large (Dana & Dawes, 2004).

Another result evident in Figure 1-7 is the trend in cross-validation performance across *p*-value cutoffs. Though there is a great deal of variation, the peak in predictive accuracy occurred at the very generous cutoff of $p \leq 0.5$ (with best R^2 slightly above 0.7%). This is a bit surprising, since one might expect that the peak would occur at a more stringent cutoff, and that most SNPs with $p \leq 0.5$ would be irrelevant noise. Peak R^2 occurred at stricter cutoffs for the three replication cohorts of Benyamin et al. (2013), including the one from MCTFR (at $p \leq 0.01$), which is a subsample of the present study sample. Likewise, in Lango Allen et al.'s (2010) report on height, the average R^2 across five validation samples was highest at $p \leq 0.001$. However, both Benyamin et al. and Lango Allen et al. had the advantage of larger calibration samples than we did here. With larger calibration samples, estimates of SNP weights have less sampling error, and a given non-null effect size will produce a smaller *p*-value.

GCTA

Figure 1-8 graphs $\hat{h}^2_{SNP}$ as a function of genetic-relatedness ceiling, with error bars representing ± 1 standard error. As the ceiling increased, more participants were included in the analysis, and the statistical precision of $\hat{h}^2_{SNP}$ increased. More importantly, the point estimate itself increased as well. Below a ceiling of 0.015, sample size was less than 300, and the software produced nonsensical negative point estimates. A noticeable spike in $\hat{h}^2_{SNP}$ is evident around 0.5, where full siblings (including DZ twins) were introduced. Something similar happens around 1.0, where the MZ twins were introduced. Figure 1-10 shows how the sample size of the GCTA analysis increases with relatedness ceiling, and Figure 1-9 shows $\hat{h}^2_{SNP}$ as a function of sample size rather than the ceiling.

According to Yang, Lee, et al. (2011), if the purpose of the analysis is estimate how much phenotypic variance is attributable to the common SNPs on a genome-wide array, then close relatives should be excluded from analysis. They suggest a genetic-relatedness ceiling of 0.025. The reason for excluding close relatives is that, if they are instead included, then the *GCTA* estimator may overestimate the actual variance attributable to the SNPs on the array. When close relatives are included, the *GCTA* estimator functions more like a pedigree-based estimator, which captures the influence of all trait-relevant polymorphisms that contribute to familial resemblance, no matter how rare, and not just the genotyped SNPs (and other polymorphisms the SNPs tag). In the extreme case, then, a *GCTA* variance component estimated from MZ twins could even reflect the influence of *de novo* mutations, which contribute to MZ-twin resemblance but are not marked in the population by common SNPs. It would seem, then, that if one wants to estimate the *overall* heritability of a phenotype with *GCTA*, inclusion of close relatives is the way to go. It might be tempting to conclude that the $\hat{h}^2_{SNP}$ values in Figure 1-8 produced when the relatedness ceiling is above 1.0 are molecular-genetic estimates of the true, broad-sense heritability of GCA. But that is not necessarily the case: as Yang, Lee, et al. also remind us, including close relatives confounds genetic resemblance with shared-environmental influence.

Imposing Yang, Lee, et al.'s (2011) suggested genetic-relatedness ceiling of 0.025 in our dataset, $N = 3{,}322$ of our participants were included, yielding $\hat{h}^2_{SNP} = 0.35$ ($SE = 0.11$). When we systematically incremented the ceiling by regular intervals, three sample-size plateaus were evident (Figure 1-10). For ceilings between 0.1 and 0.48, N and $\hat{h}^2_{SNP}$ were steady around 3,600 and 0.45, respectively. Between ceilings of 0.56 to 0.98, they were steady at about 6,050 and

0.66. When all 7,100 GWAS participants were included at a ceiling of 1.17, GCTA yielded $\hat{h}^2_{SNP} = 0.77$ (SE = 0.01).

Discussion

The present study is a "GWAS Plus" for general cognitive ability, conducted in a sample of over 7,100 Caucasian participants from two longitudinal family studies. We conducted the GWAS *per se* using 2,546,647 SNPs: 527,829 from the Illumina 660W-Quad array, plus 2,018,818 imputed with reasonable reliability (imputation $R^2 > 0.5$). The "Plus" in "GWAS Plus" refers to our additional analyses that involve predicting the phenotype from more than one SNP at a time. These analyses were (1) gene-based association tests in *VEGAS*, (2) polygenic scoring with five-fold cross-validation, and (3) a genomic-relatedness restricted maximum-likelihood analysis in *GCTA*.

Our most disappointing results were from *VEGAS* (Figures 1-5 and 1-6). No gene achieved genome-wide significance ($p < 2.8 \times 10^{-6}$ or $-\log_{10}(p) > 5.55$), and the method appears to be slightly conservatively biased in our dataset, possibly because of differences between our actual LD structure and that of *VEGAS*' reference dataset, HapMap CEU. Running *VEGAS* with LD estimated from data is possible, but it seems doubtful that the LD misspecification could be so severe as to suppress a robustly significant association signal. Certainly the most *a priori* plausible gene, *FNBP1L*, is not supported in our sample (*p* = 0.662).

Polygenic scoring is another way of combining the predictive power of multiple SNPs. At best, the polygenic score could predict 0.7% of variance in our analyses, which occurred with a *p*-value cutoff of $p \leq 0.5$. Presumably, better results could be obtained at stricter *p*-value cutoffs when the calibration sample is larger. Interestingly, our cross-validation analysis showed that signed unit SNP weights performed about as well as GWAS regression weights. This suggests that, at least when the calibration sample is relatively small, there is negligible loss in predictive accuracy when fixing all SNP effects to the same absolute magnitude, and using GWAS merely to determine the direction of each SNP's effect. We are somewhat surprised at the relative performance of the polygenic score at inclusive vis-à-vis exclusive *p*-value cutoffs. Evidently, our most significant SNPs had limited predictive power, but a heap of non-significant SNPs can better contribute to prediction in the aggregate. Polygenic scores calculated from all 527,829 genotyped SNPs at best account for about 0.7% of phenotypic variance, a value that contrasts sharply with parameter estimates from *GCTA*, even though both represent the proportion of variance attributable to literally every SNP on the array.

No single SNP has yet been replicably associated with human intelligence at genome-wide significance levels, and our GWAS results do not change that fact. This is not surprising, though, in light of our GWAS' limited power. Given a conservative estimate of our effective sample size, we would have slightly above 80% power to detect a SNP accounting for 2% of phenotypic variance, which constitutes rather poor power. Even given an aggressive estimate of effective sample size, we would have slightly above 80% power to detect a SNP accounting for 0.6% of variance. But if realistic effect sizes are even smaller, like on the order of 0.2% to 0.4% (Butcher et al., 2008; Davis et al., 2010), this would still be inadequate.

Even though we lacked sufficient power to detect a realistic SNP effect at genome-wide significance levels, the overall distribution of our test statistics and *p*-values differs slightly but appreciably from the null. This kind of genomic inflation can reflect population stratification (e.g., Marchini, Cardon, Phillips, & Donnelly, 2004) which is doubtful in our case, because we carefully ensured that all our participants are White, and included 10 principal components from EIGENSTRAT as covariates. Instead, we interpret this genomic inflation as evidence of the massive polygenicity of GCA (Yang, Weedon, et al., 2011).

We regard our *GCTA* results as the most impressive and informative. The performance of our polygenic score at inclusive *p*-value cutoffs, plus the genomic inflation evident in our GWAS, suggest that there is a very large number of trait-relevant polymorphisms, each with a very small individual effect. But our results from GCTA—which were similar to those of earlier studies (Davies et al., 2011; Benyamin et al., 2013)—make that clear beyond any doubt. We surmise that few behavior geneticists, once they understood the GREML method, were surprised that almost half the variance in cognitive ability and in height (Yang et al., 2010) could be attributed to genotyped SNPs on a chip. But, that is precisely why *GCTA* is so monumental: it has furnished molecular genetics with the result that quantitative genetics has predicted for decades. Its theoretical import cannot be overstated: it vindicates the classical theory of polygenic inheritance. We see now how truly Fisher (1918) wrote when he penned these words: "the statistical properties of any feature determined by a *large number* of Mendelian factors have been successfully elucidated…In general, the hypothesis of cumulative Mendelian factors seems to fit the facts very accurately" (p.432-433, emphasis supplied).

At low genetic-relatedness cutoffs, *GCTA* provided us with h^2_{SNP} estimates around 40%. This effect size seems to be typical for cognitive ability (Davies et al., 2011; Benyamin et al., 2013). But, biometrical heritability estimates for GCA are typically in the range of 50% to 70%. This outcome, that through GREML methods, common SNPs on a genome-wide array can account for most but not all of the heritability of a trait, also appears typical for cognitive ability (Davies et al.; Benyamin et al.) and for that archetypal polygenic quantitative trait, height (Yang et al., 2010). For such traits, this is the current state of what is known as the problem of "missing heritability" (Maher, 2008). What, then, might be the molecular basis for the heritability that is not captured by *GCTA* estimates? Since h^2_{SNP} represents the proportion of phenotypic variance attributable to common SNPs on the array (and variants in tight LD with them), it stands to reason that the missing heritability might be due to polymorphisms that are not common, or are not SNPs, such as copy-number variants. In any event, if specific polymorphisms underlying variation in GCA are to be discovered, massive sample sizes will be necessary. But in the meantime, we can conclude that there are a great many unspecified polymorphisms associated with GCA, each with a very small effect—general cognitive ability is indeed “heritable [and] highly polygenic” (Benyamin et al., p. 1).

Table 1-1. Descriptive characteristics of Study #1 sample.

	Parents	Twins (17yo)	Twins (11yo)	Non-twin Biological Offspring	Adoptees	Step-parents
N	3,264	1,146	2,080	414	116	80
Female(%)	60.2%	55.3%	50.1%	52.2%	46.6%	8.8%
Mean Age at Intake (SD)	43.3 (5.46)	17.5 (0.45)	11.8 (0.43)	14.9 (1.89)	15.1 (2.17)	40.6 (7.45)
Mean FSIQ (SD)	105.8 (14.2)	100.4 (14.1)	103.6 (13.5)	108.5 (13.1)	105.7 (14.3)	103.4 (15.7)

Table notes: Total *N* = 7100, in 2376 families. FSIQ = Full-Scale IQ; 17yo = 17-year-old cohort; 11yo = 11-year-old cohort. For a minority of twins (38%), FSIQ represents a within-person average of FSIQ scores from more than one assessment (see text). FSIQ range: 151 – 59 = 92. Parental intake age range: 65 – 28 = 37. Offspring intake age range: 20 – 10 = 10.

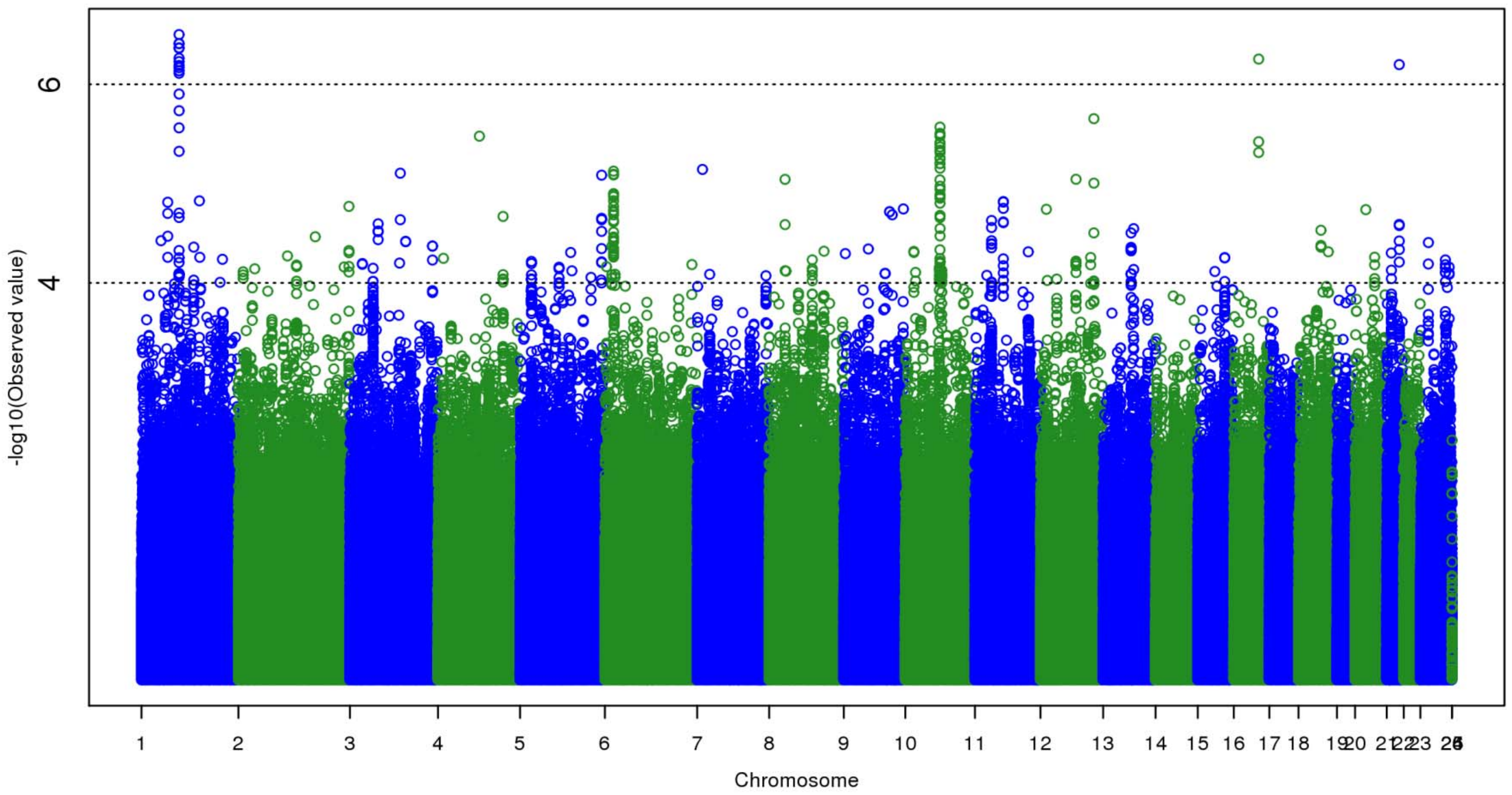
Table 1-2. *RFGLS* parameter estimates for regression of FSIQ onto covariates only.

Fixed Effects		
Parameter	Estimate	SE
Intercept	115.19	0.84
Sex	-2.81	0.32
Birth Year	-0.09	0.01
PC1	0.11	0.14
PC2	-0.23	0.13
PC3	-0.12	0.13
PC4	-0.05	0.13
PC5	-0.14	0.13
PC6	-0.25	0.14
PC7	0.17	0.13
PC8	0.12	0.13
PC9	0.22	0.13
PC10	-0.12	0.13

Random Effects		
Parameter	Estimate	SE
Correlation, Spousal	0.33	0.06
Correlation, Bio Mother-Child	0.43	0.04
Correlation, Bio Father-Child	0.42	0.05
Correlation, MZ twin	0.80	0.06
Correlation, Full Siblings	0.48	0.07
Correlation, Adopt Mother-Child	0.17	0.26
Correlation, Adopt Father-Child	0.49	1.09
Correlation, Adoptive Siblings	0.33	0.25
Variance, Offspring	193.95	0.03
Variance, Mothers	186.35	0.03
Variance, Fathers	211.47	0.04
Variance, Stepparents	266.48	0.16

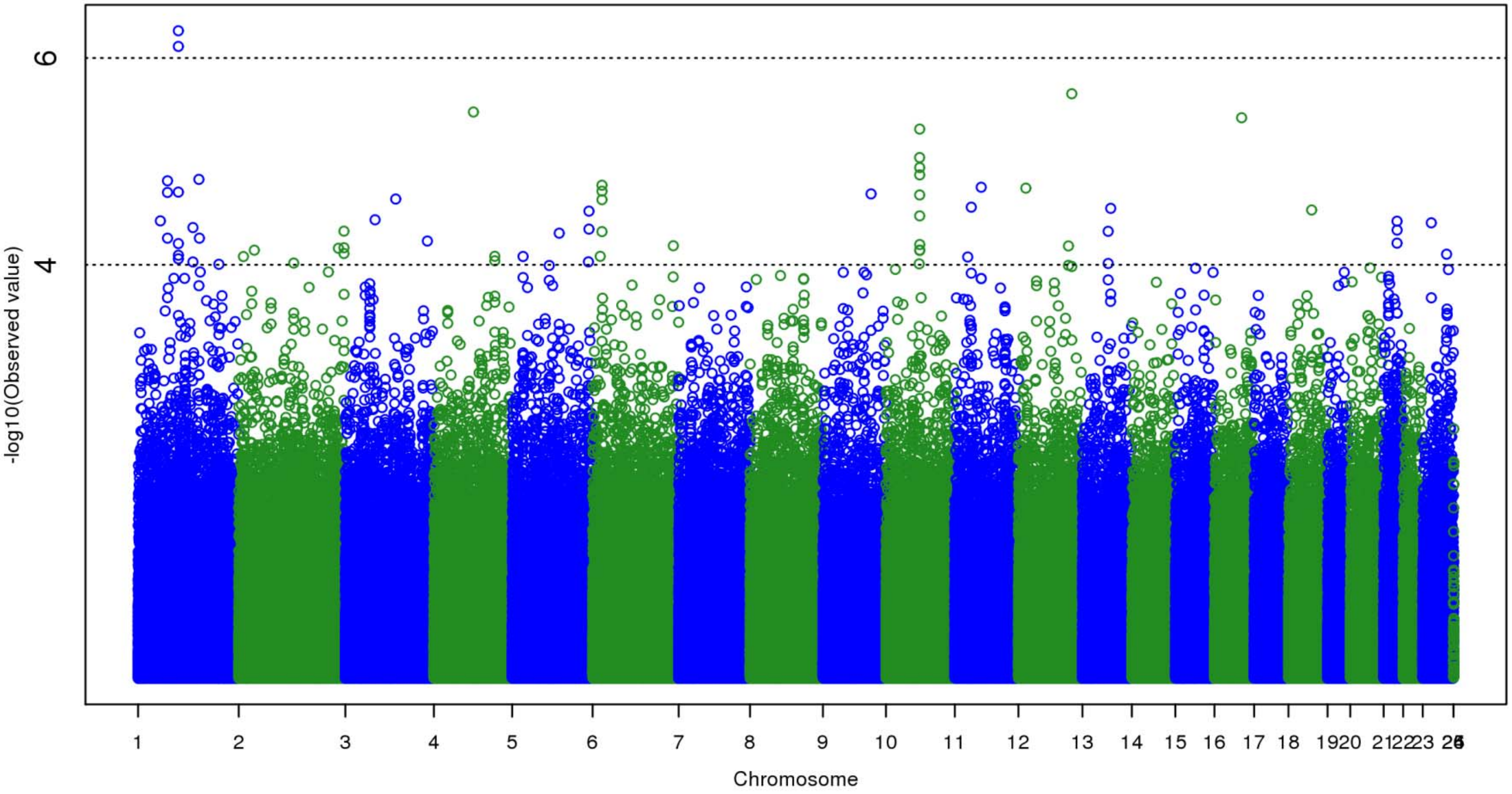
Table notes: Covariates were birth year (2 digits), dummy variable for female sex, and the first 10 EIGENSTRAT principal components. The random effects are the parameters *RFGLS* uses to assemble the residual covariance matrix.

Figure 1-1. Manhattan plot of GWAS *p*-values, all 2,546,647 observed and imputed SNPs.



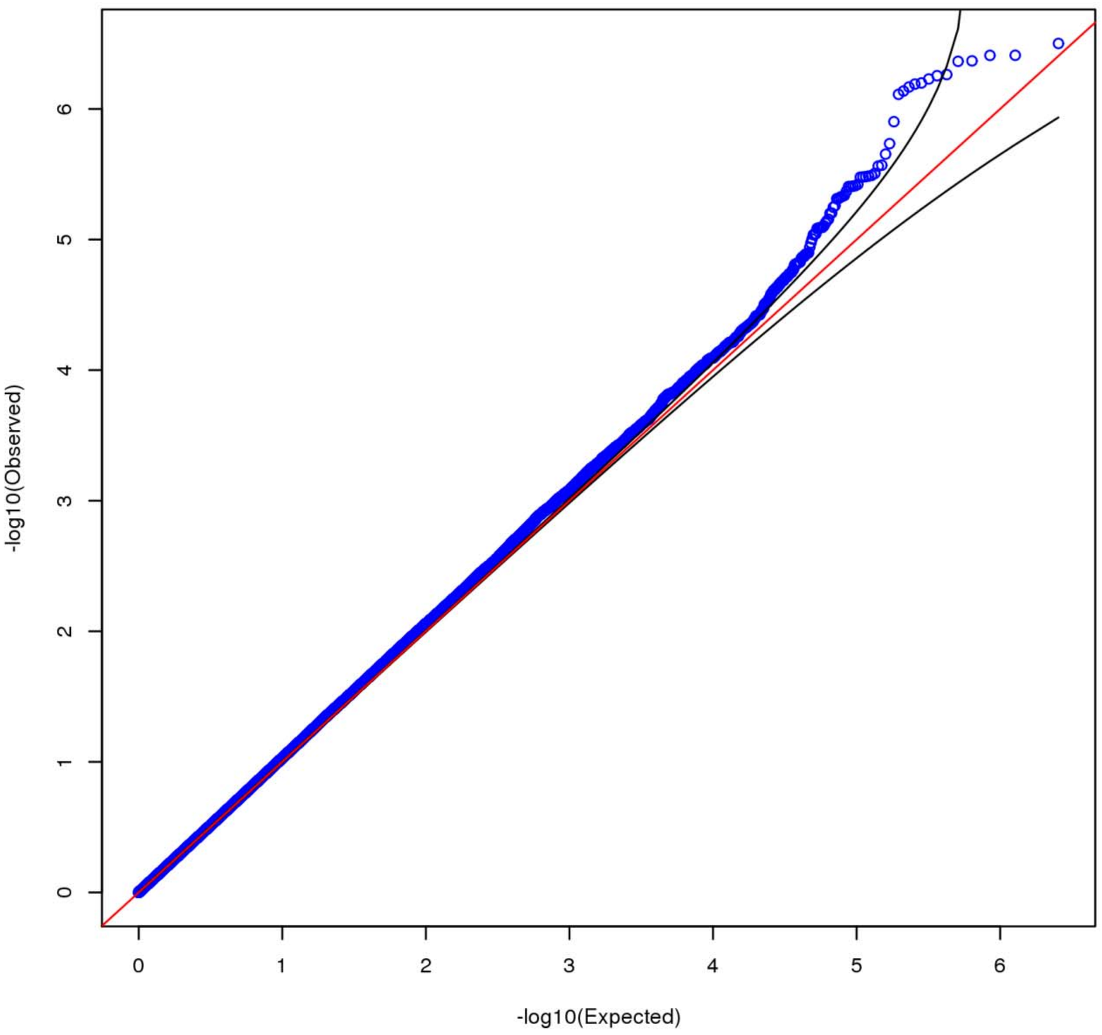
Chromosome 23 = X chromosome, chromosome 24 = Y chromosome, chromosome 25 = pseudoautosomal region of sex chromosome. Chromosome 26 indicates mitochondrial DNA. Genome-wide significance is $-\log_{10}(p) > 7.30$, which no SNP reaches.

Figure 1-2. Manhattan plot of GWAS *p*-values from 527,829 observed SNPs only.



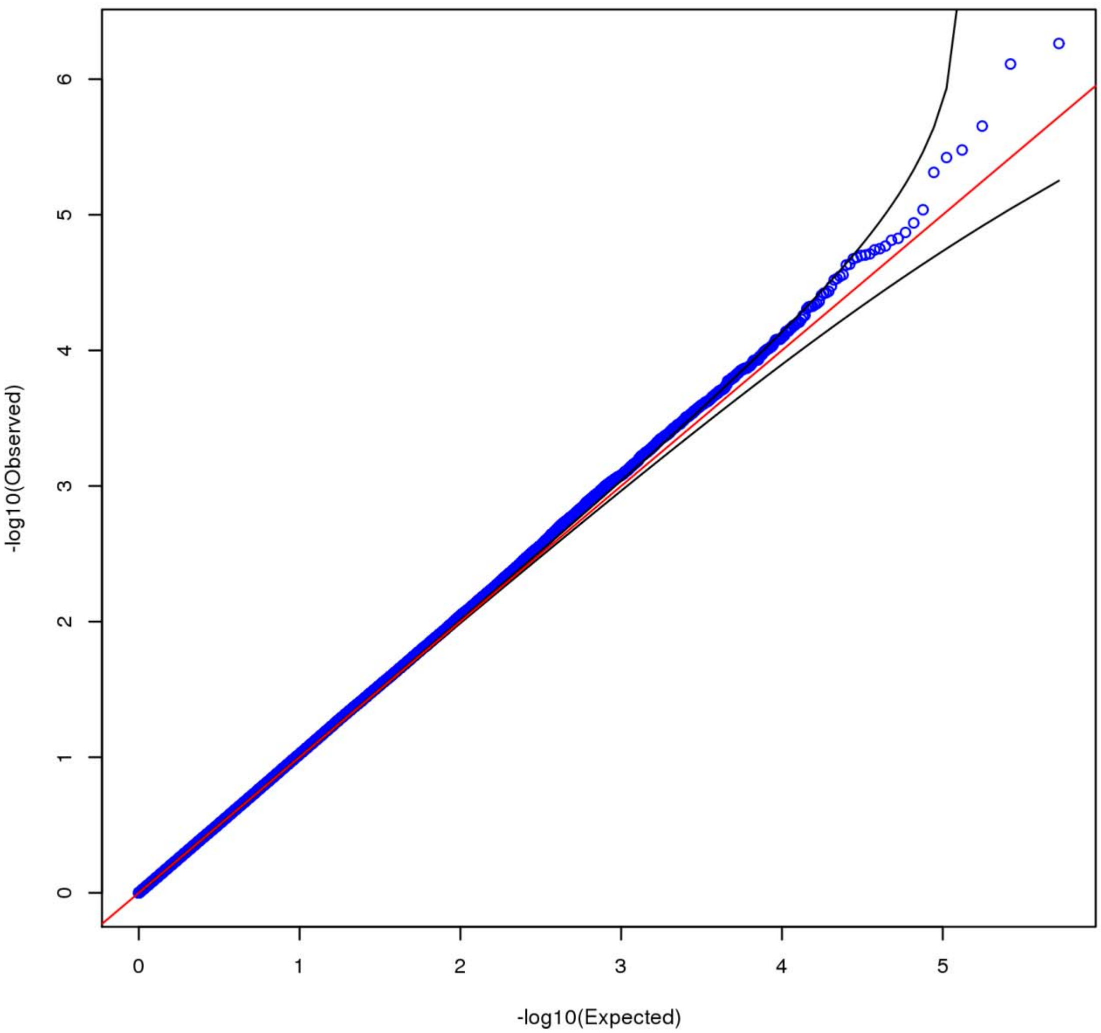
Chromosome 23 = X chromosome, chromosome 24 = Y chromosome, chromosome 25 = pseudoautosomal region of sex chromosome. Chromosome 26 indicates mitochondrial DNA. Genome-wide significance is $-\log_{10}(p) > 7.30$, which no SNP reaches.

Figure 1-3. Uniform quantile-quantile plot of GWAS *p*-values, all 2,546,647 observed and imputed SNPs.



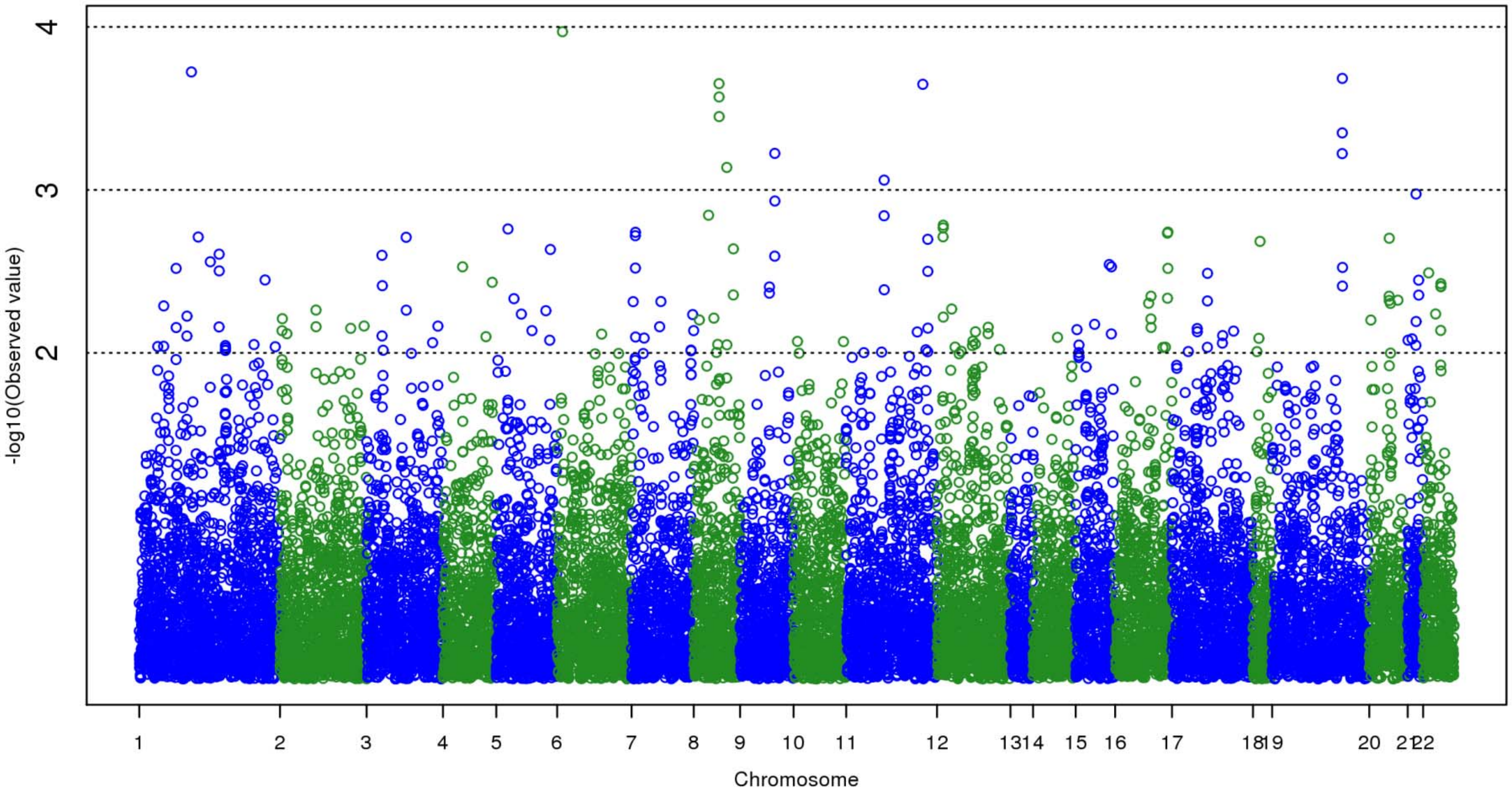
The black curves delineate 95% confidence limits.

Figure 1-4. Uniform quantile-quantile plot for GWAS *p*-values from 527,829 observed SNPs only.



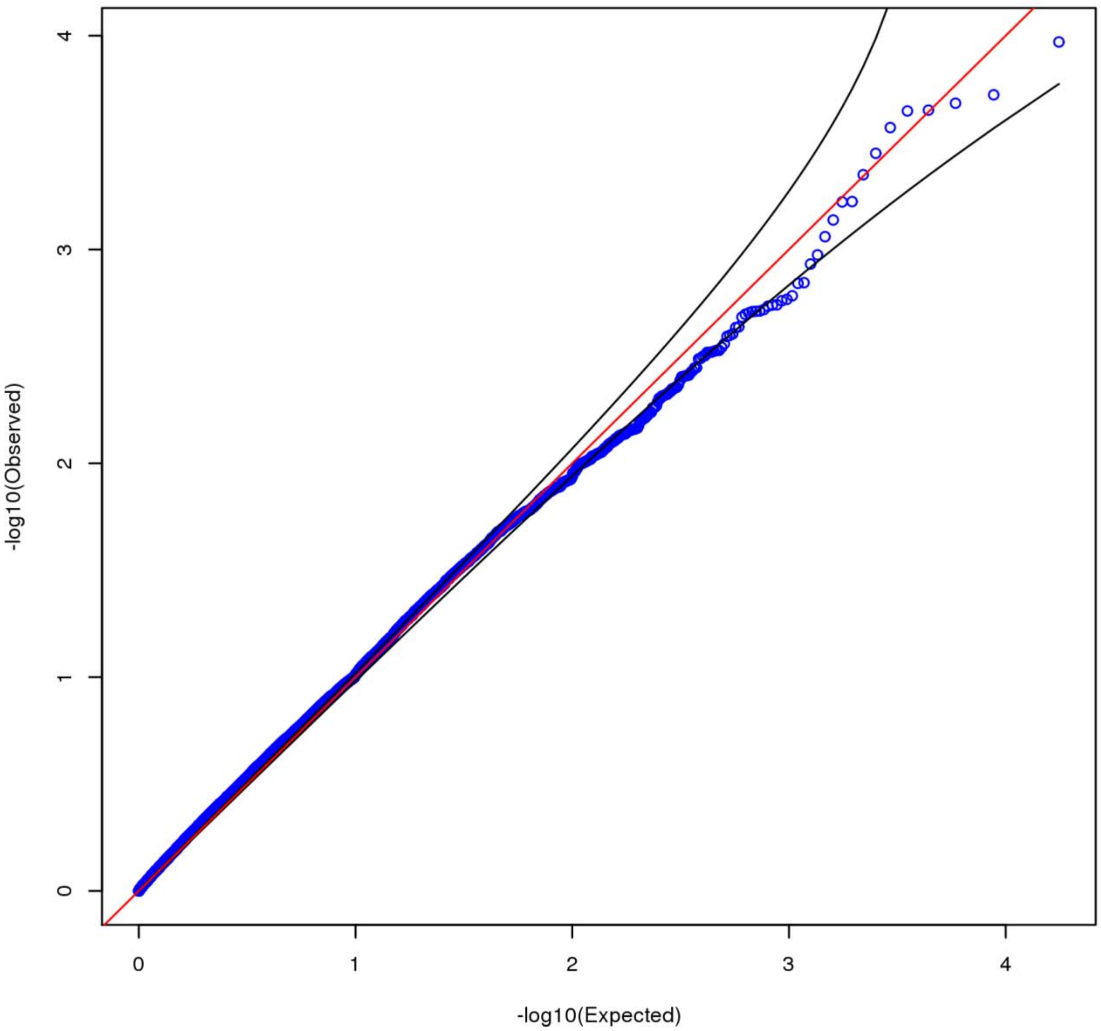
The black curves delineate 95% confidence limits.

Figure 1-5. Manhattan plot for gene-based *p*-values from *VEGAS*.



Analysis input was GWAS *p*-values from 515,385 autosomal SNPs on the Illumina array. Abscissa position of each point is the gene's beginning base-pair position, NCBI genome build 36. Genome-wide significance is $-\log_{10}(p) > 5.55$, which no gene reaches.

Figure 1-6. Uniform quantile-quantile plot for gene-based *p*-values from *VEGAS*.



Analysis input was GWAS *p*-values from 515,385 autosomal SNPs on the Illumina array. Black curves delineate 95% confidence limits.

Figure 1-7. Five-fold cross-validation of polygenic score, predicting FSIQ residuallized for covariates.

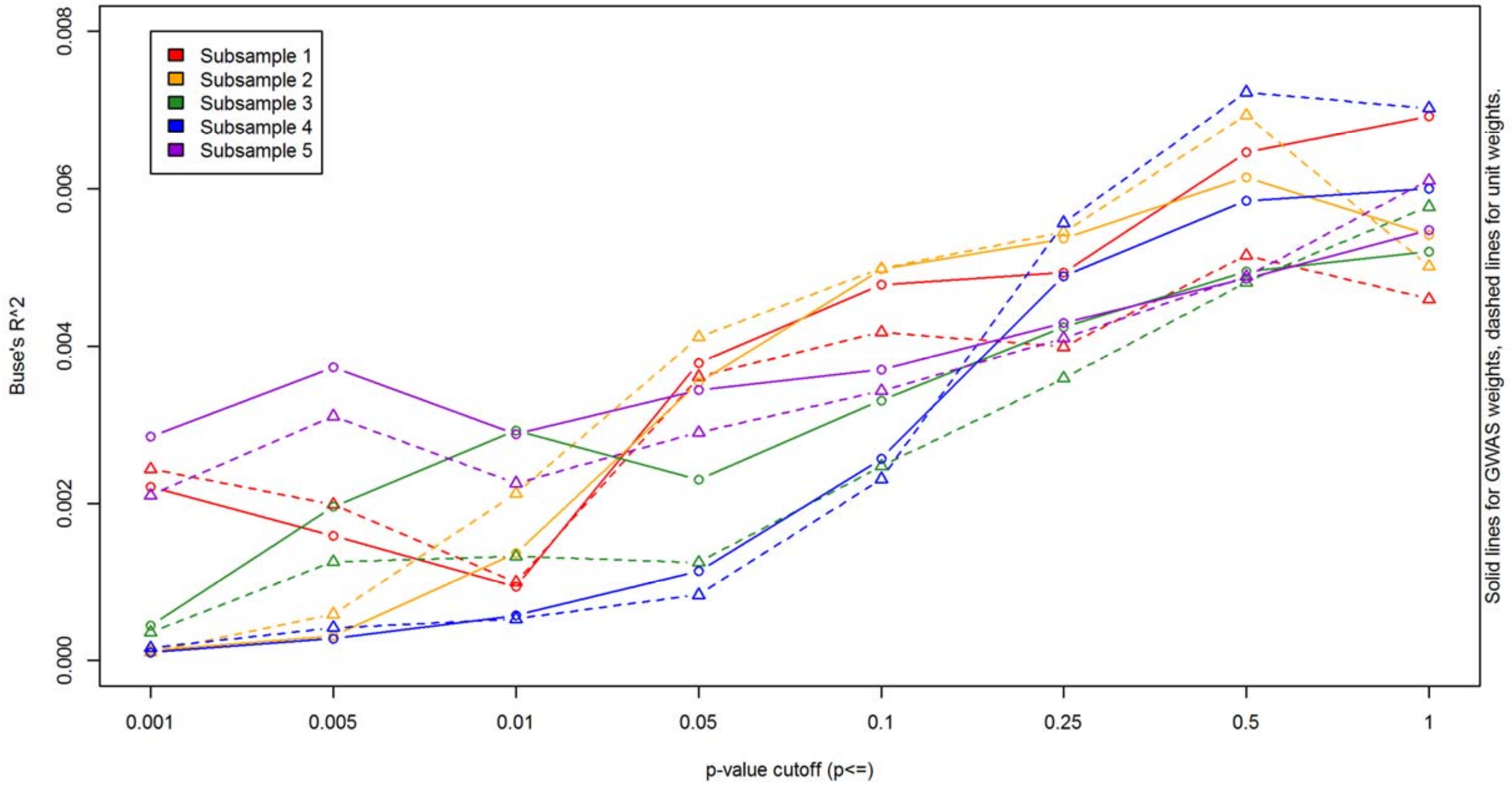
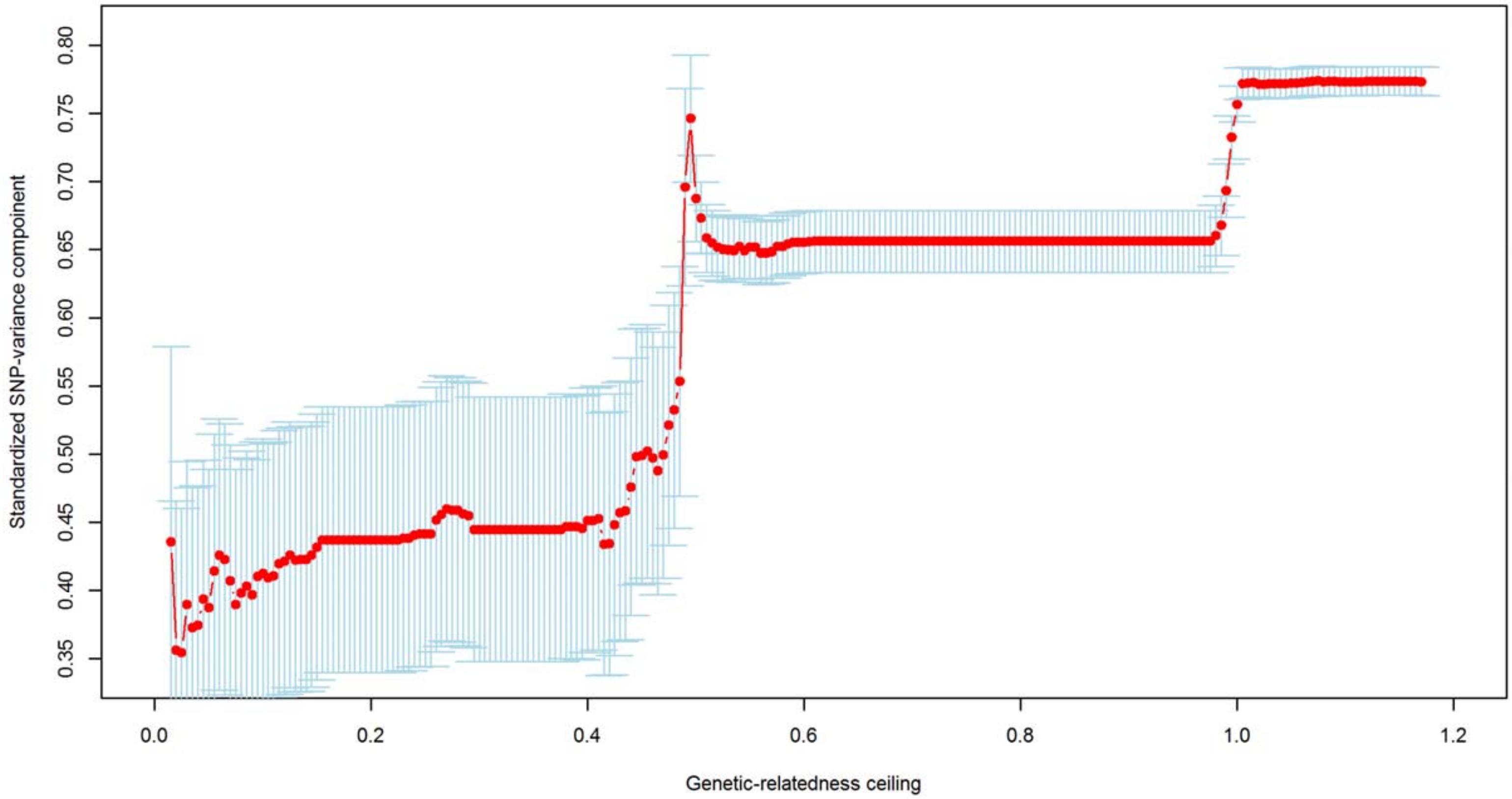


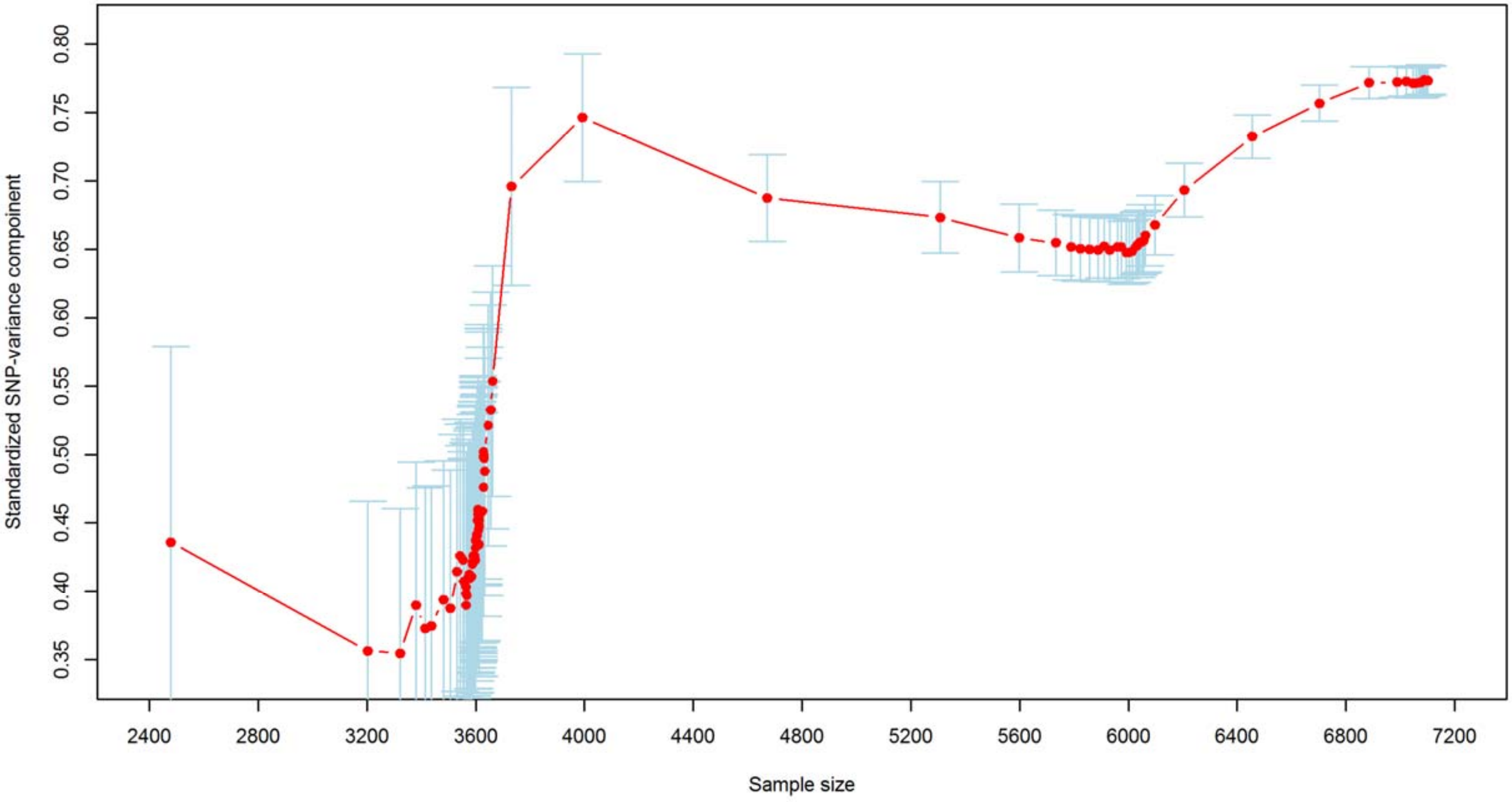
Figure depicts cross-validation R^2 (Buse, 1973) for predicting residuallized FSIQ in the indicated subsample from polygenic score calculated from GWAS regression weights obtained in the other 4 subsamples. "*P*-value cutoff" dictated how small a SNP's *p*-value had to be in the calibration GWAS to be included in calculating polygenic score for the validation sample. Polygenic score was either calculated directly from the GWAS weights (solid lines) or from signed unit weights (dashed lines; see text).

Figure 1-8. *GCTA* $\hat{h}^2_{SNP}$ as calculated at different genetic-relatedness ceilings.



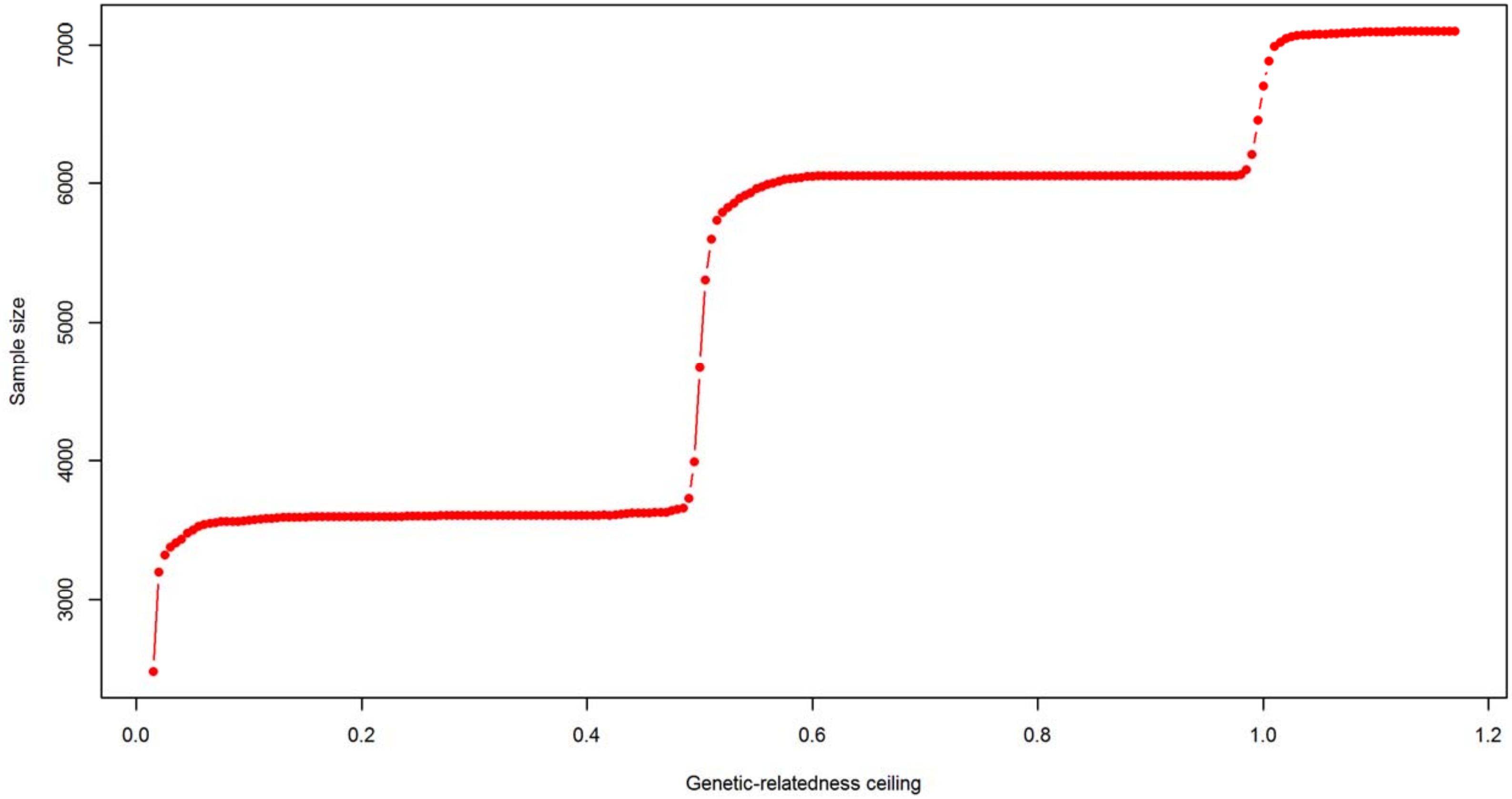
Error bars are ±1 standard error. Genetic-relatedness ceiling is the maximum degree of genetic relationship allowed among participants entered into analysis.

Figure 1-9. *GCTA* $\hat{h}^2_{SNP}$ as calculated at different genetic-relatedness ceilings, graphed as a function of sample size.



Error bars are ±1 standard error. Sample size was not manipulated directly, but was instead a consequence of genetic-relatedness ceiling. Genetic-relatedness ceiling is the maximum degree of genetic relationship allowed among participants entered into analysis.

Figure 1-10. *GCTA* sample size as function of genetic-relatedness ceiling.



Genetic-relatedness ceiling is the maximum degree of genetic relationship allowed among participants entered into analysis.

Study #2

Background

General cognitive ability (GCA) is the theoretical construct involved to some extent in every cognitively demanding task. Frequently identified with Spearman's (1904) general factor *g*, it is posited to contribute some part of the variance in scores on all mental-ability tests. GCA correlates appreciably with other variables in a striking variety of domains (Gottfredson, 2003; Herrnstein & Murray, 1996; Jensen, 1998; Deary, 2012).

Decades of twin, adoption, and family studies have firmly established via biometric methods that general cognitive ability is substantially heritable (Bouchard & McGue, 2003; Deary, Johnson, & Houlihan, 2009). At the same time, the search for underlying genetic factors via molecular methods was stymied by the limited power of linkage analysis (Risch & Merikangas, 1996) and the persistent problem of replication failure in candidate-gene association studies (Ioannidis, Ntzani, Trikalinos, & Contopoulos-Ioannidis, 2001; Hirschorn, Lohmueller, Byrne, & Hirschorn, 2002; Chabris, Hebert, Benjamin, Beauchamp, Cesarini, et al., 2012). As a result, investigators have more recently adopted the genome-wide association study (GWAS) as a routine method in molecular-genetic research. GWAS in the modern sense entails genotyping human subjects on at least 300,000 single-nucleotide polymorphisms (SNPs) throughout the genome (Balding, 2006). To date, several GWAS for human intelligence have been published (Butcher, Davis, Craig, & Plomin, 2008; Davis et al., 2010; Davies et al., 2011; Benyamin et al., 2013).

One of the lessons to be learned from GWAS with quantitative phenotypes is just how small the effect of any single SNP is going to be. For example, the largest effect sizes reported from the Twins Early Development Study GWAS are $r^2 = 0.004$ (Butcher et al., 2008) and $r^2 = 0.0024$ (Davis et al., 2010). The same lesson is apparent from GWAS of the archetypal quantitative phenotype, human height (Visscher, 2008). However, there are classes of genetic variation other than SNPs, which might instead more powerfully account for heritable variance in complex traits. One of these is the copy-number variant (CNV). According to Scherer et al.'s (2007) taxonomy of genomic variation, a CNV is a submicroscopic structural variant of at least 1000 base pairs that appears a different number of times in the genomes of different individuals. CNVs include deletions, in which case the individual will have fewer than the typical two copies of the DNA sequence, as well as duplications, where more than two copies are present.

Previous research has implicated CNVs in psychiatric diseases, including autism (Sebat et al., 2007; Pinto et al., 2010) and schizophrenia (The International Schizophrenia Consortium, 2008). Autism is highly comorbid with mental retardation (American Psychiatric Association, 1994), and the connection between schizophrenia and pre-morbid cognitive deficits is well-documented (Woodberry, Giuliano, & Seidman, 2008). This suggests the possibility that CNVs underlie part of the heritable variance of GCA. Indeed, large-scale, cytogenetically visible structural duplications and deletions of chromosomal material can cause syndromal forms of mental retardation, the textbook example being trisomy 21 (Down syndrome), caused by a redundant copy of an entire chromosome. CNVs (as defined here) are smaller-scale and submicroscopic, but recent technological advances have enabled identification of a growing list of specific deletions and duplications associated with syndromal intellectual disability (Mefford, Batshaw, & Hoffman, 2012), as well as neurodevelopmental disorders more generally (Glessner, Connolly, & Hakonarson, 2012; Coe, Girirajan, & Eichler, 2012). Thus, one research strategy would be to search for specific CNVs contributing to normal-range variation in cognitive functioning. However, if the same lessons learned from SNPs for quantitative phenotypes also hold for CNVs, the effect-sizes of relatively common deletions and duplications will be quite small. This is not to say that large-effect CNVs are unlikely to exist for GCA, but the reliability of calls for low-baserate CNVs can be limited (Cooper & Mefford, 2011), and analyses of rare variants will generally be underpowered unless the sample size is quite large. Therefore, another strategy is to aggregate CNVs from across the genome, into one or more mutational burden scores. Both the International Schizophrenia Consortium (2008) and Pinto et al.'s autism study (2010) exploited this strategy, and report significant case-control burden differences. Further, a recent review of the role of CNVs in neurodevelopmental disorders (Coe et al., 2012) concluded that a greater burden of larger and/or more numerous small CNVs typically corresponds to greater phenotypic severity.

Contemporary neurobiological theory of general intelligence recognizes the key role of a distributed network of frontal and parietal brain structures (Jung & Haier, 2007; Gläscher et al., 2010), and much evidence supports the hypothesis that the brains of more-intelligent individuals function more efficiently, in terms of energy consumption during a cognitively demanding task (Neubauer & Fink, 2009). GCA seemingly depends upon the efficient functioning and coordinated directed effort of a distributed array of neural structures, and the functioning of such a system is far easier to disturb than enhance. Approximately 84% of human genes are expressed in brain (Hawrylycz et al., 2012), and in light of CNVs' role in pathological cognitive deficits, it seems reasonable to hypothesize that they could contribute to normal-range variation as well. Specifically, individuals whose genomes show greater deviation from "typical" reference copy-number states would more likely to harbor detrimental trait-relevant mutations, and have correspondingly lower cognitive ability scores.

An organism's total load of mutations relative to the "typical" reference genome could reflect developmental instability, particularly if the mutations in question are rare. In turn, developmental instability is frequently operationalized as fluctuating asymmetry (deviation from morphological bilateral symmetry; Gangestad & Thornhill, 1999), which correlates negatively with GCA (Banks, Batchelor, & McDaniel, 2010). Therefore, one might hypothesize that mutational burden would also associate negatively with GCA. This evolutionary-biological hypothesis motivated a small-sample study of CNVs and IQ, by Yeo, Gangestad, Liu, Calhoun, and Hutchinson (2011). The study participants entered into analysis were $N = 74$ adults diagnosed with alcohol dependence, who had been assessed with the Wechsler Abbreviated Scale of Intelligence. From DNA samples, Yeo et al. detected a total of 13,557 rare CNVs, 7,249 of which were deletions, and 6,308 of which were duplications. Detected copy-number deletions in the sample ranged from ~8kb to ~626kb in length; the average copy-number deletion was ~210kb long ($SD \approx$ 14kb). The length (in kilobases) of the copy-number deletions participants carried correlated negatively with their Full-scale IQ (FSIQ) scores ($r = -0.30$, $p = 0.01$). In contrast, participants' counts of deletions carried correlated *positively* with FSIQ, though not significantly so ($r = 0.21$, $p = 0.08$). The number of deletions carried ranged from 1 to 25 in the sample, with an average of 10.95 ($SD = 5.48$). Neither the length nor count of copy-number duplications correlated significantly with FSIQ, and both correlations were less than 0.10 in absolute magnitude. Yeo et al. (2011) interpret their result for copy-deletion length as consistent with the hypothesis that individual differences in cognitive ability result partly from individual variation in the total burden of detrimental mutations carried.

Three recent studies with larger samples have attempted to replicate Yeo et al. (2011). One of these (MacLeod et al., 2012) was a study of both fluid and crystallized intelligence in a sample of over 3,000 older British adults genotyped on the Illumina 610-Quadv1 chip. MacLeod et al. called CNVs using both *PennCNV* (Wang et al., 2007) and *QuantiSNP* (Colella et al., 2007), retaining only those calls produced by both. No association was observed with rare-CNV burden. Suggestive evidence of association with fluid intelligence was observed for a specific CNV overlapping with *SHANK3* (permutation-corrected $p = 0.01$). But, of the three mutant carriers in the sample, two had duplications and one had a deletion, which MacLeod et al. regard as counter-intuitive.

Bagshaw et al. (2013) reported a study of IQ and academic achievement conducted in a sample of 717 participants from the longitudinal Christchurch Health and Development Study in

New Zealand. These participants were genotyped on the Illumina 660W-Quad chip. Bagshaw et al. called CNVs with *PennCNV*, and conducted a rare-CNV burden analysis and a genome-wide scan of common CNVs. They observed no strong evidence of association, and the only suggestive association signals were for academic achievement, and not IQ, which we regard as a superior measure of the GCA construct.

The third recent study of interest is McRae et al. (2013), which was conducted in a sample of 800 Australian adolescents, who were IQ-tested around age 16, and genotyped on the Illumina 610K Quad array. McRae et al. called CNVs using *QuantiSNP*, and performed a mutational burden analysis and a genome-wide scan with rare CNVs, obtaining only null results.

Below, we report the results from our study of CNVs and IQ, which was conducted in an ethnically homogeneous community-based sample, larger ($N > 6000$ participants) than that of Yeo et al. (2011) or the three recent studies. Our participants were genotyped on the Illumina 660W Quad array, and we called CNVs with *PennCNV*. We report the results of mutational burden analyses, and—since we have reasonable power to detect a specific CNV of moderate-to-large effect—we supplement the burden analyses with genome-wide association scans of single CNVs.

Methods

GWAS Sample

Participants.

Both studies involved—the Sibling Interaction and Behavior Study (SIBS; McGue, Keyes, Sharma, Elkins, Legrand, et al., 2007) and the Minnesota Twin Family Study (MTFS; Iacono, Carlson, Taylor, Elkins, & McGue, 1999; Iacono & McGue, 2002; Keyes et al., 2009)—and the collection, genotyping, and analysis of DNA samples, were approved by the University of Minnesota Institutional Review Board's Human Subjects Committee. Written informed assent or consent was obtained from all participants; parents provided written consent for their minor children.

Our participants came from two longitudinal family studies conducted by the Minnesota Center for Twin and Family Research (MCTFR): SIBS and MTFS. The two age cohorts of the MTFS, the 11-year-old and 17-year-old cohorts, are named for the target ages of their constituent twins at the intake assessment. MTFS is a longitudinal study of a community-based sample of same-sex twins born between 1972 and 1994 in the State of Minnesota, and their parents. SIBS is an adoption study of sibling pairs and their parents; its community-based sample contains families where both siblings are adopted, where both are biologically related to the parents, or where one is adopted and one is biologically related. As required by SIBS inclusion criteria, any sibling in the sample who was adopted into the family will not be biologically related to his or her co-sibling, which has been genomically verified (Miller et al., 2012). For the purposes of our analysis, the sample comprises six distinct family-types:

1. Monozygotic- (MZ) twin families (*N* = 3590, in 1095 families),
2. Digyzotic- (DZ) twin families (*N* = 1898, in 618 families),
3. SIBS families with two adopted offspring (*N* = 268, in 214 families),
4. SIBS families with two biological offspring (*N* = 426, in 178 families),
5. "Mixed" SIBS families with 1 biological and 1 adopted offspring (*N* = 180, in 100 families),
6. Step-parents (*N* = 77).

Table 2-1 provides some descriptive characteristics of our sample. As explained below, our method of analysis accounted for the clustering of individual participants within families. However, family-type #6, step-parents, do not fit neatly into a four-member family unit; we treated them as independent observations in our analysis.

Genotyping.

Participants who provided DNA samples were typed on a genome-wide set of markers with the Illumina Human660W-Quad array. The array comprises 657,366 markers, including 95,876 intensity-only markers that were designed to increase coverage in between SNPs and to tag previously identified regions known to contain CNVs. Both DNA samples and markers were subject to thorough quality-control screens (Miller et al., 2012). The present analyses restricted the sample only to Caucasian participants of European ancestry (i.e., "White" participants), who were identified based upon self-reported ancestry as well as principal components from EIGENSTRAT (Price et al., 2006). After excluding DNA samples that failed quality-control screening, the GWAS sample of 7702 White participants was identified. Details concerning the studies (MTFS and SIBS), genotyping, quality-control, and ancestry determination can be found in Miller et al. (2012).

Phenotypic measurement.

Measurement of GCA was included in the design of the intake assessment for most participants, by way of an abbreviated form of the Wechsler Intelligence Scale for Children-Revised (WISC-R) or Wechsler Adult Intelligence Scale-Revised (WAIS-R), as age-appropriate (that is, 16 or younger, and older than 16, respectively). The short forms consisted of two Performance subtests (Block Design and Picture Arrangement) and Verbal subtests (Information and Vocabulary), the scaled scores on which were prorated to determine Full-Scale IQ (FSIQ). FSIQ estimates from this short form have been shown to correlate 0.94 with FSIQ from the complete test (Sattler, 1974). Parents in the SIBS sample are an exception, in that they were not tested with this short form of WAIS-R until the first SIBS follow-up assessment. By design, only one parent per SIBS family returned for this follow-up, which was usually the mother. As a result, IQ data for SIBS fathers is very limited in its availability.

IQ-testing was also included in the design of the second follow-up for both cohorts of MTFS twins, and for the fourth follow-up for the 11-year-old cohort. At these assessments, twins received a further abbreviated form of WAIS-R, consisting only of the Vocabulary and Block

Design subtests, the scaled scores on which were again prorated to determine FSIQ. Of the 2,914 twins entered into our analysis, 827 were tested twice, and 312 were tested three times. Multiple testing occasions were spaced approximately seven years apart. To achieve a more reliable assessment of the phenotype, we simply averaged all available measures of FSIQ for each participant, and used these single within-person averages in analysis. FSIQ among participants entered into analysis ranged from 59 to 151 (also see Table 2-1). Ten participants with FSIQ of 70 or below were included in analyses. Despite their low scores, these participants were not visibly impaired and were capable of completing the multifaceted MTFS/SIBS assessment during their visit. They are therefore unlikely to meet diagnostic criteria for mental retardation (*DSM-IV*), and instead, merely represent the low end of the normal-range distribution of GCA.

CNV calling.

We called CNVs using *PennCNV* (Wang et al., 2007), which implements a hidden Markov model to resolve CNVs from genome-wide normalized intensity data[5]. To briefly summarize, within a given participant, *PennCNV*'s algorithm treats each SNP's combined probe intensity (known as the log R ratio) and relative B-allele intensity as observed realizations of a random vector with distribution defined by the unknown ("hidden") copy-number state of the SNP. Conditional on some copy-number state, this vector's two elements are independent, and their marginal distributions have closed-form expression. Thus, the inferential problem is to determine which copy-number state in this mixture model has highest posterior probability. The copy-number state of a given SNP i only depends upon the states of other SNPs by way of the immediately adjacent SNP $i - 1$ (hence, the "Markov" in "hidden Markov model"). The probabilities of states at SNP i depend upon the state at SNP $i - 1$ and the distance separating the SNPs; state changes are more probable with greater distances. Copy-number variants within a

[5] Readers who require an accessible introduction to CNV detection with genome-wide SNP arrays are referred to Cooper and Mefford (2011).

participant are identified as runs of SNPs with the same mutant state. *PennCNV* further provides an index of the overall quality of a participant's DNA sample for CNV-calling purposes. This is the standard deviation (across SNPs, within-person) of the log-*R* ratio, corrected for guanine-cytosine (GC) content (hereinafter, LRRSD). Prior to calling CNVs, we re-clustered intensities from the Illumina array only using data from the 6,110 participants with raw (i.e., not GC-corrected) LRRSD below 0.30.

We only retained CNVs that spanned 20 or more markers and had confidence scores greater than 10.0, which, assuming 1:1 prior odds, corresponds to posterior odds in excess of 20,000 to 1 favoring the called copy-number state over the second-most-likely state. These are somewhat conservative thresholds intended to reduce risk of false-positive CNV calls[6]. We further excluded CNVs near the centromeres and telomeres, and those overlapping immunoglobulin genes. We made these two exclusions because CNVs identified in these regions have a high probability of being artifactual (Need et al., 2009), the latter because previous research has found that CNV calls in immunoglobulin regions are not representative of germline DNA and tend to depend substantially on DNA source—e.g., saliva versus blood (Need et al., 2009) or lymphoblastoid cell lines versus blood (Wang et al., 2007; Sebat et al., 2004). We did not filter CNVs by mutation frequency. To be included in the analysis, a participant had to have LRRSD below 0.30 (approximately 98th percentile; see Figure 2-1), and have provided a blood sample for DNA extraction. Finally, during CNV calling, 7 participants were identified with apparent somatic mosaicism for partial trisomy or monosomy of >80% of an arm of one chromosome. These were identified by plotting the standard deviation and the inter-quartile

[6] In a recent association study of CNVs and familial Parkinson disease (Pankratz et al., 2011), *PARK2*, a known true positive, evidenced a genome-wide significant signal only when analysis was restricted to *PennCNV* calls spanning 20 or more markers, and not when using a less stringent threshold nor when using different software (and therefore, not when using a "consensus-call" approach, which only retains CNVs called by more than one software algorithm). Furthermore, analysis of calls made with the conservative threshold did not yield genome-wide significant signal at two loci, implicated in other analyses, that ultimately failed molecular validation.

range of the B-allele frequency for each chromosomal arm of each participant and then visually inspecting the intensity data (see Figure 2-2). These participants were excluded from analysis.

We hypothesized that genome-wide mutational burden would negatively associate with FSIQ. To operationalize "mutational burden," we scored participants on eight different variables (hereinafter, the "CNV variables") from CNV calls:

1. Total (i.e., both deletions and duplications) CNV count,
2. Total CNV length,
3. Deletion count,
4. Deletion length,
5. Duplication count,
6. Duplication length,
7. Count of homozygous deletions,
8. Length of homozygous deletions.

Here, "count" refers to the number of distinct identified CNVs of the specified type, whereas "length" refers to the sum of lengths (in kilobases) of CNVs in the specified type. Table 2-2 presents descriptive statistics for the CNV variables, including zero-order correlations with FSIQ. Notably, the average participant in the sample harbored two homozygous deletions—the mean count was 2.2 (SD = 1.3), and the modal count was 2.

To conduct genome-wide association scans for specific CNVs, we first coded the copy-number states for the 6,439 participants, at markers across the 22 autosomes, under three coding schemes: code "0" for reference copy-number state, otherwise, code "1" for (a) deletion or duplication; (b) deletion only; (c) homozygous deletion. There were numerous runs on the genome-wide array across which no participant's copy-number state changed from one adjacent marker to the next. These runs were each collapsed into a single site, identified by the first marker in the run. "Monomorphic" loci, with no copy-number variation among participants, were dropped from analysis. Thus, the genome-wide scans were conducted across 17,262 sites counting deletions and/or duplications, 10,634 sites counting deletions only, and 499 sites counting homozygous deletions only.

Type I error correction.

We report nine burden analyses in all, one for each of the eight CNV variables, plus a base model with covariates only. Thus, we tested the significance of eight regression coefficients. A Bonferroni or Šidák correction of the per-comparison Type I error rate would be straightforward, but overly conservative, since the CNV variables are correlated with one another. We therefore applied the Cheverud-Nyholt correction (Nyholt, 2004), which is a Šidák correction for the "effective" number of independent statistical tests, m_{eff}, as calculated from the spectral decomposition of *m* variables' correlation matrix. Let $\lambda_1, \dots, \lambda_m$ denote the eigenvalues of that correlation matrix, ordered from greatest to least. Then,

$$m_{eff} = 1 + (m - 1)\left(1 - \frac{V_\lambda}{m}\right) \tag{2.1}$$

where

$$V_\lambda = \frac{1}{m - 1}\sum_{i=1}^{m}(\lambda_i - 1)^2 \tag{2.2}$$

For *m* = 8 CNV variables, we calculate a m_{eff} = 6.7604. At the conventional familywise Type I error rate of α_{FW} = 0.05, we therefore have a per-comparison Type I error rate (significance threshold) of $\alpha_{PC} = 1 - (1 - \alpha_{FW})^{1/m_{eff}} = 0.0076$.

We similarly applied this correction to the three genome-wide scans by calculating the correlation matrix for all sites under the relevant coding scheme on a given chromosome, and obtaining a m_{eff} for that chromosome. Summing the 22 m_{eff} values then provided the effective number of statistical tests for each scan, from which corrected values for α_{PC} were calculated; these were 3.03×10^{-6}, 4.95×10^{-6}, and 1.12×10^{-4}, counting deletions and/or duplications, deletions only, and homozygous deletions only, respectively.

Analyses

Statistical Power.

Because our participants are clustered within families, the effective number of independent observations (N_{eff}) in our sample was less than 6,439. We conducted two sets of power calculations in *Quanto* (Gauderman & Morrison, 2006), one that assumed 6,439 independent participants (surely an overestimate of our N_{eff}), and one that assumed 2,209 independent participants, which is equal to the number of families in the analysis, and represents a very conservative estimate of our N_{eff}. In the burden analyses, if N_{eff} = 6,439, we would have at least 80% power for a CNV variable that accounts for 0.2% or more of phenotypic variance, and if N_{eff} = 2,209, for a CNV variable that accounts for 0.56% of variance. In the genome-wide scans, if N_{eff} = 6,439, we would have at least 80% power to detect a mutation accounting for at least 0.48%, 0.46%, or 0.35% of variance, when respectively counting deletions and/or duplications, deletions only, or homozygous deletions only. If instead N_{eff} = 2,209, we would have at least 80% power to detect a mutation accounting for at least 1.37%, 1.32%, or 1% of variance, when respectively counting deletions and/or duplications, deletions only, or homozygous deletions only. For further details about the power of the genome-wide scans, see Figures 2-3 through 2-5.

Mutational Burden.

A total of 6439 Caucasian participants, from 2209 families, provided valid data both for FSIQ and CNVs, and were entered into analysis. Since our participants are clustered within families, observations were not sampled independently of one another. Moreover, the expected covariance structure among family members depends upon family type. We therefore conducted our multiple-regression analysis with *RFGLS*, a package for the R statistical computing environment, which conducts feasible generalized least-squares (FGLS) regression in datasets with complicated family structures (Li, Basu, Miller, Iacono, & McGue, 2011).

Our analyses included four covariates: sex, birth year[7], and the first-two principal components from EIGENSTRAT (Price et al., 2006), to control for population stratification. The correlations among these covariates were uniformly small, in no case exceeding 0.035 in magnitude. Similarly, the correlations among the covariates and the CNV variables were small, the largest in magnitude being that between duplication count and birth year (r = 0.062). Since these analyses simultaneously estimated the fixed-effects' regression coefficients and the residual covariance matrix, the model contained 18 free parameters altogether: the regression intercept, the regression coefficient for the CNV variable, regression coefficients for the four covariates, four residual variance parameters (one each for offspring, mothers, fathers, and step-parents), and residual correlation parameters for eight relationships between family members (MZ twin, DZ twin and full sibling, adoptive siblings, spouses, biological mother-offspring, biological father-offspring, adoptive mother-offspring, adoptive father-offspring). We corroborated our burden analyses by fitting an equivalent model in *Mx* (Neale, Boker, Xie, & Maes, 2003), the results of which lead to the same overall conclusions.

Genome-wide association scans.

We conducted the association scans with *RFGLS*, using its "rapid-FGLS" approximation, which works as follows. We first calculated the residual covariance matrix from a FGLS regression of FSIQ onto the covariates only (sex, birth year, and the two principal components to control for population stratification). Then, we "plugged in" this matrix for use in all subsequent

[7] We covaried out birth year, rather than age, for three reasons. First, IQ tests are age-normed to begin with. Second, the FSIQ scores of a minority of our twins are within-person averages of FSIQs from more than one testing occasion. In a sense, these twins have multiple ages at testing, but of course, they have only one birth year. Third, the nuisance confound for which we want to correct is the Flynn Effect (first reported by Flynn, 1984, 1987), which is the secular trend of increasing IQ with each generation, and is directly related to birth year, and not to age *per se*. Surprisingly, our data seem to contradict the Flynn Effect at a glance. In the covariates-only FGLS regression, the estimated coefficient for birth year was -0.089 (*SE* = 0.010), indicating that later birth year predicted lower IQ. The zero-order Pearson correlation of birth year and FSIQ was in the same direction, r = -0.097 (bootstrap *SE* = 0.012).

single-site FGLS regressions of FSIQ onto copy-number state at the site, with covariates. This approximation only requires that a single residual covariance matrix be calculated, which greatly reduces computation time and produces negligible bias in the resulting *p*-values, as long as no site accounts for more than 1% of phenotypic variance (Li et al., 2011). We resolved to follow-up any suggestive signals from the scans with a slower but more precise FGLS regression analysis of the implicated site that would simultaneously estimate the regression coefficients and residual covariance matrix, as was done in the burden analyses.

Results

Table 2-2 presents the results of our *RFGLS* regression analyses. The rightmost column provides the increase, relative to the covariates-only model, in Nagelkerke's (1991) generalized R^2. In the present context, it is interpretable as the proportion of variance in FSIQ attributable to the CNV variable and covariates. The covariates alone accounted for approximately 2.2% of the phenotypic variance.

None of the eight CNV variables we analyzed showed even nominally significant association with FSIQ. The most suggestive result was for deletion count (p = 0.057), which was positively associated with FSIQ, and therefore not in the hypothesized direction. Its coefficient ($\hat{b} \approx 0.04$) suggests that, holding covariates constant, one would predict a 1-point increase in IQ for approximately every 25 copy-number deletions that a person harbors. Even when the CNV set was restricted to homozygous deletions, the regression coefficients, though in the expected direction, did not differ significantly from zero. Although the magnitude of the coefficient for count of homozygous deletions stands out, predicting a 1-point decrease in IQ for about every 5 such deletions, its wide confidence interval shows that it is estimated with little statistical precision. All CNV variables provided a negligible increase in the proportion of phenotypic variance explainable by the regression model.

Further, we obtained no evidence that any specific CNV is associated with FSIQ (see Figures 2-6 through 2-8). In the genome-wide association scan counting homozygous deletions, the only p-value more extreme than 0.001 was p = 4.60×10^{-4} ($\hat{b}$ = 31.39, SE = 8.96, for a mutation on chromosome 14 starting at rs1950943, for which only two participants were homozygous), which would not be considered significant with correction for multiple testing, and contrary to hypothesis, is in the positive direction. The follow-up analysis for this mutation, which simultaneously estimated the regression coefficients and residual covariance matrix, yielded p = 4.81×10^{-4} ($\hat{b}$ = 31.33, SE = 8.97) [8]. No site in either of the two other scans yielded a p-value more extreme than 0.0001, let alone approaching the adjusted α_{PC}.

Discussion

The present study obtained no evidence for our hypothesis that copy-number variants contribute to the heritable variance of GCA. Contrary to hypothesis, mutational burden was not negatively associated with FSIQ, and genome-wide scans failed to identify any CNV significantly related to the phenotype. Thus, we did not replicate the mutational-burden results of Yeo et al. (2011). Instead, our results were much like those of recent studies (MacLeod et al., 2012; Bagshaw et al., 2013; McRae et al., 2013), all of which had larger samples that Yeo et al. (2011). Both our data and those of these recent studies suggest it is unlikely that CNVs will provide the missing piece in the puzzle of what has become known as "missing heritability" (Maher, 2008) for this particular quantitative trait.

In particular, the results of our aggregate mutational burden analyses with CNVs appear especially unimpressive compared to recently developed methods that leverage genome-wide

[8] We also conducted an exact permutation test of the null hypothesis that the group-mean difference between mutant-state individuals and reference-state individuals in FSIQ, once covariates are partialled out, is zero. This entailed computing a group difference in mean residuallized FSIQ, using each of $C(6439{,}2)$ = 20,727,141 possible mutant-state groups that could be generated from the dataset, and comparing the absolute value of each difference to that observed in the actual data. The proportion that equal or exceed the observed difference, 196,486 ÷ 20,727,141 ≈ 0.009, is a two-tailed p-value not exceeding the corrected Type I error rate for the homozygous-deletion scan.

SNP data by predicting the phenotype from the combined effect of multiple SNPs at once. One such approach is to select SNPs that manifest suggestive signal in a GWAS, and weight participants' reference-allele counts for each selected SNP by its GWAS regression coefficient. Each participant's sum of weighted allele counts constitutes his/her "polygenic score." Polygenic scores are then employed as predictors of the phenotype, and cross-validated in (ideally) an independent sample not used in the GWAS. This genetic scoring approach has been used to predict, under cross-validation, as much as 3% of disease risk for schizophrenia (International Schizophrenia Genetics Consortium, 2009) and as much as 16.8% of the variance in height (Lango Allen et al., 2010). More impressive results can be achieved by a method of analysis that treats the influence of every SNP in the GWAS as a latent random effect. Some (e.g., Benjamin et al., 2012) refer to this method as GREML, for "genomic-relatedness-matrix restricted maximum-likelihood," though it is widely identified with the software written to implement it, *GCTA* (Genome-wide Complex Trait Analysis; Yang, Lee, Goddard, & Visscher, 2011). Very briefly summarized, it is accomplished by calculating a genetic relationship matrix between classically unrelated individuals and performing best linear unbiased prediction of the SNP variance component in a mixed linear model fit by restricted maximum likelihood. Via GREML, common SNPs can account for 22% to 51% of the variance in intelligence (Davies et al., 2011; Benyamin et al., 2013) and 45% of the variance in height (Yang et al., 2010). Biometric heritability estimates for adult IQ, in the normal range of variation in Western countries, typically range from 50% to 70% (Bouchard & McGue, 2003; Deary, Johnson, & Houlihan, 2009), somewhat above these GREML estimates. However, it appears from our results that CNVs—a different class of genetic polymorphism from SNPs—show little potential to help "fill in the gap" between classical heritability estimates and the somewhat lower GREML variance components.

On the other hand, while no SNP has been reliably identified as associated with GCA, researchers have identified specific CNVs associated with syndromal intellectual disability and other neurodevelopmental pathology (Mefford et al., 2012). This elicits the hypothesis that there might exist specific CNVs contributing to normal-range variation in cognitive ability. But, our three genome-wide association scans, each using a different coding scheme for copy-number state, failed to discover any association signal exceeding our precisely adjusted significance threshold, and provide no support for the hypothesis.

However, we readily acknowledge that these conclusions are mitigated by several limitations of our study. First, our sample size of over 6,000 participants may have afforded insufficient power to reliably detect true CNV effects of quite small magnitude. For the sake of comparison, the GIANT Consortium (Lango Allen et al., 2010) required about 180,000 participants—a sample size of truly Brobdingnagian proportions—to confirm 180 SNPs associated with stature. We are also bound by the inherent limitations of our Illumina 660W-Quad array. A SNP array with denser coverage of the genome would allow for higher-fidelity CNV calls. Despite the very modest sample size in Yeo et al.'s (2011) study, one of its strengths, as both MacLeod et al. (2012) and McRae et al. (2013) point out, is their use of the Illumina Human 1M Duo BeadChip array. This chip allowed them to genotype their participants on over 1 million SNPs and achieve denser coverage of the genome than did we or any of the three recent studies.

Another limitation is that we may not have called small, low-baserate CNVs. To reduce the risk of spurious calls, we have taken a conservative approach: we only retained CNVs with a confidence score greater than 10.0 and spanning 20 or more markers, and dropped CNVs occurring in artifact-prone regions of the genome. McRae et al. (2013) also imposed a minimum confidence score and CNV length, and remark that, based on manual inspection of data, those minima served to reduce false-positives and false-*negatives*, respectively. A previous study by Pankratz et al. (2011) reported that analysis of *PennCNV* calls subjected to similar filters had highest power to detect a known true positive, and had the lowest rate of significant association signals for loci that subsequently failed molecular validation. Importantly, if Pankratz et al. had only retained those CNVs called by more than one software algorithm, they would have missed the signal for *PARK2*. Guided by their experience, we elected not to use "consensus calls," which would entail retaining only those CNVs in the intersection of the set of *PennCNV* calls and the set(s) of calls from other software. However, we restricted analysis to participants who had provided adequate-quality DNA from a blood sample, and who passed a screen for somatic mosaicism. We regard our filters on CNV calls and DNA samples as adequate safeguards against spurious CNV calls. Nonetheless, they commensurately diminish the sensitivity of CNV detection for short and/or rare variants. Therefore, from our data, we cannot rule out the possibility of such CNVs of large effect. Such mutations are difficult to call reliably, and attempting to discover them would require acceptance of a higher error rate. If only rare or small-effect CNVs contribute to variation in GCA, further studies will require still larger samples to make headway along this line of research.

It is rather remarkable that the typical participant in our normal-range-IQ, mostly community-based sample has two homozygous deletions in their genome, many of which overlap genes. But, for the sake of comparison, next-generation sequencing has revealed that the genome of James D. Watson, an undeniably high-GCA individual who co-discovered the structure of DNA, contains 23 (9 duplications, 14 deletions) large (26 kb to 1.6 Mb) CNVs altogether encompassing 34 genes (Wheeler et al., 2008). Because GCA was likely under selection pressure in humans' ancestral environment of adaptation (Jensen, 1998), one would anticipate that most mutations, particularly those with the potential to cause loss-of-function, would detrimentally affect the trait (Gangestad, 2010). But, the present study—which utilized a large, ethnically homogeneous sample, assessed on an individually administered IQ test—found no evidence that greater mutational burden, however operationalized, corresponds to reduced GCA. In fact, the most suggestive signals from the burden analyses and the genome-wide scans were for deletions positively associated with FSIQ. Although CNVs can cause mental retardation and other forms of neurodevelopmental psychopathology, normal-range variation in general cognitive ability is “highly heritable and polygenic” (Davies et al., 2011), and its genetic architecture may also be more redundant and more robust to mutational disruption than previously guessed.

Table 2-1. Descriptive characteristics of Study #2 sample.

	Parents	Twins (17yo)	Twins (11yo)	Non-twin Biological Offspring	Adoptees	Step-parents
N	2985	1066	1848	366	97	77
Female(%)	60.7%	55.5%	50.8%	49.7%	44.3%	9.1%
Mean Age at Intake (SD)	43.29 (5.51)	17.5 (0.45)	11.8 (0.43)	14.9 (1.86)	15.2 (2.17)	40.5 (7.45)
Mean FSIQ (SD)	106.0 (14.2)	100.3 (14.1)	103.8 (13.6)	108.9 (13.0)	105.7 (14.3)	103.0 (15.7)

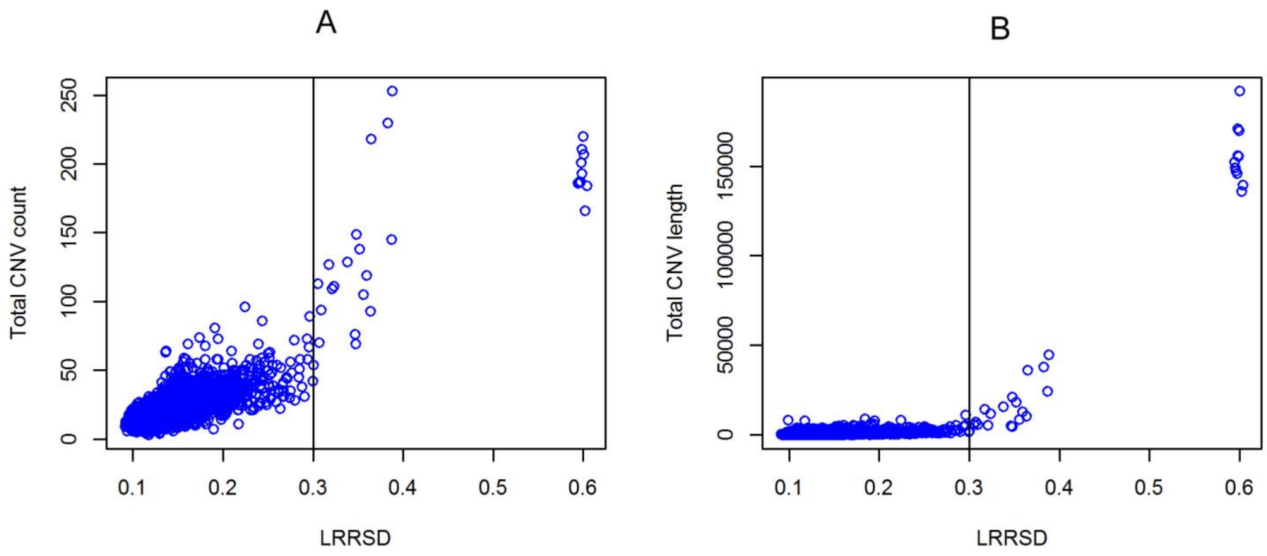
Table notes: Total *N* = 6439, in 2209 families. FSIQ = Full-Scale IQ; 17yo = 17-year-old cohort; 11yo = 11-year-old cohort. For a minority of twins (39%), FSIQ represents a within-person average of FSIQ scores from more than one assessment (see text). FSIQ range: 151 – 59 = 92. Parental intake age range: 65 – 28 = 37. Offspring intake age range: 20 – 11 = 9.

Table 2-2. Descriptive statistics for CNV variables and results of mutational burden analyses.

CNV Variable	Mean (SD)	Min – Max	Correlation with FSIQ	Regression Coefficient	95% CI	P-value	ΔR^2
Count, All	19.3 (8.0)	3 – 96	0.008	0.0326	-0.0078, 0.0731	0.114	0.0009
Length, All	722.4 (565.1)	32 – 8992	-0.010	-1.568×10^{-5}	-0.0006, 0.0006	0.959	2.622×10^{-5}
Count, Deletions	14.1 (8.2)	1 – 68	0.007	0.0389	-0.0012, 0.0789	0.057	0.0012
Length, Deletions	430.1 (420.0)	5 – 8746	-0.012	0.0002	-0.0006, 0.0010	0.567	0.0006
Count, Duplications	5.2 (4.7)	0 – 62	0.001	-0.0208	-0.0928, 0.0512	0.572	0.0001
Length, Duplications	292.3 (398.1)	0 – 7263	-0.001	-0.0003	-0.0012, 0.0005	0.453	0.0001
Count, Homozygous Deletions	2.2 (1.3)	0 – 8	-0.032	-0.2362	-0.5053, 0.0330	0.085	0.0011
Length, Homozygous Deletions	26.4 (27.7)	0 – 210	-0.026	-0.0105	-0.0227, 0.0017	0.092	0.0010

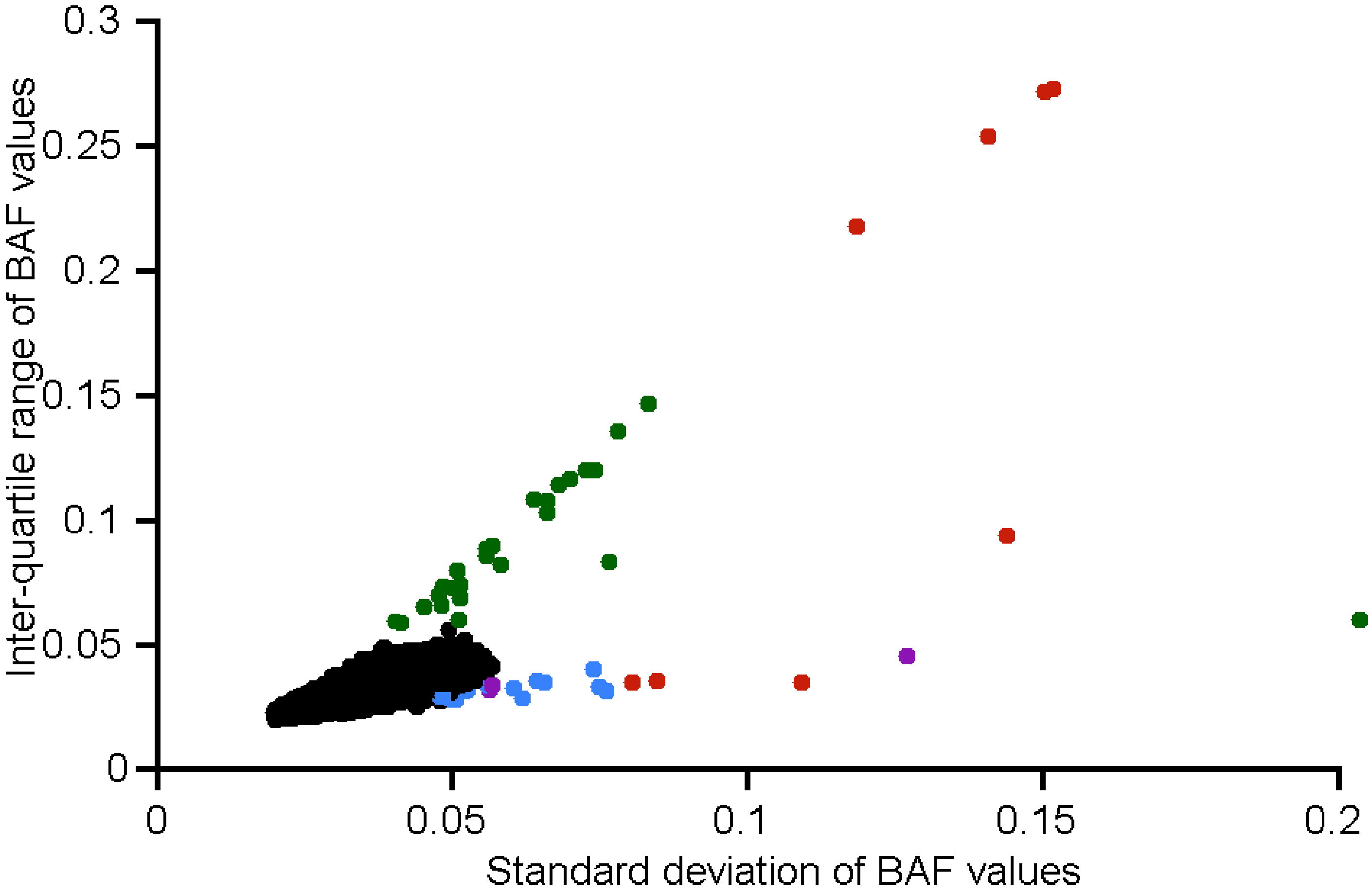
Table notes: "Count" refers to number of identified CNVs of the given type; "length" refers to their combined size in kilobases. Correlations are zero-order correlations with Full-Scale IQ. Regression covariates are sex, birth year, and first two principal components from EIGENSTRAT (Price et al., 2006). Confidence intervals and p-values are from Student's t distribution on 6433df. The rightmost column reports increase in Nagelkerke's (1991) R^2 over a covariates-only model, for which $R^2 = 0.0224$.

Figure 2-1. Participants' total CNV count (A) and length (B), as function of LRRSD.



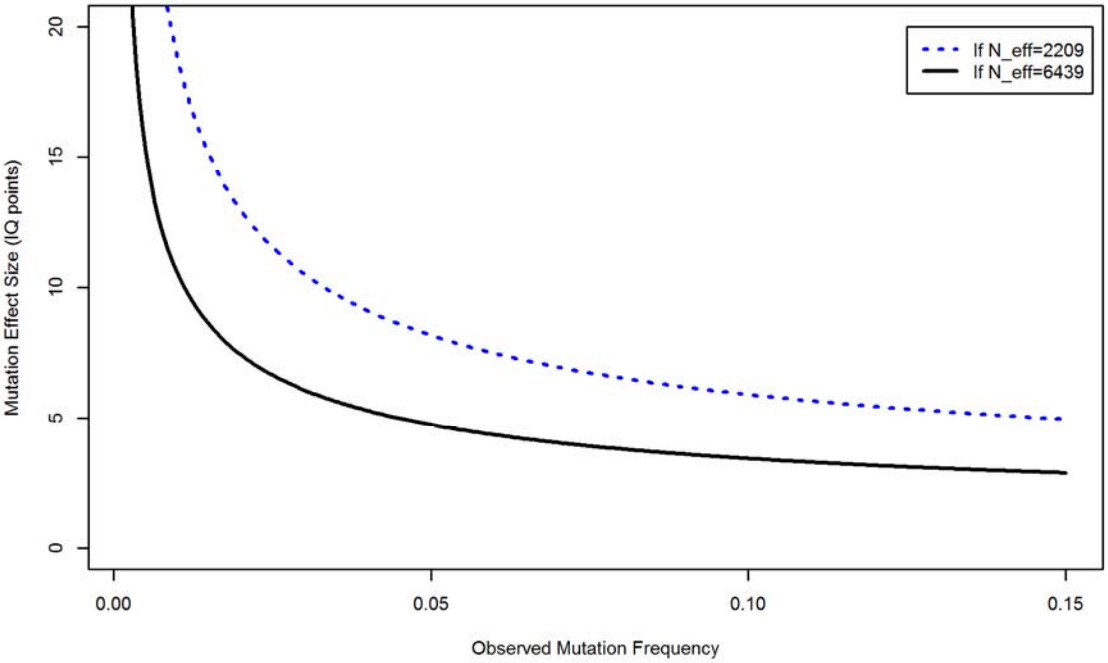
The mean level and degree of variation for both of these CNV variables become markedly higher above LRRSD = 0.30 (vertical lines). Therefore, participants with LRRSD greater than 0.30 were not included in analyses. Participants excluded from analyses due to any other filters, such as mosaicism or providing a non-blood sample, are not represented in these figures.

Figure 2-2. Identification of mosaic chromosomal arms using B-Allele Frequency Distributions.



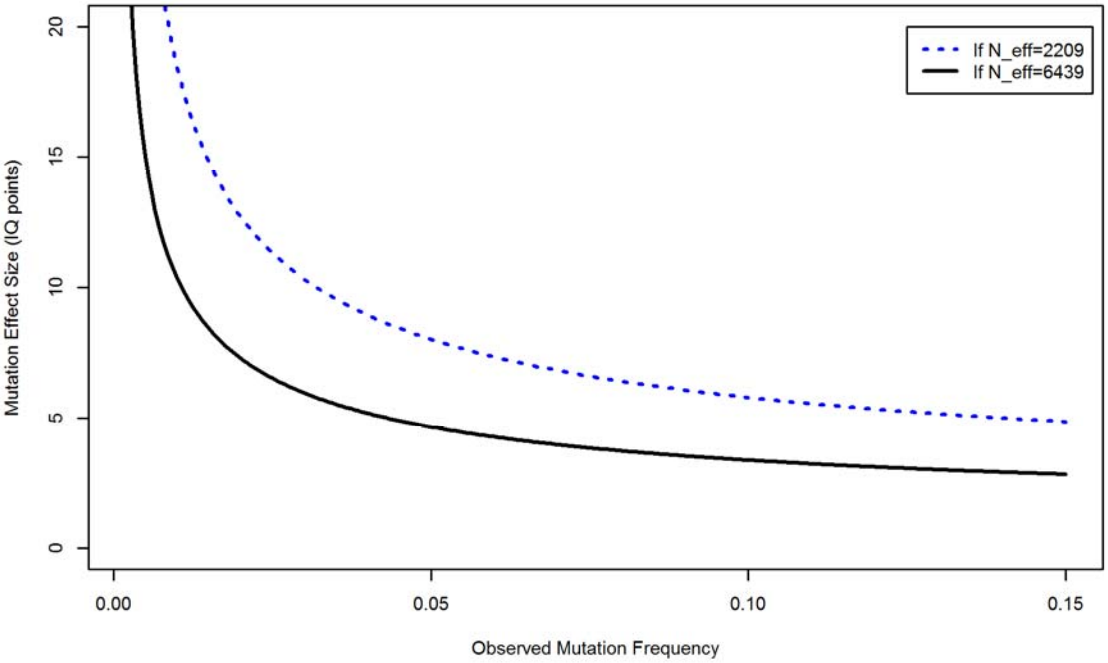
Red = contains severe mosaicism with lots of CNV calls in the region and were therefore excluded from analyses (*n*=7 samples, 8 chromosomal arms); Green = contains mild mosaicism, but not severe enough to lead to an intensity change called by *PennCNV* or to be excluded from analyses; Purple = contains large tracts of homozygosity that are biasing the BAF statistics; Blue = contains large duplications (0.5-6 Mb in size); Black = normal chromosomal arm BAF distributions.

Figure 2-3. Deletion/Duplication Effect Size Detectable with 80% Power ($\alpha = 3.03 \times 10^{-6}$), as Function of Observed Mutation Frequency



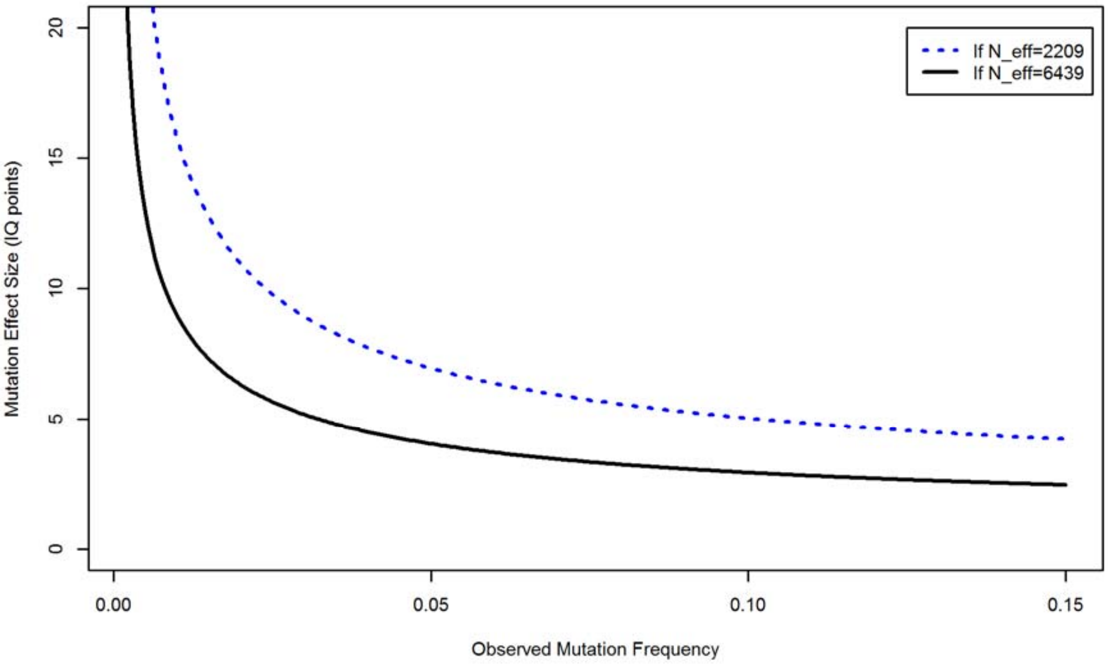
The solid curve assumes a sample of 6,439 independent participants, whereas the dotted curve assumes 2,209 independent participants. "Mutation effect size" refers to mean difference in FSIQ between the reference-state and mutant-state groups. Each point was plotted by first calculating what the sizes of the reference-state and mutant-state groups would be, given the N_{eff} and the proportion of participants having the mutant copy-number state (the "observed mutation frequency"). Then, the mutation effect size (the ordinate position of the point) was calculated from the power function for a two-tailed, independent-samples *t*-test , conditional on 80% power, a Type I error rate of 3.03×10^{-6}, and the two group sizes. The lowest plotted mutation frequency is 0.002, corresponding to about 4 of 2,209 participants harboring the mutation, and to a mutation effect size outside the bounds of the plot (25.90 if N_{eff} = 6,439, and 59.28 if N_{eff} = 2,209). At the other extreme, given a mutation frequency of 0.50, we would have power to detect an effect of 3.53 if N_{eff} = 2,209, or 2.06 if N_{eff} = 6,439.

Figure 2-4. Deletion Effect Size Detectable with 80% Power ($\alpha = 4.95 \times 10^{-6}$), as Function of Observed Mutation Frequency



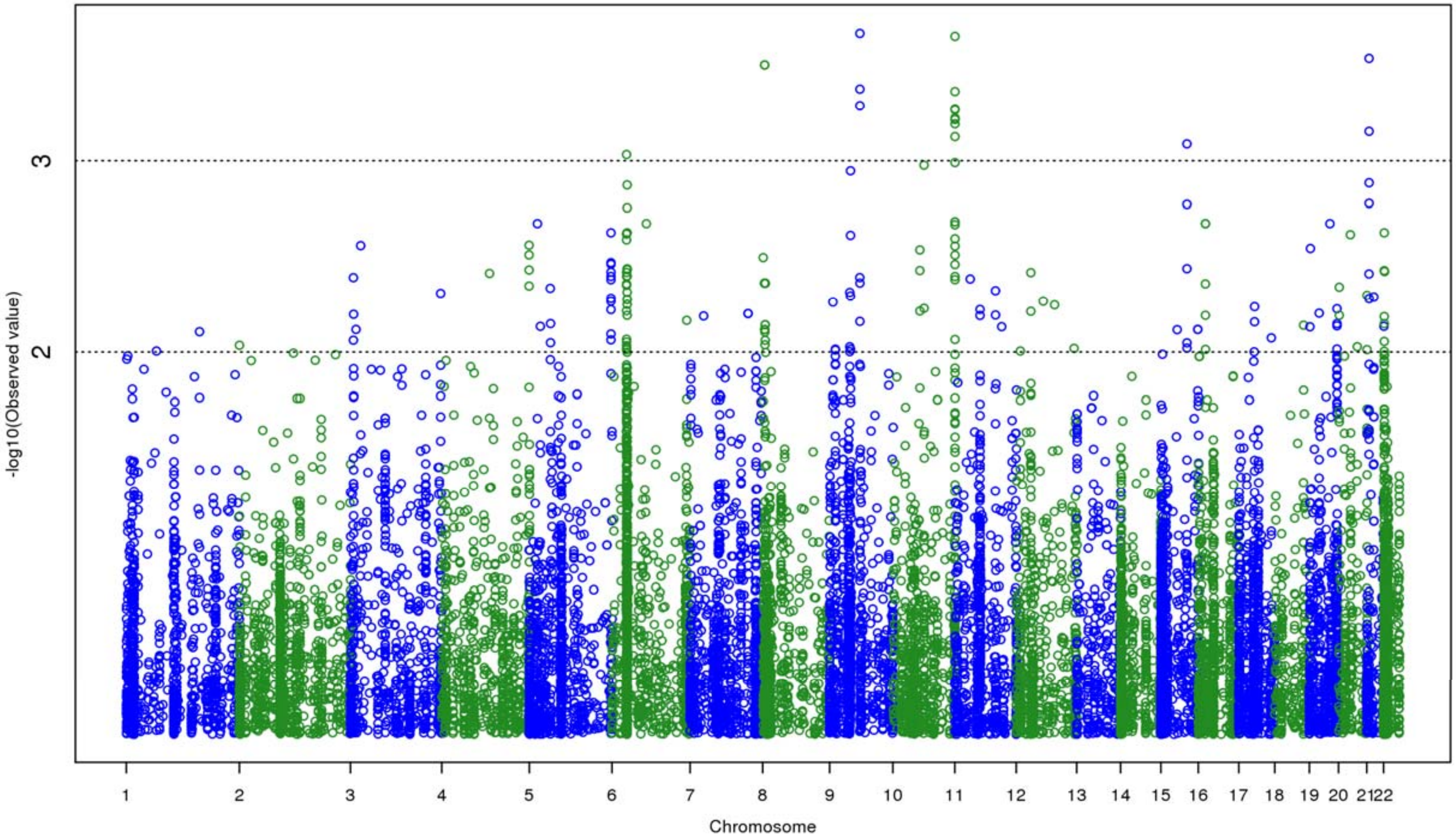
The solid curve assumes a sample of 6,439 independent participants, whereas the dotted curve assumes 2,209 independent participants. "Mutation effect size" refers to mean difference in FSIQ between the reference-state and mutant-state groups. Each point was plotted by first calculating what the sizes of the reference-state and mutant-state groups would be, given the N_{eff} and the proportion of participants having the mutant copy-number state (the "observed mutation frequency"). Then, the mutation effect size (the ordinate position of the point) was calculated from the power function for a two-tailed, independent-samples *t*-test , conditional on 80% power, a Type I error rate of 4.95×10^{-6}, and the two group sizes. The lowest plotted mutation frequency is 0.002, corresponding to about 4 of 2,209 participants harboring the mutation, and to a mutation effect size outside the bounds of the plot (25.29 if N_{eff} = 6,439, and 57.06 if N_{eff} = 2,209). At the other extreme, given a mutation frequency of 0.50, we would have power to detect an effect of 3.46 if N_{eff} = 2,209, or 2.02 if N_{eff} = 6,439.

Figure 2-5. Homozygous Deletion Effect Size Detectable with 80% Power ($\alpha = 1.12 \times 10^{-4}$), As Function of Observed Mutation Frequency



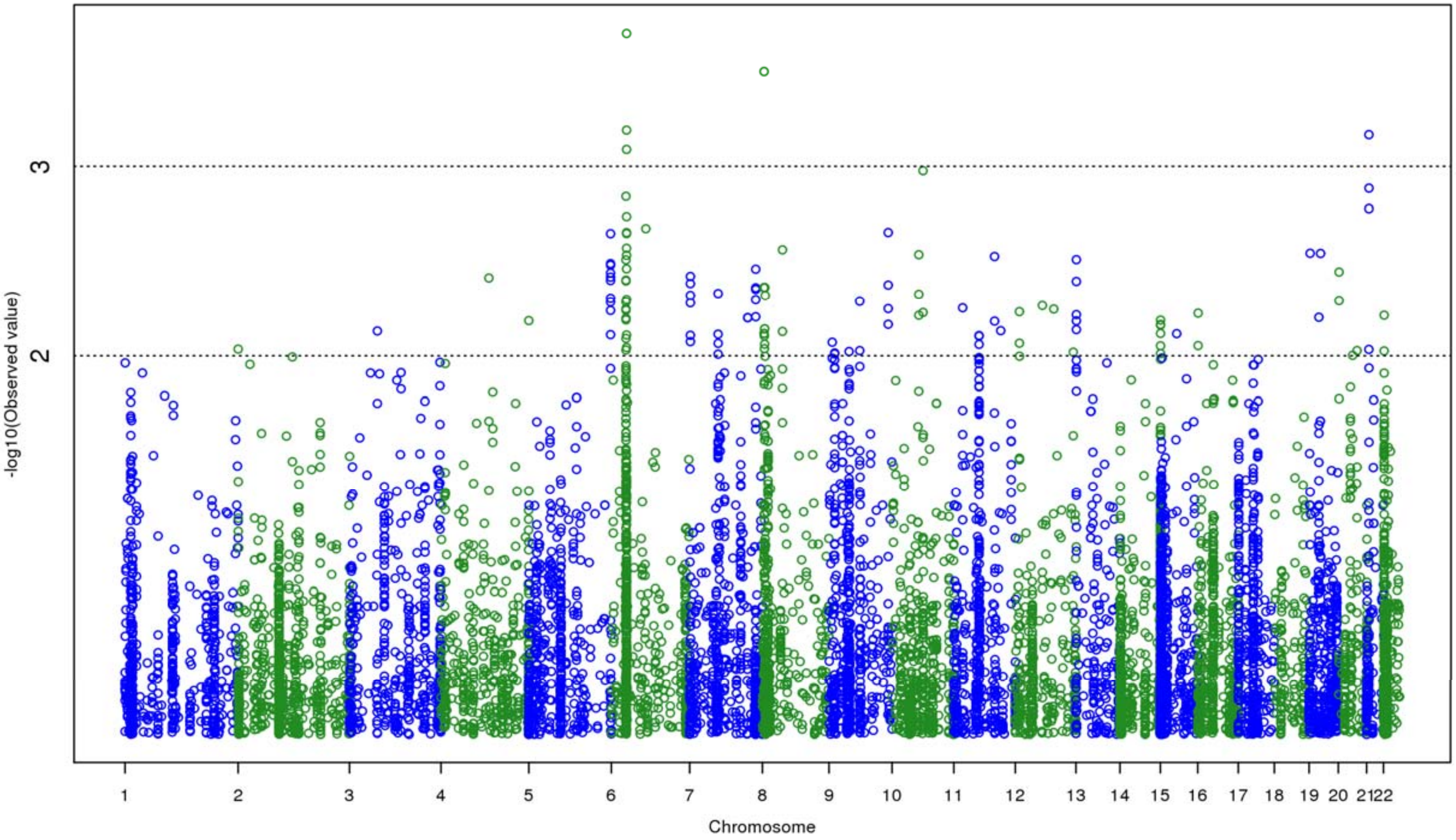
The solid curve assumes a sample of 6,439 independent participants, whereas the dotted curve assumes 2,209 independent participants. “Mutation effect size” refers to mean difference in FSIQ between the reference-state and mutant-state groups. Each point was plotted by first calculating what the sizes of the reference-state and mutant-state groups would be, given the N_{eff} and the proportion of participants having the mutant copy-number state (the “observed mutation frequency”). Then, the mutation effect size (the ordinate position of the point) was calculated from the power function for a two-tailed, independent-samples *t*-test , conditional on 80% power, a Type I error rate of 1.12×10^{-4}, and the two group sizes. The lowest plotted mutation frequency is 0.002, corresponding to about 4 of 2,209 participants harboring the mutation, and to a mutation effect size outside the bounds of the plot (21.30 if N_{eff} = 6,439, and 44.04 if N_{eff} = 2,209). At the other extreme, given a mutation frequency of 0.50, we would have power to detect an effect of 3.01 if N_{eff} = 2,209, or 1.76 if N_{eff} = 6,439.

Figure 2-6. Manhattan Plot for Genome-Wide Scan of Deletions and Duplications.



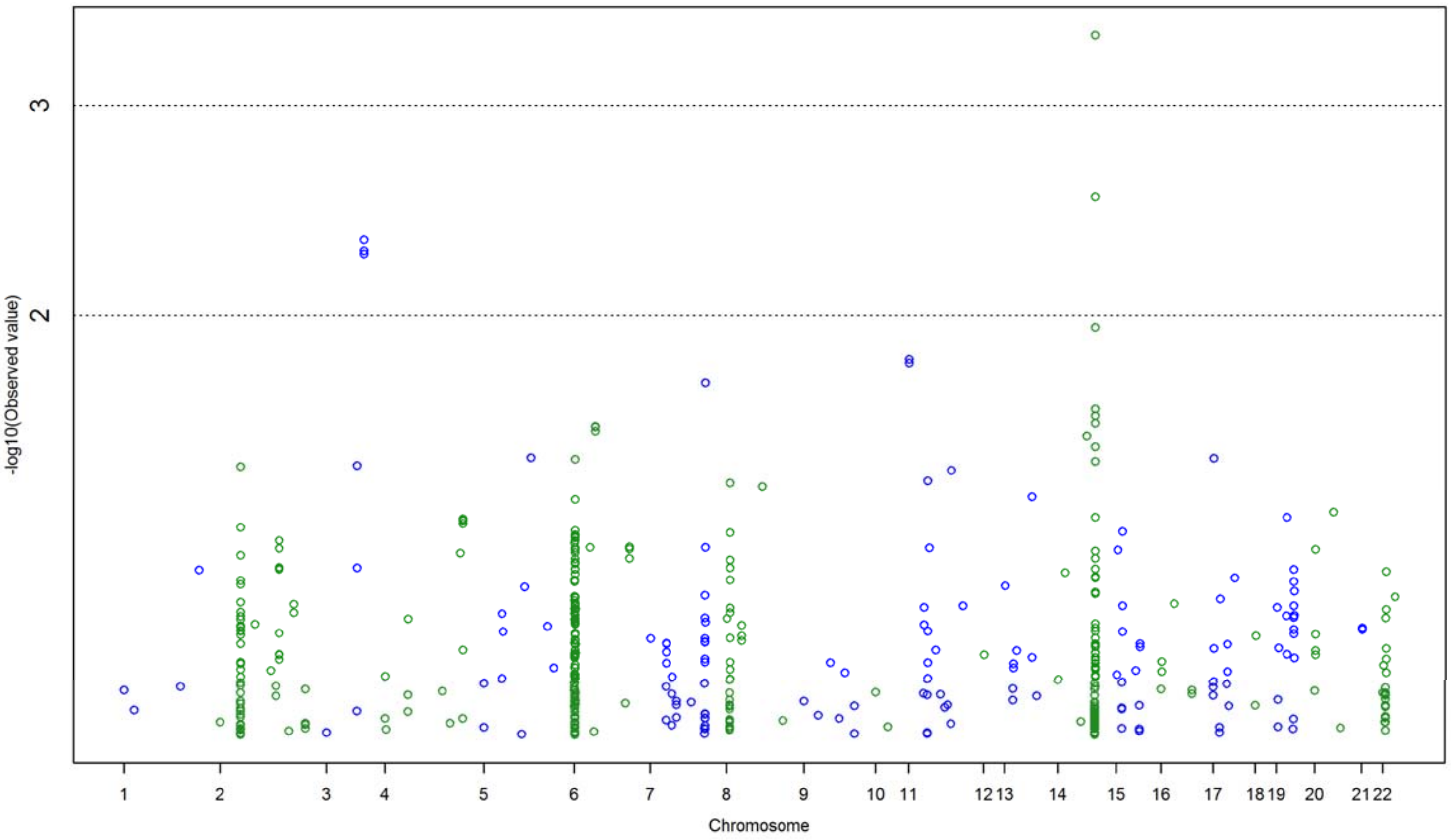
The threshold for genome-wide significance is about 5.52, outside the bounds of the plot.

Figure 2-7. Manhattan Plot for Genome-Wide Scan of Deletions.



The threshold for genome-wide significance is about 5.31, outside the bounds of the plot.

Figure 2-8. Manhattan Plot for Genome-Wide Scan of Homozygous Deletions.



The threshold for genome-wide significance is about 3.95. To ensure that all chromosome numbers appear on the *x*-axis of this figure, a gap of 50Mb has been added to separate the end of each chromosome from the start of the next. The most extreme point depicted here is $-\log_{10}(p) \approx 3.34$ for a deletion starting at rs1950943 on chromosome 14, for which only two participants were homozygous.

Study #3

Background

Biometric Modeling in Behavior Genetics

For decades, to better understand phenotypes of interest, behavior geneticists have applied biometric modeling to data from family, twin, and adoption studies. We use the term "biometric modeling" herein to broadly refer to the analysis-of-variance that decomposes phenotypic variation into different components attributable to heredity and environment (Fisher, 1918; Jinks & Fulker, 1970). Perhaps the simplest example is the well-known Falconer model, applied to the classical twin study. From the intraclass correlations for MZ and DZ twins reared together, one can calculate an estimate of (additive or narrow-sense) heritability a^2, shared-environmentality c^2, and unshared-environmentality e^2, which are respectively the proportions of phenotypic variance attributable to additive-genetic factors, between-family environmental factors, and within-family environmental factors. These quantities are standardized variance components, because $a^2 + c^2 + e^2 = 1$; their unstandardized (or "raw") counterparts—V_A, V_C, and V_E, respectively—together sum to the total phenotypic variance. When the Falconer model is applied to the powerful design of twins reared apart, the MZ-twin correlation provides a direct estimate of broad-sense heritabilityh_B^2. Broad-sense heritability h_B^2 differs from additive (narrow-sense) heritability a^2 because it is the proportion of phenotypic variance attributable to *all* hereditary factors, not merely those acting in an additive manner. However, since an estimate of a^2 may be regarded as a lower-bound estimate of h_B^2, in practice the distinction is not often emphasized, and discussion often proceeds in terms of one heritability parameter, h^2. This approximation is especially apt if non-additive genetic variance appears to be negligible.

Nowadays, biometric modeling is often implemented in a structural-equation-modeling framework, in which the raw variance components are postulated to reflect the effects of latent variables and can be estimated via numerical maximum-likelihood methods (Neale & Maes, 2004; Neale, 2009). Figure 3-1 depicts a path diagram of a simple biometric model, the monophenotype ACE model in the classical twin study (twins reared together). It is assumed here that all latent and observable variables have expectation zero, and that the latent variables further have unit variance. The observable variables are the paired phenotype scores, Y_1 and Y_2, for twin #1 and #2 in a pair, respectively. The latent variables A_1 and A_2 refer to the additive-genetic endowment of twins #1 and #2, respectively. The covariance between A_1 and A_2, r_A, is 1 for MZ pairs, and is fixed at its expected value of 0.5 for DZ pairs. The latent variable C represents the shared, or common, environment—the sources of environmental variance that contribute to between-pair variance but not within-pair variance. To simplify the diagram, we will represent it as a single variable; generally, it could be represented as C_1 and C_2, having covariance 1 or 0 depending on whether or not the twins were reared together. Finally, E_1 and E_2 are the unshared environments of twins #1 and #2, respectively. These represent environmental sources of variance that contribute to within-pair variance (including random measurement error of the phenotype). By tracing the paths in the diagram, the expected variances and covariances may be derived. The structural-equation-modeling approach to biometric analysis may be extended to larger pedigrees and/or kinships other than twins, and to multiple phenotypes (Neale & Maes, 2004; Posthuma, 2009).

Even in the monophenotype case, the model depicted in Figure 3-1 is likely to be an oversimplification. A more complete model would treat phenotypic variance as the sum of independent genetic and environmental components, plus component(s) attributable to some function of *both* heredity and environment (Jinks & Fulker, 1970). The possible forms of this function include *gene-environment correlation* (symbolized r_{GE}) and *gene-environment interaction* (symbolized $G \times E$) (Plomin, DeFries, & Loehlin, 1977). Generally speaking, r_{GE}

occurs when trait-relevant genetic and environmental factors tend to systematically coincide in the population. It comes in three flavors (Plomin et al., 1977; Scarr & McCartney, 1983). Passive r_{GE} occurs when the trait-relevant characteristics of individuals' genetic endowments and rearing environments correlate. Reactive (or evocative) r_{GE} occurs when individuals' partly-heritable behaviors elicit phenotype-relevant inputs from their environments. Active r_{GE} occurs when individuals' volitional selection of environments is partly under genetic influence. For instance, physically aggressive parents might transmit a genetic predisposition to aggression to their children, and also serve as role-models of aggressive behavior (passive r_{GE}), an aggressive child may elicit harsh physical discipline from parents and provoke fights with peers (reactive r_{GE}), and an individual with aggressive tendencies may seek out environments congenial to this tendency—say, by taking up boxing (active r_{GE}).

Figure 3-1 does not include double-headed arrows connecting the latent A factors with any environmental factors, which would constitute r_{GE}. Nor does Figure 3-1 allow for the magnitude of the A factor loadings to depend upon the value of another factor, which would constitute $G \times E$. But of course, such embellishments would make the model unidentified with only monophenotype twin data. Therefore, one approach to researching $G \times E$ is to add more observable variables, by measuring specific environmental and/or genetic variables hypothesized to be phenotype-relevant. In molecular genetics, one can assay the actual genotype of putative trait-relevant polymorphisms, and estimate the extent to which genotype-phenotype association depends upon some environmental variable(s). We will not discuss this approach further herein. Another approach is to estimate how much the magnitude of the genetic variance depends upon some observed environmental variable(s). The present work is concerned with an extension of this approach: estimate how much the magnitudes of *all* biometric variance components depend upon environmental variable(s). We will use "biometric moderation" to refer to the phenomenon that the biometric decomposition of a phenotype varies as a function of some observable variable, the "biometric moderator." We will specifically be concerned with biometric moderation in general cognitive ability (the phenotype) by family-of-origin socioeconomic status (the moderator). When we refer to a moderation *effect*, we mean "effect" in the statistical sense, and not necessarily in the causal sense.

Biometric Modeling of General Cognitive Ability

General cognitive ability (GCA) is that ability which is tapped by all cognitively demanding tasks. Often identified with Spearman's (1904) *g*, it can be operationalized as a composite score from a battery of tests that adequately samples the domain of cognitive tasks and specific abilities—for example, Full-Scale IQ (FSIQ) from an individually administered IQ test. Decades of research (to say the least—see Galton, 1869) have made clear that general cognitive ability is substantially heritable trait. Estimates of its heritability typically range from 0.50 to 0.70 (Bouchard & McGue, 1981, 2003; Deary, Spinath, & Bates, 2006), and are sometimes as high as ~0.80 (Rijsdijk, Vernon, & Boomsma, 2002).

As we described above, an important principle in contemporary behavior-genetic research is that the magnitude of a biometric variance component may depend upon other variables (moderators). The heterogeneity of heritability estimates for GCA may reflect the influence of such moderators. The role of one of them, age, has been well replicated: the general trend is that, from early childhood through late adolescence or early adulthood, IQ's heritability increases while its shared-environmentality decreases (Bouchard & McGue, 2003; Deary et al., 2006). Bouchard and McGue review IQ correlations for unrelated siblings reared together, from eleven reports from adoption studies. Consistent with the trend, the weighted average of the six correlations calculated using children was 0.26, whereas that of the five correlations using adults was 0.04. Using the twin IQ correlations then available in the published literature, McGue et al. (1993) calculated Falconer estimates of heritability and shared-environmentality for different age groups. The variance decomposition ranged from $h^2 \approx 0.4$ and $c^2 \approx 0.2$ around age 5, to $h^2 \approx 0.8$ and c^2 near zero in adulthood. Data from longitudinal adoption studies, such as the Colorado Adoption Project (Plomin, Fulker, Corley, & DeFries, 1997), consistently show that, in terms of cognitive ability, adoptees gravitate toward their biological kin and away from their adoptive relatives. Variance-component estimates from longitudinal twin studies provide perhaps the clearest evidence that heritability increases from early childhood to early adulthood. Such a study of Dutch twins (Boomsma, de Geus, van Baal, & Koopmans, 1999; Bouchard & McGue, 2003) yielded steadily increasing heritability estimates, from around 0.30 at age 5, to more than 0.80 at age 27. In contrast, shared-environmentality declined sharply, from around 0.35 at age 5 to near zero at age 16. Recent studies (Davis, Haworth, & Plomin, 2009; Haworth et al., 2010; van Soelen et al., 2011) replicate an increasing heritability through childhood, though they also report shared-environmentality estimates that differ reliably from zero. In old age, the heritability of GCA may decline somewhat. In their review of elder twin studies, Lee, Henry, Trollor, and Sachdev (2010) noted that heritability is quite high ($h^2 \approx 0.80$) for individuals during their mid-60s, and lower during their mid-80s ($h^2 \approx 0.60$).

A more tentative biometric moderator is the socioeconomic status (SES) of the family-of-origin. Two theoretical perspectives—that of Sandra Scarr (1992) and that of Bronfenbrenner & Ceci (1994)—predict that cognitive abilities will be more heritable among children from higher-SES families. The two perspectives are similar in many respects, but differ somewhat in their emphasis, so they make that same prediction for somewhat different reasons. Scarr's theory emphasizes r_{GE}, and asserts that individuals' environments in large part originate causally from their genotypes because of it. Her theory predicts that heritability of most traits will increase with age from childhood to adulthood, because of the increasing role of active r_{GE}, since people are better able to select environments consonant with their genetically influenced preferences as they grow up. It further asserts that as long as rearing environment is minimally supportive of normal development, the differential effects of that rearing environment *per se* will be trivial; conversely, the direct effects of rearing environment will only be appreciable at the most unfavorable extremes. Therefore, Scarr's theory would predict a lower influence of heredity, relative to that of the shared environment, only among those from very low-SES families. However, it allows for an absolute increase in genetic variance with increasing SES, due to increasing active r_{GE}, since more-affluent families can provide their children with access to a wider range of environments.

Bronfenbrenner & Ceci's theory is somewhat more nuanced. It proposes that genetic potential for psychologically adaptive development is actualized by certain "proximal processes" (which are conceptually distinguishable from the environmental context in which they occur). Bronfenbrenner & Ceci concur with Scarr that children are able to modify their own environment, and become better able to do so as they mature. However, they point out that in early childhood, parents are in a much better position to influence their children's environment and facilitate the action of proximal processes. In particular, parents are better able to engender adaptive development when they possess relevant knowledge and access to more resources, and are able to provide the stable environment that proximal processes require to actualize children's potential over time. For cognitive abilities, then, their theory predicts a continuous gradient of increasing heritability with increasing SES, since heritability can only represent the proportion of variance attributable to *actualized* genetic potential.

Scarr (Scarr-Salapatek, 1971) was the first to investigate whether the heritability of children's GCA might vary as a function of their family SES, using a sample of ~1000 twin pairs from the Philadelphia public school system. Scarr's results were consistent with higher heritabilities at higher SES levels. But, this early study has been criticized for major limitations (e.g., Turkheimer et al., 2003; Hanscombe et al., 2012), such as its indirect resolution of zygosity in same-sex twin pairs, and the use of neighborhood census-tract data (rather than data from the twins' actual parents) to operationalize SES. A later study of Swedish twins (Fischbein, 1980) did not suffer from these limitations. Fischbein stratified twins by their father's occupational status (3 levels) and reported the MZ and DZ intraclass correlations by stratum. These results were also consistent with higher heritabilities at higher SES levels, though the sample size was rather small (<300 twin pairs), and Fischbein reported no inferential statistics. More importantly, results were only presented for two ability tests, separately, and not for any composite from multiple tests, so the twins should not be considered assessed for *general* cognitive ability.

Subsequently, Rowe and co-authors reported two studies (van den Oord & Rowe, 1998; Rowe, Jacobson, & van den Oord, 1999) investigating gene-environment interaction in cognitive abilities. Unlike those of Scarr and Fischbein, both studies employed data-analysis methods allowing statistical inference about moderation effects. van den Oord and Rowe utilized a genetically informative design by identifying pairs of siblings, half-siblings, and cousins in the National Longitudinal Survey of Youth. Their sample comprised ~3300 children who had been assessed with an achievement test composed of a mathematics, word-recognition, and reading-comprehension subtest. Participants' home environments had been measured on eleven variables, including family poverty status and each parent's educational attainment. Multilevel regression provided little evidence that the eleven variables moderated heritability of total achievement scores or subtest scores. Instead, most significant (at $p < 0.01$) interactions indicated that either total variance, or unshared-environmental variance only, were smaller in favorable home environments compared to less-favorable home environments. In contrast, Rowe et al. reported that parental education significantly moderated both heritability and shared-environmentality of picture-vocabulary scores in their sample of ~1900 secondary-school-age sibling pairs from the Add Health dataset. Their double-entered DeFries-Fulker regression showed that heritability increased, and shared-environmentality decreased, with increasing parental education. However, neither of these studies assessed their participants on a broad spectrum of abilities. The achievement test in van den Oord & Rowe's study, of course, only represents abilities heavily dependent on formal education, whereas Rowe et al.'s participants only took a Vocabulary test. These instruments should be regarded as measuring a narrower construct than GCA.

The Turkheimer Effect

In an important study that has generated much interest[9], Turkheimer et al. (2003) reported that the biometric decomposition of FSIQ (from the Wechsler Intelligence Scale for Children) varied as a function of parental SES in a sample of 319 pairs of 7-year-old twins. They operationalized SES as a composite of parental education, income, and occupational status. At the upper extreme of the SES variable, IQ variance decomposed into ~80% additive-genetic variance and near-zero shared-environmental variance, whereas at the lower extreme of the SES variable, it decomposed into near-zero additive-genetic variance and ~60% shared-environmental variance. Further, unshared-environmental variance decreased with SES. However, judging by what is mentioned in the title and abstract of Turkheimer et al.'s article, it is the moderation of genetic variance (a specific form of $G \times E$, which we will designate as $A \times SES$) that is of primary interest, with the moderation of shared-environmental variance (shorthand, $C \times SES$) of secondary interest. Therefore, we will here define the "Turkheimer effect[10]" to refer to the phenomenon of increasing additive-genetic variance, and possibly decreasing shared-environmental variance, with increasing family SES.

Turkheimer et al. (2003) was the first study of its kind, that is, the first to apply the continuous-moderator model of Purcell (2002) to the question of SES's role as a biometric moderator of GCA. The continuous-moderator model estimates the moderation effect as an interaction between the moderator and the latent biometric factor, in an analysis of

[9] As of 5/12/13, it had 613 citations in Google Scholar, 80 in PubMed, 67 in PsycINFO, and 315 in Web of Science.

[10] We are certain that we are not the first to coin the term "Turkheimer effect" in the sense meant here, but have not yet been able to track down an earlier usage.

unstandardized phenotype. As its name implies, it retains the quasi-continuous nature of the moderator, by formulating the statistical analysis as a regression of the phenotype onto the observed moderator and the latent biometric factors. It is therefore preferable to coarsening the moderator into a smaller number of strata and separately estimating biometric variance components within each stratum (as in, say, Fischbein 1980). Further, it is also preferable to double-entry DF regression, because it estimates how the magnitudes of the *raw* variance components, rather than the *standardized* components, vary as a function of the moderator.

Figure 3-2 presents a path diagram of Purcell's (2002) continuous-moderator model, depicting only twin #1's half of the full diagram. Here, it is no longer assumed that the phenotype has zero mean. Rather, the phenotypic mean now depends upon an intercept, and a slope from regression onto the moderator, *M* (which would be SES in Turkheimer et al., 2003). Additionally, the coefficients on the paths connecting the phenotype to the latent biometric factors now also depend upon *M*. It is important to recognize that the variance actually being biometrically decomposed is the residual phenotypic variance remaining after partialling out *M*. It is also important to recognize that (if *M* were SES) the variance being attributed to the main effect of SES would otherwise be attributed to the shared environment, *C*. This is a consequence of the way the model is identified in twin data: strictly speaking, the latent *C* factor really represents all sources of variance that contribute equally to MZ- and DZ-twin similarity. Since family SES is the same for both twins in a pair, irrespective of zygosity, it is effectively part of the shared environment as far as this model is concerned. However, the association between family SES and children's GCA is surely at least partly genetically mediated, as evident from the larger associations between family characteristics and offspring ability in biological families vis-à-vis adoptive families (e.g., Scarr & Weinberg, 1978; Kirkpatrick, McGue, & Iacono, 2009). This is an example of passive r_{GE} (specifically, correlation between genes and the common environment, or r_{AC}): parental cognitive ability and SES are positively correlated, and higher-ability parents pass on their trait-relevant genes to their children as well as provide them with an enriched rearing environment. The presence of r_{GE} can lead to spurious detection of $G \times E$, and indeed, part of the justification for regressing out the main effect of SES is to control for the presence of r_{GE} (Purcell, 2002).

We are aware of four studies interpretable as attempts at replicating the Turkheimer effect on GCA (Harden, Turkheimer, & Loehlin, 2007; van der Sluis, Willemsen, de Geus, Boomsma, & Posthuma, 2008; Grant, Kremen, Jacobson, Franz, Xian, et al., 2010; Hanscombe et al., 2012). Each of these studies applied the continuous-moderator model to unstandardized scores on a test composed of both verbal and non-verbal subtests. Thus, we do not consider studies of SES-moderation in specific abilities, such as reading achievement (e.g., Kremen et al., 2005) or mathematics achievement (e.g., Tucker-Drob & Harden, 2012), or separate verbal and non-verbal ability factors (e.g., Asbury, Wachs, & Plomin, 2005). We also do not consider studies of cognition in very young children (e.g., Tucker-Drob, Rhemtulla, Harden, Turkheimer, & Fask, 2011), because cognitive tests prior to age 2 correlate weakly with cognitive ability measured at later ages (Wilson, 1983) and sample a limited amount of verbal content (Bouchard & McGue, 2003), rendering them questionable as measures of the GCA construct. General cognitive ability does not emerge as a stable trait until around age 4 or 5.

Harden et al. (2007) reported an analysis using a sample of 839 pairs of adolescent twins who sat for the 1962 National Merit Scholarship Qualifying Test (NMSQT). The NMSQT, an aptitude test designed to assess scholarship candidates' potential for success in future educational endeavors, comprised five subtests: English Usage, Mathematics Usage, Social Science Reading, Natural Science Reading, and Word Usage / Vocabulary. Evidently, the NMSQT sampled relatively little content disconnected from mastery of a formal educational curriculum or involving nonverbal cognitive processes. It is therefore with some reservation that we regard it as a measure of GCA. Indeed, Harden et al. acknowledged that the NMSQT is not an IQ test like that used in Turkheimer et al. (2003).

The actual phenotype biometrically decomposed in Harden et al.'s (2007) analysis was the latent common factor extracted from the five NMSQT subtests. Harden et al. used two SES variables, mid-parental education level and family income. A comparison of the observed income distribution to that of the 1960 U.S. census revealed that their twins disproportionately came from relatively affluent families. Harden et al. conducted separate moderation analyses with each SES variable, and in both analyses estimated $A \times SES$ and $C \times SES$ effects. The analyses yielded a significant ($p \approx 0.02$) effect, in the hypothesized direction, of income on additive-genetic influence only, but also suggestive evidence of the other moderation effects (all other ps < 0.1). However, none of the point estimates of the $C \times SES$ effect were in the hypothesized direction. Thus, Harden et al. (2007) constitutes a replication of the primary $A \times SES$ element of the Turkheimer effect: increasing additive-genetic variance with increasing SES.

Van der Sluis et al. (2008) reported a replication attempt in an adult sample of Dutch twins. The sample comprised two age cohorts: a younger cohort (mean age ~27, $N = 385$) and an older cohort (mean age ~49, $N = 370$). Participants were assessed on an abbreviated form of the Dutch WAIS-III consisting of 9 subtests—a broad sampling of the domain of cognitive tasks. FSIQ scores were age- and sex-corrected prior to analyses. Midparental education level was scored as a binary indicator for completing higher education; on average, twins in the younger cohort had more highly-educated parents. Because of this cohort difference (and because of sex differences with respect to other variables not germane to the present discussion), van der Sluis et al. fit moderation models separately by cohort and sex. All biometric-moderation effects of parental education could be dropped without significantly disturbing model fit, apart from that for shared-environmental variance among older males. This effect size translates into $c^2 = 0.47$ among older males with higher-educated parents, but $c^2 = 0$ otherwise. Clearly, the results of van der Sluis et al. (2008) did not support the Turkheimer effect.

Grant et al. (2010) conducted a replication attempt with a sample of ~6000 adult-male twins from the Vietnam-Era Twin Registry, whose general cognitive ability had been assessed with the Armed Forces Qualification Test upon their induction to the military. Midparental education level was calculated for each pair as the average of the twins' reports of their mother's and father's years-of-education completed. Then, these averages were converted to a scale on which the lowest level of parental education was coded 0, and the highest level was coded 1. But, moderation analysis provided no evidence that any of the three biometric variance components varied as a function of parental education. Thus, Grant et al. (2010) also did not replicate the Turkheimer effect.

Most recently, Hanscombe et al. (2012) attempted to replicate the Turkheimer effect in a longitudinal dataset from the Twins Early Development Study. The sample was composed of 8716 British twin pairs in which at least one twin had been IQ-tested at any of the target ages (2, 3, 4, 7, 9, 10, 12, and 14; assessments differed appropriately by age), and for which at least one of the three indices of parental SES was available. These three indices were (1) parental education level and occupational status, as recorded at recruitment, when the twins were 18 months old; (2) parental education level and occupational status as recorded when the twins were 7 years old; and (3) parental income as measured when the twins were 9 years old. Altogether, Hanscombe et al. presented results of separate biometric model-fitting at for the 17 different combinations of age and SES index. We will here consider their results for ages 7 and older[11].

For SES index #1, the AIC-preferred model included only a $C \times SES$ effect at ages 9 and 14, only an $A \times SES$ effect at age 10, and no SES-moderation at ages 7 and 12. For SES index #2, the AIC-preferred model included only a $C \times SES$ effect at ages 9, 10, and 12, and no SES-moderation at ages 7 and 12. For SES index #3, the AIC-preferred model included only a $C \times SES$ effect at ages 9, 10, 12, and 14. Thus, the moderation effect most consistently implicated is moderation of the shared-environmental component. Although the $C \times SES$ effect is in the hypothesized direction in almost all cases, the only clear effect on the additive-genetic component (with SES index #1 at age 10) was in the direction opposite that of the Turkheimer effect (i.e., decreasing additive-genetic variance with increasing SES). Thus, Hanscombe et al. do not replicate the primary $A \times SES$ element of the Turkheimer effect, merely the secondary $C \times SES$ element.

Let us here briefly summarize our preceding review. The primary $A \times SES$ element of the Turkheimer effect is that additive-genetic variance increases with increasing family-of-origin SES; the secondary $C \times SES$ element is that shared-environmental variance decreases with increasing SES. It was first observed in Turkheimer et al.'s (2003) study of a U.S. sample of 7-year-old twins. Harden et al. (2007) replicated the primary $A \times SES$ element in a U.S. adolescent sample when operationalizing SES as parental income (and obtained suggestive results when operationalizing SES as midparental education). Two studies of adult twins, one from the Netherlands (van der Sluis et al., 2008) and one (all males) from the U.S. (Grant et al., 2010), failed to replicate the effect; both operationalized SES as midparental education. Most recently, a longitudinal study of British twins, assessed through childhood and into early adolescence, reported evidence supporting only the secondary $C \times SES$ element of the Turkheimer effect.

[11] Hanscombe et al. (2012) reported correlations between IQ scores at different ages in their Table 2. IQ at ages 2, 3, and 4 correlate significantly, though modestly, with IQ at the later ages—in no case did the correlation exceed 0.30. We regard this as supporting evidence for our assertion that GCA does not emerge as a stable trait until age 4 or 5.

Assuming that the effect is real, how can this mixed replication record be explained? One obvious possibility is that the Turkheimer effect is age-dependent, and declines in magnitude with increasing age. As Hanscombe et al. (2012) point out, since shared-environmental variance declines with increasing age, there may not be enough of it among adults for the second element of the Turkheimer effect to be detectable (especially if some of it is being partialled out by the main effect of SES). But, Hanscombe et al.'s point estimates and graphs do not reveal any clear age-related trend for the $A \times SES$ effect nor for the $C \times SES$ effect of two of the three SES indices, and the $C \times SES$ effect of SES index #3 actually becomes more pronounced with increasing age. If the Turkheimer effect is present in childhood but absent in adulthood, perhaps its age-related decline occurs mainly during adolescence, that is, in a slightly older age-range than that of the TEDS twins.

Other possible explanations involve the SES variable, specifically, how it was measured, and what its relations with other variables might be. We note that the two replication failures (van der Sluis et al., 2008; Grant et al., 2010) had parental education as the only available SES measure, whereas Turkheimer et al. (2003) and Harden et al. (2007) had parental income available as well. Also, Hanscombe et al.'s clearest evidence for the $C \times SES$ effect came from analyses using SES index #3, parental income, though this could be because index #3 was measured in closer temporal proximity to the ages at which it was used in analysis. Further, Hanscombe et al. (2012) pointed out that the developmental implications of family-of-origin SES, and the adequacy of a particular operationalization, are presumably different in different countries. We note that the samples supporting the primary $A \times SES$ element of the Turkheimer effect (Turkheimer et al. and Harden et al.) were from the United States, the samples failing to replicate the effect were from the U.S. (Grant et al.) and the Netherlands (van der Sluis et al.), and a British sample replicated the secondary $C \times SES$ element (Hanscombe et al.).

Of course, the Turkheimer effect could be spurious. Its primary $A \times SES$ element seems less plausible from sample-size considerations alone, since it is only supported in samples of fewer than 1000 twin pairs (Turkheimer et al., 2003; Harden et al., 2007). Several phenomena can lead to detection of spurious $G \times E$. One of these is differential heritability (or shared-environmentality) by phenotype level. If the influence of A increases, or the influence of C decreases, with increasing GCA, then this heterogeneity may appear to be a biometric moderation effect of SES, simply because SES and GCA are positively correlated[12]. Another complication is assortative mating: if there is greater assortative mating for GCA at lower SES levels, then DZ twins from lower-SES families will share more of their trait-relevant alleles identical-by-state (Loehlin, Harden, & Turkheimer, 2009). Then, DZ-twin covariance will be higher at lower SES levels, but the MZ-twin covariance will not vary by SES. Thus, additive-genetic variance will appear to increase with increasing SES. Yet another pitfall is the fact, discussed above, that the main effect of family SES necessarily partials out variance that would otherwise be shared-environmental, even though, due to passive r_{GE}, the correlation between SES and offspring cognitive ability is at least partly genetic in nature. In a follow-up to Harden et al. (2007), Loehlin et al. (2009) evaluated the consequences of assortative mating, and of correlation between family income and the latent A factor (representing passive r_{GE}). Through simulations, they showed that income's estimated $A \times SES$ effect changed little in value by manipulating passive r_{GE} and SES-independent assortative-mating parameters, and remained statistically significant except at high levels of passive r_{GE}. Further, they showed that DZ-twin correlations, and spousal correlations for education level, did not vary by family income levels, which suggests that SES-dependent assortative mating is unlikely to be present in their sample. Loehlin et al. are to be commended for making this effort to evaluate the sensitivity of their conclusions to the

[12] See Tucker-Drob, Harden, & Turkheimer (2009) and McCallum & Mar (1995) for discussion of how quadratic trends may be mistaken for multiplicative interactions.

assumptions of their analyses.

Finally, there is the issue of the specificity of biometric-moderation effects. Under the continuous-moderator model, biometric moderation may be thought of simply as heteroskedasticity in the regression of the phenotype onto the putative moderator. The specificity issue concerns how well an analysis can resolve which and how many biometric-moderation effects are nonzero—that is, which biometric variance components are heteroskedastic. Purcell (2002) remarked on this issue when discussing a substantial estimate of a $C \times SES$ effect in simulated data when the true generating model only had an $A \times SES$ effect. Both Turkheimer et al. (2003) and Hanscombe et al. (2012) refer to the issue as well. It is also evident in Harden et al.'s Table 5: the estimate of the $A \times SES$ effect when the $C \times SES$ effect was fixed to zero was very similar to the estimate of the $C \times SES$ effect when the $A \times SES$ effect was fixed to zero. This calls to mind an essential fact: inference about a parameter is model-dependent; specifically, it depends upon which other parameters are free to be estimated in the model at hand. Harden et al. did not consider a model with an $E \times SES$ effect, even though such an effect is not preposterous and was reported by Turkheimer et al. One wonders what Harden et al.'s conclusions would have been had they also, say, allowed for $E \times SES$—how much of the heteroskedasticity would the unshared-environmental component have "soaked up?"

The Information-Theoretic Approach and Multimodel Inference

It bears repeating: inference about a parameter is model-dependent. Of course, this is something that researchers already know (and likely have known since they were students and, say, saw a regression coefficient drastically change when another predictor was added or removed from the model). But we wish to highlight this point in order to emphasize two others. The first of these is that, although the distinction may be blurry at times (particularly in Bayesian contexts), model selection is distinct from, and should mostly precede, statistical inference.

Failing to follow this principle is usually venial, though it can at times lead to plans of analysis and interpretations of results that are less trenchant and more muddled than they otherwise could be. But we wish to bring attention to one specific objectionable practice that is widespread in behavior genetics: that of selecting a single model that is best (by some criterion), and then basing inference only on that model, as though no others had ever been considered. Breiman (1992, p. 738) has called this practice "a quiet scandal." We instead provide an alternative approach with our third point: inference about parameters can be based on *multiple* models; in fact, one can (in a sense) select many or even *all* models under consideration, each only to the extent that it is supported by the data. Much of the present study is conducted using methods of multimodel inference. Awareness of these methods is not as widespread as we believe it should be, which is why we describe them here. Our description mostly follows that of Burnham & Anderson (2001, 2002, 2004), whose work we recommend for further details.

Kullback & Leibler's important 1951 paper concerns, *inter alia*, derivation of a metric representing how well one probability distribution is approximated by another. Specifically, it is the expected amount of information (in Kullback & Leibler's generalized Shannon-Wiener sense) lost when one probability distribution is approximated by another. This metric has become known as Kullback-Leibler (KL) divergence. A sensible objective of model selection, then, is to choose the model that has the smallest KL divergence from full reality. Full reality, of course, is not known, and may not even be knowable in principle; possibly, any complete description of full reality would be infinitely long. If we accept the possibility that no statistical model can completely describe full reality, then the premise of a "true model" that generated the data becomes a rather dubious notion. These issues pose no problem, however, if one is only interested in the *relative* divergence of different models, since the unknown constants depending upon full reality cancel out from subtraction.

In a series of important contributions in the 1970s, Hirotugu Akaike[13] showed that the maximized joint loglikelihood of a model's parameters estimates how relatively "close" (in a KL-divergence sense) the model is to full reality, except that this estimator is biased upward, because it represents the fit of the model in the same data from which its parameters were estimated. Akaike further showed that, in large samples, the magnitude of this bias is in fact approximately equal to *k*, the number of free parameters. Subtracting *k* from the loglikelihood thus serves to estimate the expected loglikelihood of the model when "plugging in" parameter estimates previously obtained from a separate, independent sample of the same size. Akaike multiplied this bias-adjusted loglikelihood by -2 (to turn it into a bias-adjusted deviance), obtaining what has become known as Akaike's Information Criterion,

$$AIC = -2logL\left(\widehat{\boldsymbol{\theta}} \mid M, x\right) + 2k \tag{3.1}$$

where $\widehat{\boldsymbol{\theta}}$ is the vector of maximum-likelihood estimates of model *M*'s *k* parameters, as estimated from dataset *x*. In theory, the candidate model with the smallest AIC is the model that best approximates full reality, conditional on sample size *N* and the set of candidate models considered. The expected relative KL divergence of two candidate models may be estimated simply by subtracting their AICs.

As is evident from the previous paragraph, AIC is a penalized fit index. The unpenalized model deviance, $-2logL\left(\widehat{\boldsymbol{\theta}} \mid M, x\right)$, by itself is a poor measure of a model's merit, as it may be made arbitrarily small by adding parameters and increasing model complexity. AIC's penalty is the approximate amount by which model deviance is underestimated when assessing the model in the same sample in which its parameters are being estimated. In other words, AIC has deep theoretical connections to cross-validation (discussed further by Stone, 1977; Shao, 1997; and

[13] Unfortunately, several important primary sources by Akaike are inaccessible to us, due to being conference presentations or being written in Japanese. We do not cite sources we cannot read. Here, we rely on secondary sources by Burnham & Anderson (2001, 2002, 2004).

Browne, 2000). Specifically, in large samples, it is expected to select that model in the candidate set which minimizes error of prediction in new samples of the same size from the population, where error is based on a loglikelihood function (Hastie, Tibshirani, & Friedman, 2009). Since maximizing normal likelihood is equivalent to minimizing quadratic loss, and since many analyses assume (at least implicitly) a normal distribution, in many contexts AIC is expected to select that model in the candidate set which minimizes *mean squared* error of prediction. We therefore phrase our interpretations of AIC in terms of “efficiency” or “performance”—shorthand for *expected relative* efficiency or performance—rather than “fit,” because, again, one can just add more parameters to improve model fit to the data at hand.

However, one of AIC’s appealing qualities is that it allows the expected relative efficiency of *all* the models in the candidate set to be compared to one another. Unlike the likelihood ratio test (LRT), AIC can be used to compare multiple models to one another and rank them in terms of their merit; they need not be a sequence of nested models. In fact, different models’ AICs will be comparable to one another provided that the models all (1) have the same dataset (and in particular, the same *N*), (2) have the same endogenous variable(s) (which are no longer considered “the same” if they have been transformed in any way), and (3) either have likelihood functions from the same family of distributions *or* use fully normalized likelihood functions (Burnham & Anderson, 2002).

We now describe how AIC can be used to weight the results of multiple models under consideration, and obtain model-averaged point estimates and sampling variances. Let AIC_{min} denote the smallest AIC in a set of *m* comparable models. Then, those models’ AICs can be re-expressed relative to AIC_{min}. For some model *l*, let $\Delta_l = AIC_l - AIC_{min}$. Then, model *l*’s Akaike weight can be calculated as

$$w_l = \frac{\exp{(-\Delta_l/2)}}{\sum_{i=1}^{m} \exp{(-\Delta_i/2)}} \tag{3.2}$$

Do this for all models $l = 1, \dots, m$. The resulting Akaike weights are normalized (sum to 1); each is interpretable as the posterior probability that its model is the one that minimizes K-L divergence from full reality in the population (again conditional on *N* and the candidate set of comparable models; Burnham & Anderson, 2002). The implicit prior probability on each model in the set calculated is not equal for all models. Instead, it is a "savvy prior" that takes into consideration the number of free parameters relative to sample size (see Burnham & Anderson, 2004).

Once Akaike weights are computed for all comparable models in the candidate set, a pragmatic way to proceed is to average each parameter's estimates, and their corresponding sampling variances, across those models in which the parameter is free to be estimated[14] (Burnham & Anderson, 2002). For purposes of model-averaged estimates, the Akaike weights need to be re-normalized so that they sum to 1 within the subset of models in which the parameter of interest is free. If some parameter θ is a free parameter in some subset $\mathcal{S}$ of the comparable set of models, then for some model *l* within that subset, the re-normalized Akaike weight w_l^* equals

$$w_l^* = \frac{w_l}{\sum_{i \in \mathcal{S}} w_i} \tag{3.3}$$

Do this for all models *l*, $l \in \mathcal{S}$. With the re-normalized weights, the model-averaged point estimate of θ can be calculated:

$$\hat{\theta}. = \sum_{i \in \mathcal{S}} w_i^* \hat{\theta}_i \tag{3.4}$$

where $\hat{\theta}_i$ is the maximum-likelihood estimate of θ, conditional on model *i*. In a sense, when computing $\hat{\theta}.$, one is "integrating out" the model-dependence of the point estimates by averaging across models informative about the parameter, each contributing to the average in proportion to

[14] It may be objected that basing inference about a parameter only upon those models in which it is freely estimated ignores evidence about the parameter conveyed by those models in which it is fixed. If one's objective is regression prediction rather than inference, Burnham & Anderson (2002) do recommend calculating the model-averaged regression coefficient from models in which it is fixed, as well as those in which it is free. However, as Bartels (1997, footnote 11) points out, a model-averaged estimate computed in this way will not have a normal sampling distribution.

its relative weight-of-evidence. The model-averaged point estimate $\hat{\theta}.$ has estimated sampling variance equal to (Burnham & Anderson, 2004):

$$v\hat{a}r(\hat{\theta}.) = \sum_{i \in S} w_i^* \left[v\hat{a}r_i(\hat{\theta}_i) + (\hat{\theta}_i - \hat{\theta}.)^2 \right] \tag{3.5}$$

where $v\hat{a}r_i(\hat{\theta}_i)$ is the estimated sampling variance of the MLE of θ, conditional on model i. Thus, the model-averaged sampling variance represents a weighted average of within-model variance estimates and between-model variance estimates. In the simplest application, one uses the square root of $v\hat{a}r(\hat{\theta}.)$ as the standard error to form confidence intervals and test null hypotheses, assuming asymptotic normality of $\hat{\theta}.$, which is what we do herein.

The chief advantage of multimodel inference is that it enables the researcher to base inference about parameters on all models under consideration, allowing each model to contribute in proportion to how well it is supported by the data (Burnham & Anderson, 2002). Even if, say, the best-approximating model has the shared-environmental effect fixed to zero, it does not necessarily follow that the best estimate of the effect is zero, especially if other models under consideration had AICs close to that of the best model. The multimodel approach attempts to avoid the biased estimation and inference that result from conditioning one's conclusions on a single best model (Lukacs, Burnham, & Anderson, 2009). In applied contexts, information-theoretic model-averaging can also improve predictive accuracy (e.g., Kapetanios, Labhard, & Price, 2008).

Study Overview

Our study, which attempts to replicate the Turkheimer effect, improves upon previous replication attempts in several ways. First, our large sample is composed of twins, non-twin biological siblings, and adoptive siblings, assessed at a range of ages spanning the teenage years. A prior study of IQ in a substantially identical sample has been reported (Kirkpatrick, McGue, & Iacono, 2009). The presence of adoptees provides us with a "backstop" against artifacts stemming from passive r_{GE} and assortative mating, and allows us to directly estimate shared-environmental variance (and, in principle, variance due to covariance between the A and C factors). Second, we also have parental phenotype—IQ scores for the parents of the twins and siblings—and therefore can estimate assortative mating, both SES-independent and SES-dependent. Third, we have data on the same three SES indices used in the original Turkheimer (2003) report. Our study addresses the following research questions, in turn:

1. Which sources of biometric variance should be represented in our model? The ACE model is quite plausible *a priori* from existing literature (reviewed above), especially for an adolescent (rather than adult) sample, and in light of the dearth of evidence for non-additive genetic variance in the domain of cognitive abilities (Bouchard, 2004). However, we can estimate more than two sources of familial variance in our sample. One possibility would be twin-specific environmental effects ("twin effects"), which would contribute to between-family variance among twins but not among non-twin siblings. One limitation of Kirkpatrick et al.' s (2009) study is that they assumed an ACE model *a priori*, but a glance at the correlations in their Table 2 suggests the possibility of twin effects. Of course, this could be because Kirkpatrick et al. did not correct for age and sex prior to calculating those correlations (McGue & Bouchard, 1984). Another possible source of variance is assortative mating. We need not assume that the additive-genetic correlation between full siblings is 0.5—we can estimate it from the data, because an ACE model would be identified by MZ twins and adoptees alone.
2. (a) Does the biometric decomposition of IQ vary as a function of trait level? (b) Is there SES-dependent assortative mating among parents? Answering Questions 2a and 2b serves to test for two possible sources of spurious biometric-moderation effects. Studies pertinent to Question #2(a) have been conducted in children younger than 4 (Cherny, Cardon, Fulker, & DeFries, 1992; Petrill, Saudino, Cherny, Emde, Fulker, et al., 1998),

with mixed results. In a sample of older children (ages 6 to 12), Detterman, Thompson, & Plomin (1990) reported that, with decreasing ability, shared environmentality decreased while heritability decreased. No differential effects were found in a study of adult twins (Saudino, Plomin, Pedersen, & McClearn, 1994). Previously, Kirkpatrick et al. (2009) found no evidence that high ability *per se* is differentially heritable or shared-environmental, so we anticipate that our data will answer #2(a) in the negative. There is much less existing research pertaining to Question #2(b), though Loehlin et al. (2009) reported results that are not consistent with SES-dependent assortative mating.

3. <u>Can we replicate an increase in additive-genetic variance, and a decrease in shared-environmental variance, with increasing age among adolescents?</u> As we described above, biometric-moderation effects of age in GCA are well-documented. It would be surprising if we do not replicate them.
4. <u>Can we replicate the Turkheimer effect?</u> We will attempt to replicate both its primary $A \times SES$ and secondary $C \times SES$ elements, that is, both increasing additive-genetic variance and decreasing shared-environmental variance with increasing SES.
5. <u>Is the Turkheimer effect age-dependent?</u> If so, we anticipate that it would weaken through adolescence.

Thus, we first (Questions #1 and #2) resolve basic questions of model specification, allowing for estimation of certain parameters not possible in samples of only twins. Then (Question #3), we seek to fine-tune our model specification by allowing for biometric-moderation effects with strong *a priori* evidence. Then (Question #4), we attempt to replicate the Turkheimer effect, which is the confirmatory analysis of primary interest. Finally (Question #5), we conduct exploratory analyses that provide a context in which to better understand our replication results.

Sample and Measurements

Sample

The primary sample consists of twins from the Minnesota Twin Family Study ("MTFS"; Iacono, Carlson, Taylor, Elkins, & McGue, 1999; Iacono & McGue, 2002; Keyes et al., 2009), and non-twin sibling pairs from the Sibling Interaction and Behavior Study ("SIBS"; McGue et al. 2007). MTFS is a longitudinal study of a community-based sample of same-sex twins, born between 1972 and 1994 in the State of Minnesota, and their parents. SIBS is an adoption study of sibling pairs and their parents; its community-based sample includes families where both siblings are adopted, where both are biologically related to the parents, or where one is adopted and one is biologically related. Per the SIBS inclusion criteria, any sibling in the sample who was adopted into the family will not be biologically related to his or her co-sibling. For adopted siblings, the mean age at placement was 4.7 months (SD = 3.4 months).

Between MTFS and SIBS, 2,504 families have visited the Minnesota Center for Twin & Family Research. For the purposes of our analyses, the sample (*N* = 4,973, in 2,494 families) comprises five distinct family-types:

1. Monozygotic- (MZ) twin families (*n* = 2401, in 1204 families),
2. Digyzotic- (DZ) twin families (*n* = 1348, in 675 families),
3. SIBS families with two adopted offspring (*n* =567, in 285 families),
4. SIBS families with two biological offspring (*n* = 415, in 208 families),
5. "Mixed" SIBS families with 1 biological and 1 adopted offspring (*n* = 242, in 122 families).

This sample is predominantly Caucasian of European ancestry ("White," 84.4%). Most of the adoptees were adopted internationally from Korea, so 10% of the sample is Asian. The remaining 5.6% of the sample report their ancestry as other than White or Asian.

The MTFS sample is composed of two cohorts, an eleven-year-old cohort (10-13 years old at intake; mean age = 11.78) and a seventeen-year-old cohort (16-18 years old at intake; mean age = 17.47). The age range of the siblings at intake was 10-19 for the younger siblings, and 12-20 for the older. Written informed assent or consent was obtained from all participants, with parents providing written consent for their minor children.

For the present study, we used parental data only from parents who were the “original rearing” parents in the family. Usually, the original rearing parents would be the biological parents of the family’s offspring, unless it was known that one of them had limited contact with the children while they were growing up (due to divorce, etc.). In the case of families with only adopted offspring, the original rearing parents would be those with whom the offspring were originally placed for adoption, unless again it was known that one of them had limited contact with the children.

Measurements

Zygosity.

Twin zygosity was assessed at intake using three indicators: a standard, parent-completed zygosity questionnaire, staff judgment of physical similarity, and an algorithm that used various anthropometric measures. Zygosity determination for putatively DZ twins who provided DNA samples has subsequently been verified from genome-wide marker data (Miller et al., 2012).

Cognitive ability.

Measurement of GCA was included in the design of the intake assessment for most participants, by way of an abbreviated form of the Wechsler Intelligence Scale for Children-Revised (WISC-R) or Wechsler Adult Intelligence Scale-Revised (WAIS-R), as age-appropriate (that is, 16 or younger, and older than 16, respectively). The short forms consisted of two Performance subtests (Block Design and Picture Arrangement) and Verbal subtests (Information and Vocabulary), the scaled scores on which were prorated to determine Full-Scale IQ (FSIQ). FSIQ estimates from this short form have been shown to correlate 0.94 with FSIQ from the complete test (Sattler, 1974). Parents in the SIBS sample were an exception, in that they were not tested with this short form of WAIS-R until the first SIBS follow-up assessment. By design, only one parent per SIBS family returned for this follow-up, which was usually the mother. As a result, IQ data for SIBS fathers is very limited in its availability.

SES.

Our analysis used three family-level SES variables: (1) the higher of the parents' occupational statuses, (2) midparental educational attainment, and (3) annual household income. We only used the occupational and educational data of the original rearing parents. If data were available only for one of the parents, we took that parent's occupation and education as the higher occupational status and the average education level of the couple, respectively. After exclusions, at least one family-level SES variable was observed for 2,501 families.

Mothers' and fathers' occupational status was assessed during the recruitment phone interview with families' mothers. Occupational status was coded on the Hollingshead scale (Hollingshead, 1957). We reverse-scored the Hollingshead scale so that higher values, on a scale of 1 to 7, represent higher status. We coded as missing the occupational status of those who did not work full-time in their reported occupation, those who reported their occupation as "homemaker," and those reported to be retired, disabled, or institutionalized.

Mothers' and fathers' educational attainment was also assessed during the phone interview. We harmonized educational attainment from the slightly different phone interview given to different subsamples into a five-point scale (1 = less than high school, 2 = high school, 3 = some post-secondary education, 4 = four-year college degree, 5 = graduate/professional degree).

Annual household income was collected by parental report at the intake assessment of MTFS, and at the first follow-up visit of SIBS. Income was measured on an ordinal scale representing income brackets: 0 = "less than $10,000," 1 = "$10,001 to $15,000," and so forth, up to a maximum of 12 = "Over $80,000." Unfortunately, this maximum was not very high, and there was a noticeable ceiling effect: about 25% of families in the sample had an income score of 12, the modal score.

Of the 2,501 families, the percentages missing data on each family-level SES variable were 7.4% for occupational status, 0.7% for educational attainment, and 8.4% for household income. Around 85% of families had no missing observations, 14% had one missing observation, and 1% had two missing observations. As did Turkheimer et al. (2003) and Myrianthopoulos & French (1968), we converted each family's score on the three SES variables into a cumulative proportion (from that variable's empirical CDF), and then averaged the available proportions, producing an SES score for each family (if only one proportion was available, it was taken as the family's SES score). There were 2,494 families having both an SES score and FSIQ for at least one of the offspring. There were 2,382 families in which SES and at least one parent's FSIQ score were available.

Analyses and Results

Unless stated otherwise, all analyses were conducted in *OpenMx* (Boker et al., 2011), via full-information maximum-likelihood (FIML) estimation from raw data. FIML estimation makes use of all available information in incomplete data, and is not biased by missing data as long as the missing-data mechanism is missing-at-random (Rubin, 1978; Shafer & Graham, 2002). In most of our analyses, the endogenous variable was offspring IQ, which is assumed to follow a bivariate normal distribution.

For model comparison and multimodel inference, we used Hurvich & Tsai's (1989) sample-size-corrected version of Akaike's Information Criterion, AICc:

$$AICc = -2logL\left(\widehat{\boldsymbol{\theta}} \mid M, x\right) + 2k + \frac{2k(k+1)}{N-k-1} \tag{3.6}$$

In large samples, AICc differs little from AIC. However, some (e.g., Burnham & Anderson, 2004) argue that AICc should always be used in practice, and that AIC's reputation for overfitting has resulted partly from failure to use AICc in simulation studies.

We first estimated IQ standard deviations and sibling correlations, separately by family type, while correcting for age and sex (McGue & Bouchard, 1984), which is especially important in the present case since members of a sibling pair from SIBS were not necessarily the same age and sex, whereas MTFS twins were. From these estimates (Table 3-2), we can see that the DZ-twin correlation and SD were greater than those of the bio sibs (type 4), suggesting the possibility of twin effects. The presence of adoptees also enables us to estimate r_{GE}. Specifically, this would be a correlation between latent *A* and *C* factors (r_{AC}), for biological offspring of the parents only. However, it is evident that the phenotypic variance among adoptees was greater, not less than, the variance among biological offspring. This indicates that the correlation between *A* and *C* would be *negative*—in other words, that a typical person's genes and shared environment affect IQ in opposite directions. On its face, this is a difficult conclusion to accept, particularly because covariance between *A* and *C* is distinguished from shared-environmental variance only by the difference in phenotypic variance between adoptees versus biological offspring, and not by any of the phenotypic covariances. The MTFS twins, who constitute the majority of the biological offspring, were ascertained differently from the SIBS adoptees, and represent a somewhat different population, which alone could account for the differing variances. We therefore decided not to fit any models including an r_{AC} estimate.

We proceeded by fitting models to answer each research question in turn, assessing model performance via AICc, and using the performance of previously fitted models to guide specification of subsequent ones. In our report below, we refrain from reporting parametric inference until all models informative about a particular parameter have been fitted, and—so that Akaike weights can be used—until all AIC-comparable models have been fitted as well. At that point, if more than one model informative about the parameter had been fitted, we computed model-averaged point estimates, with confidence intervals and *p*-values[15] from the model-

[15] We consider effect sizes and their interval estimates to be more scientifically interesting and informative than hypothesis tests. However, our confidence intervals only have a *marginal* 95% coverage probability; their joint

averaged standard error, under the assumption of normal sampling distribution. The AICcs of the models we fit to address Questions #1, #3, #4, and #5 are all comparable, so it will facilitate exposition if we begin with Questions #2(a) and #2(b).

Question #2(a): Does the biometric decomposition of IQ vary as a function of trait level?

This question is important to address because, if yes, then the heterogeneous decomposition may appear as spurious $A \times SES$ or $C \times SES$ effects. DeFries-Fulker regression (DeFries & Fulker, 1985, 1988) with double-entered data (Rodgers & McGue, 1994; Rodgers & Kohler, 2005) has been used to answer Question #2(a) in other studies (e.g., Cherny et al., 1992). With double-entered data, phenotype scores are mean-centered within kinship groups, and then each sibling pair (twins being a special case of siblings) is entered into the dataset twice, with the labels "sibling #1" and "sibling #2" reversed for each entry. Since our data support the use of a model with the ACE biometric components (see Question #1 below), the DeFries-Fulker regression equation we used is

$$K_1 = b_1 K_2 + b_2(K_2 R) + b_3(K_2^2) + b_4(K_2^2 R) + b_5(Age_1) + b_6(Sex_1) \tag{3.7}$$

where K_1 is the phenotype score of sibling #1, K_2 is the phenotype score of sibling #2, R is the coefficient of relationship (1 for MZ twins, 0.5 for full siblings, and 0 for adoptive siblings), Age_1 is the age of sibling #1, and Sex_1 is a dummy variable for whether or not sibling #1 is female. In this model, the interaction coefficients b_3 and b_4 estimate how much the shared-environmentality and heritability, respectively, depend upon trait level.

This type of DeFries-Fulker regression requires complete data within sibling pairs. There were 2,479 pairs in which FSIQ was available for both members. We conducted the regression

coverage probability is presumably smaller. Also, not every free parameter we estimated is an easily interpretable effect size, and further, the null hypothesis is indeed of interest and somewhat plausible for certain parameters. We therefore report p-values as well, and when making decisions about null hypotheses, compare them to the conventional significance level of $\alpha = 0.05$. P-values are also easier than confidence intervals for the reader to adjust for "multiple testing." We report 17 of them altogether. A Bonferroni correction would almost certainly be too conservative, but skeptical readers are free to hold our results to its standard of $\alpha = 0.0029$.

represented by Eq. (3.7) via an implementation of Kohler & Rodgers' (2001) "efficient DF estimation" in the R statistical computing language. The interaction estimates were both small and statistically indistinguishable from zero: $\hat{b}_3 = -2.87 \times 10^{-5}$ (95% CI: -2.36×10^{-3}, 2.30×10^{-3}; p = 0.9807) and $\hat{b}_4 = 7.40 \times 10^{-4}$ (95% CI: -1.72×10^{-3}, 3.21×10^{-3}; p = 0.5560). Further, the joint test of the two interactions was not significant (Wald $\chi^2(2) = 1.05$, p = 0.5913). This DeFries-Fulker regression required exclusion of incomplete sibling pairs, and was only informative about the standardized, not raw, additive-genetic and shared-environmental variance components. Nonetheless, we regard it as reasonably good evidence that the additive-genetic and shared-environmental components do not linearly vary across the FSIQ continuum.

Question #2(b): Is there SES-dependent assortative mating among parents?

This is another inquiry about a potential source of spurious $A \times SES$ effects. We addressed this question by model-fitting in *OpenMx*. We modeled parental IQ with a bivariate normal distribution, having a different mean and standard deviation for mothers and fathers. We fit two models, one in which the spousal correlation was allowed to vary linearly with SES, and one in which it was constant with respect to SES. The former model estimated that the spousal correlation would be 0.42 at the bottom of the SES distribution, and 0.30 at the top—a change of -0.12 (95% CI: -0.29, 0.06), which was not statistically distinguishable from zero (LRT $\chi^2(1) = 1.66$, p = 0.1972). The estimate of the spousal correlation from the latter model (constant across SES) was moderate, and very close to the meta-analytic average reported over 30 years ago (Bouchard & McGue, 1981): r = 0.35 (95% CI: 0.30, 0.39). Obviously, it differed significantly from zero (LRT $\chi^2(1) = 44.45$, $p = 2.54 \times 10^{-11}$).

This analysis indicated that parental assortative mating is not SES-dependent, and is moderate in magnitude. As explained by Kirkpatrick et al. (2009, footnote 1), if we assume a high heritability for adult IQ, that spouses select mates for psychometric IQ *per se*, and that the phenotypic spousal correlation perfectly reflects a genetic spousal correlation, then the classical twin model would underestimate heritability by about 28%, and commensurately overestimate shared-environmentality. However, these are “worst-case scenario” assumptions, and are generally not true. Further, in our dataset, the ACE variance components are identified by the adoptees and MZ twins alone, whose covariances are not affected by the true genetic correlation between full siblings. As described in the next section, we actually estimated this genetic correlation.

Question 1: Which sources of biometric variance should be represented in our model?

To answer Question #1, we fit Block #1, consisting of Models #1through #4, each of which is some variation on the path diagram depicted in Figure 3-3. These four models represented the four combinations of twin effects fixed (to zero) versus free and r_A for full siblings fixed (to 0.5) versus free. All four included the main effects of sex and age, and three separate intercept parameters, one each for twins, biological SIBS offspring, and adoptees. Their AICcs are presented in Table 3-3. Because Block #1 was the first part of a series of comparable models, Table 3-3 also includes their Akaike weights, which are calculated relative to the AICcs of all models in this comparable set.

From Table 3-3, it can be seen that the best-approximating model within this block is #2, which has both r_A and γ_T fixed to their null values. The near-zero Akaike weights in this block make clear that more models, achieving much better expected efficiency, are yet to come. However, the only models in which we consider twin effects or estimate r_A as a free parameter are in this block. Once model-fitting was complete, we calculated their model-averaged estimates, which are reported in Table 3-7. As anticipated, we conclude from Block #1 that the biometric ACE components are sufficient to describe our data, which is somewhat fortunate, since model-fitting with additional variance components would likely complicate interpretation of results and comparison to the existing literature. In the previous section, we conclude that the spousal correlation for IQ is not SES-dependent, and we report here that the genetic correlation for full sibs differs trivially from 0.5. On the basis of the foregoing, we resolved here to assume in further analyses that the effects of assortative mating are negligible.

The models of Block #1 are the only ones we fit that provide single estimates for the ACE variance components when age and sex are regressed out, since all others (save Model #19, described further below) include some kind of biometric-moderation effect, and therefore in a sense estimate different component values at different levels of the moderator. The model-averaged point estimates are $\hat{V}_A$ = 113.00, $\hat{V}_C$ = 30.34, and $\hat{V}_E$ = 43.00, which sum to total residual variance 186.34, and respectively yield standardized estimates $\hat{a}^2$ = 0.61, $\hat{c}^2$ = 0.16, and $\hat{e}^2$ = 0.23.

Question #3: Can we replicate an increase in additive-genetic variance, and a decrease in shared-environmental variance, with increasing age among adolescents?

To address Question #3, we fit Block #2, consisting of Models #5 through #19. The models of this block were all special cases of that depicted in Figure 3-4, which is an extension of the Purcell (2002) continuous-moderation model. The ACE path coefficients in Figure 3-4 represent the sum of several terms:

$$\begin{aligned}\gamma_A^* &= \gamma_{A0} + \gamma_{A1}(Age_1) + \gamma_{A2}(SES) + \gamma_{A3}(SES \times Age_1) \\ \gamma_C^* &= \gamma_{C0} + \gamma_{C1}(Age_1) + \gamma_{C2}(SES) + \gamma_{C3}(SES \times Age_1) \\ \gamma_E^* &= \gamma_{E0} + \gamma_{E1}(Age_1) + \gamma_{E2}(SES)\end{aligned} \tag{3.8}$$

Additionally, we estimated separate a separate β_0 for twins, biological SIBS offspring, and adoptees, and a separate β_{SES} for biological offspring (including twins) and adoptees. Different models of this form are distinguished from one another by which parameters are fixed to zero.

In all models within Block #2, the SES-moderation effects (γ_{A2}, γ_{C2}, and γ_{E2}) and the interaction-moderation effects (γ_{A3}, γ_{C3}, and γ_{E3}) were fixed to zero. To attempt to answer Question #3 (which concerns age-moderation) in isolation of any effect of SES, the main effect of SES (β_{SES}) was fixed to zero in models #5 through #11. However, we were moving toward answering Question #4, which concerns SES-moderation effects. We therefore also want to know something about age-moderation effects when SES is being partialled out.

In Table 3-4, the columns represent which age-moderation effects (γ_{A1}, γ_{C1}, and/or γ_{E1}) are free parameters. In the models of the first row, there was no main effect of SES, which was present in the models of the second row. Again, we estimated two separate β_{SES} regression parameters: one for adoptees ($\beta_{SES,A}$), and one for biological offspring of parents ($\beta_{SES,B}$). We did so because we anticipated a larger effect among biological offspring, since the association between SES and IQ among them would be at least partly genetically mediated due to passive r_{GE}, which would not be the case for adoptees. Model #19 provided single estimates of the ACE components, now that SES has been regressed out as well: $\hat{V}_A$ = 109.22, $\hat{V}_C$ = 23.11, and $\hat{V}_E$ = 43.21, which sum to 175.53, and respectively yield standardized estimates $\hat{a}^2$ = 0.62, $\hat{c}^2$ = 0.13, and $\hat{e}^2$ = 0.25. As expected, SES mostly accounted for variance that would otherwise be absorbed by V_C. The dramatic improvement in the AICcs of the second column over the first column shows that the two SES main effects contributed enormously to model performance.

However, within each of the table's two rows, the AICcs tell a similar story. The best-approximating model was one with age-moderation of all three variance components. But, not far behind was the second-best model, which has age-moderation of the *A* and *C* effects only, i.e. those expected from the existing literature. However, the third best model was also the most parsimonious, as it includes no age-moderation effects. From the AICcs, we would conclude that the answer to Question #3 is "yes," but we were going to fit more models informative about age-moderation effects, so we were not yet ready to draw inferences about the parameters. We therefore resolved to attempt to replicate the Turkheimer effect under three different specifications of age-moderation: one including all three effects, one including only those for *A* and *C*, and one with no age-moderation.

Question #4: Can we replicate the Turkheimer effect?

To address Question #4, we fit Block #3, consisting of Models #20 through #40. The AICcs and Akaike weights of this block are reported in Table 3-5, from which we draw several conclusions. First, the performance of any particular combination of SES-moderation effects depended little on which age-moderation effects are included in the model. Second, models that included an $A \times SES$ effect clearly fared better than those that did not, and those that included an $E \times SES$ effect fared slightly better than those that did not. But, the $C \times SES$ effect appeared quite extraneous. Third, the inclusion of *any* kind of SES-moderation effect improved model efficiency, indicating that the regression of IQ onto SES is heteroskedastic. From these results, we concluded that our data support only the primary $A \times SES$ element of the Tukheimer effect, but not the secondary $C \times SES$ effect.

Question #5: Is the Turkheimer effect age-dependent?

Perhaps the $A \times SES$ effect apparent in our data weakens with age. Perhaps there is a small $C \times SES$ effect lurking in our data that is only operative among younger participants. Certainly, if the Turkheimer effect declines with age, it would help to explain why attempts to replicate it in adults (van der Sluis et al., 2008; Grant et al., 2010) failed. To investigate these possibilities, we fit Block #4, composed of Models #41 through #43. For these three models only, we mean-centered age and SES prior to analysis. Both age- and SES-moderation effects for *A* and *C* should be included, since we are considering the moderation effects of an age × SES interaction. We also included the SES-moderation effect on *E*, since it received moderate support in Block #3. Therefore, all three models had γ_{A1}, γ_{A2}, γ_{C1}, γ_{C2}, and γ_{E2} as free parameters, which should help facilitate interpretation. Model #41 included both γ_{A3} and γ_{C3} as free parameters, whereas only γ_{A3} was free in Model #42 and only γ_{C3} was free in Model #43.

The three models' AICcs and Akaike weights are reported in Table 3-6. None of the interaction parameters contributed to model performance. On this basis alone, we conclude that the answer to Question #5 is "no." However, we are now ready to draw inferences about those interaction parameters, and a number of other parameters of interest as well. Table 3-7 lists model-averaged parameter estimates, plus corresponding confidence intervals and p-values based on the assumption of normal sampling distribution. The estimates of neither γ_{A3} nor γ_{C3} differed significantly from zero. However, as hypothesized, we did observe a significant increase in additive-genetic variance, and a significant decline in shared-environmental variance, with increasing age. Most interestingly, we replicated only the primary $A \times SES$ element of the Turkheimer effect: additive-genetic variance varied positively with family SES. The secondary $C \times SES$ effect was of similar magnitude, but not in the hypothesized direction, and estimated with little statistical precision. Finally, although the AICcs provided some support for $E \times Age$ and $E \times SES$ effects, the model-averaged results show that we do not have sufficient evidence to conclude that they differ from zero.

Discussion

Guided by existing data and theory, we fit a number of biometric models to a relatively large dataset collected from twins, non-twin biological siblings, and adoptive siblings. We compared models by their expected relative Kullback-Leibler divergence, as indexed by a sample-size corrected version of AIC, the AICc (Hurvich & Tsai, 1989). We compared models' AICcs to first resolve basic questions of specification, then to verify age-related effects that are well-supported by prior research, then to attempt to replicate the SES effect of primary interest, and finally to explore the possibility of age-dependent SES effects. We first resolved that an a

priori plausible ACE model would suffice for our purposes, and that the effects of assortative mating and of differential heritability/shared-environmentality were negligible. We then observed tentative evidence of the expected $A \times Age$ and $C \times Age$ effects, but models with an additional $E \times SES$ effect, and with no age-moderation effects at all, performed about as well. Under those three age-moderation schemes, we fit models with various SES-moderation effects, and observed evidence clearly favoring the hypothesized $A \times SES$ effect, and suggestive evidence of an $E \times SES$ effect, but no evidence for the hypothesized $C \times SES$ effect. Our exploratory analysis did not provide any evidence for age-dependent SES-moderation effects. The overall best-fitting model included $A \times SES$ and $E \times SES$ effects, but no age-moderation effects. However, the model-averaged $A \times Age$ and $C \times Age$ effects were significant and in the expected directions (positive and negative, respectively). The model-averaged $A \times SES$ effect was also significant and in the expected direction (positive), but the model-averaged $C \times SES$ and $E \times SES$ effects were not significant. Thus, our study confirms that additive-genetic variance increases with age and that shared-environmental variance decreases with age, and replicates the primary $A \times SES$ element of the Turkheimer effect (increasing additive-genetic variance with increasing SES, a form of $G \times E$).

Though the Turkheimer effect has generated much interest, it has never been fully replicated in any study of general cognitive ability applying Purcell's continuous-moderation model. It has failed replication twice (Grant et al., 2010; van der Sluis et al., 2008), and its secondary $C \times SES$ element has been replicated once (Hanscombe et al., 2012). Our study constitutes the second replication of the primary $A \times SES$ element, after Harden et al. (2007). Interestingly, the primary $A \times SES$ element has only been observed in samples of American children and adolescents, in which parental income was available as an SES variable. It has not replicated in European samples nor in an American sample in which only parental education was available. In public health, it has been shown that income and education each provide different information about health-relevant aspects of an individual's SES, and are usually not so highly correlated that entering both into a regression analysis produces multicollinearity problems (Braveman et al., 2005). Further, a given SES variable's relations with other variables can differ by country, and by demographic strata and regions within countries (Uher, Dragomirecka, & Papezova, 2006; Braveman et al., 2005). Possibly, the primary $A \times SES$ effect is a distinct moderation effect of family income in the United States. More research is needed to evaluate this tentative proposition. In the present study, we could have conducted analyses to gauge how much each of the three SES variables contributed to the $A \times SES$ effect. However, this would be a greater undertaking than it might seem, since rigorously gauging variables' relative importance can be rather involved in multiple regression (Azen & Budescu, 2003), let alone in a structural equation model involving interactions with latent variables.

Our study, Harden et al. (2007), and Turkheimer et al. (2003) were all conducted in samples of American youth in which parental income was available, but the secondary $C \times SES$ element only occurs alongside the primary element in the original study. We offer a speculative explanation for why this is so. Our sample, Harden et al.'s, and Grant et al.'s (2010) are predominantly Caucasian, but Turkheimer et al.'s is mostly (54%) African-American. Perhaps low SES is not enough to produce the extreme deprivation that, according to Scarr (1992), is necessary to amplify the differential effect of the rearing environment; perhaps low SES must be combined with membership in a disadvantaged minority group whose place in and experience of American society is unique due to the historical legacy of slavery.

The fact that the $A \times SES$ effect appears only in children and adolescents, and not in adults, suggests that it could be age-dependent. But, Hanscombe et al.'s (2012) graphs and point estimates show no clear age-related trend; further, we tested this hypothesis directly, and it was not supported. The availability of IQ data at different ages, which allowed us to directly estimate the age-dependence of SES-moderation effects, is one of several advantages our study has over some existing ones. Another advantage is that we were able to empirically check for possible sources of spurious results, including assortative mating, and differential heritability/shared-environmentality by trait level. Still another advantage was the availability of adoptees, whose data are informative about shared-environmental variance, without bias due to assortative mating, passive r_{GE}, or violations of the "equal environments assumption" for twins. We were also able to calculate different SES main effects for adoptees and biological children. The one for adoptees shows that family SES has a moderate, *environmental* effect on children's cognitive functioning, equal to a 6-point IQ advantage for children from the highest-SES families versus the lowest-SES families. Finally, we consider our use of multimodel inference to be a major advantage of our study, because it enables us to produce point estimates and confidence intervals based on all fitted models informative about a parameter, each to the extent that AICc favors it over others. This avoids the bias resulting from conditioning one's parametric inference only upon a single model (Lukacs et al., 2009).

Although multimodel inference is well-suited for inference about one parameter at time, it does not necessarily make for easy interpretation. Consider the model-averaged estimate of the $A \times SES$ effect: $\hat{\gamma}_{A2} = 2.994$. This means that, for the highest-SES families, the loading onto *A* is greater than that for the lowest-SES families by 2.994. But to really interpret this value, one would need a value for the "intercept" loading, γ_{A0}, which is not a parameter of interest. Sometimes, a meritorious model can tell a complete story in a way that model-averaging cannot easily do. For this reason, we also report point estimates and standard errors from the single most AICc-favored model, Model #36 (Table 3-8), and graph how the biometric decomposition would vary by SES according to those estimates (Figure 3-5). It can be seen that the estimates of its free parameters are quite similar to the corresponding model-averaged estimates.

We wish to temper our endorsement of multimodel inference with a few caveats. First, we must emphasize that Model #36 is not necessarily more likely to be the true model because it has the smallest AICc. Likewise, a model's Akaike weight is not the posterior probability that the model is the true model. AIC is not intended to discover the "true" model in the first place. Instead, as stated by Browne (2000, p. 129), AIC is "not appropriate for selecting the best-fitting model in some general sense independent of sampling error, but…for indicating models whose calibrations can be trusted given a specified sample size."

Second, our conclusions depend upon the candidate set of models under consideration. Certainly the reader can think of models we could have fit, but did not. For example, we did not fit any models including both twin effects and SES main effects. But, that is because in Block #1, AICc provided no support favoring models with twins effects over those without, and we judged it unlikely that twin effects would suddenly become important once SES was added to the model. Certainly, we could have fitted more models. But, we have already fitted quite a few, partly because we wanted to obtain estimates of each biometric-moderation effect from models in which other such effects were variously present or absent. We had to balance that objective with the needs to preserve interpretability and a manageable scope, to avoid empirically blind "data fishing," and keep our analyses relevant to our research questions.

It slightly complicates matters that our candidate model set evolved as our analyses proceeded, in that we used the results from previously fitted models to guide specification of subsequent ones. Also, for the sake of interpretability and maintaining a manageable scope, we proceeded from simpler to more-complicated models. In these respects, our approach bears some resemblance to stepwise forward-selection. However, we deliberately avoided some of the most objectionable aspects of stepwise analyses. We did not conduct a purely data-driven, blindly empirical analysis. Our analysis was guided by subject-matter knowledge, each block of models was intended to address a specific research question, and we saved the most exploratory analyses for last. Further, we did not use significance testing for model selection, nor did we base our conclusions solely upon the final model.

One restriction we imposed upon the candidate set is that all the biometric-moderation models we considered are of the form of Purcell's (2002) continuous-moderator model. There are other model formulations arguably more appropriate for estimating $G \times E$ in the presence of r_{GE}, such as others described by Purcell (2002), and those of Rathouz, Van Hulle, Rodgers, Waldman, and Lahey (2008) or of Price and Jaffee (2008). All of these involve biometrically decomposing the putatative moderator in some way. We decided to retain the Purcell formulation because existing studies of the Turkheimer effect have used it, and our study is intended as a replication study. Nonetheless, inclusion of SES main effects in our models is a rather vexing problem. If one thinks of the path diagram in, say, Figure 3-4 as a simultaneous regression of IQ onto both observable and latent variables, then clearly the main effect of SES must be included if any interactions of SES with latent variables are to be included as well. With data from twins only, SES will necessarily account for variance otherwise attributable to C. Our data enabled us to separately estimate the β_{SES} path coefficient for adoptees and biological offspring; the fact that it is slightly larger among biological offspring may reflect passive r_{GE} (r_{AC}, specifically). But both effect sizes are nontrivial, and possibly, enough shared-environmental variance was partialled out that the secondary $C \times SES$ element of the Turkheimer effect was rendered impossible. Different model formulations for estimating biometric moderation in the presence of r_{GE} should be explored in future research.

Our study raises several other questions that can guide future research. We have already suggested three: to what extent are SES-moderation effects dependent upon country, SES measure, or ethnic minority status? Investigators should also consider employing model formulations intended for estimating $G \times E$ in the presence of r_{GE}, such as those discussed above. Future studies could attempt to test specific hypotheses made by the Scarr (1992) and Bronfenbrenner & Ceci (1994) theories about SES-moderation. For instance, Scarr's theory predicts that $C \times SES$ effects are only likely to be observed when the lowest echelons of SES are represented in the sample. Similarly, Bronfenbrenner and Ceci emphasize the importance of environmental stability for effective development. Since family SES is correlated with stability of the rearing environment (Evans, 2004), perhaps stability is what really drives the Turkheimer effect. It would also be interesting to investigate another correlate of SES—parental *phenotype*, that is, parental cognitive ability—as a biometric moderator. Finally, behavior geneticists could attempt to replicate the Turkheimer effect when genetic factors are not latent, but measured as molecular-genetic data. Exciting avenues of $G \times E$ research remain to be explored.

Table 3-1. Descriptive characteristics of Study #3 sample.

	MZ twins, Older cohort	MZ twins, Younger cohort	DZ twins, Older cohort	DZ twins, Younger cohort	Ado-Ado Sibs	Bio-Bio Sibs	Mixed Sibs
N	832	1571	419	929	567	415	242
#families	416	789	210	469	285	208	122
Female(%)	54.46	50.10	52.98	52.64	57.67	51.33	54.13
Age at Intake							
M	17.46	11.78	17.51	11.79	14.82	15.05	14.94
SD	0.47	0.44	0.44	0.41	2.07	1.75	1.87
Mean FSIQ	100.03	103.23	99.16	103.87	106.55	107.60	108.29

Table notes: "Ado-Ado" = both siblings adopted, "Bio-Bio" = both siblings are biological offspring of parents, "Mixed" = one sibling is adopted and one is biological offspring.

Table 3-2. Age- and Sex-corrected FSIQ correlations and standard deviations, by family type.

	MZ twins (Type 1)	DZ twins (Type 2)	Ado-Ado Sibs (Type 3)	Bio-Bio Sibs (Type 4)	Mixed Sibs (Type 5)
SD (SE)	13.69 (0.25)	13.72 (0.30)	14.09 (0.38)	12.92 (0.42)	[a]
r (SE)	0.77 (0.01)	0.50 (0.03)	0.11 (0.06)	0.36 (0.06)	0.24 (0.07)

Table notes: "Ado-Ado" = both siblings adopted, "Bio-Bio" = both siblings are biological offspring of parents, "Mixed" = one sibling is adopted and one is biological offspring.

[a] In mixed families, the standard deviation of biological offspring was constrained equal to that of bio-bio sibs, and the standard deviation of adoptees was constrained equal to that of ado-ado sibs.

Table 3-3. Model-fitting results from Block #1: AICcs and Akaike weights.

	Free γ_T	Fix $\gamma_T = 0$
Fix $r_A = 0.5$	38842.01 1.01×10^{-51} (Model #1)	38840.32 2.34×10^{-51} (Model #2)
Free r_A	38843.95 3.81×10^{-52} (Model #3)	38842.12 9.52×10^{-52} (Model #4)

Table notes: r_A is the correlation between latent factors A_1 and A_2 for full siblings (including DZ twins). γ_T is the path loading for twin-specific environmental effects. AICcs are numbers greater than 30,000, whereas Akaike weights are proportions.

Table 3-4. Model-fitting results of Block #2: AICcs and Akaike weights.

	Age Moderation Effects							
	ACE	AC	AE	CE	A	C	E	None
No SES main effect	38838.36 6.24×10^{-51} (Model #5)	38839.31 3.88×10^{-51} (Model #6)	38842.39 8.32×10^{-52} (Model #7)	38843.48 4.82×10^{-52} (Model #8)	38841.88 1.07×10^{-51} (Model #9)	38845.79 1.52×10^{-52} (Model #10)	38841.48 1.31×10^{-51} (Model #11)	38840.32 2.34×10^{-51} (Model #2)[a]
SES main effects	38628.58 2.23×10^{-5} (Model #12)	38628.90 1.90×10^{-5} (Model #13)	38630.61 8.08×10^{-6} (Model #14)	38631.17 6.11×10^{-6} (Model #15)	38629.88 1.16×10^{-5} (Model #16)	38629.95 1.12×10^{-5} (Model #17)	38629.16 1.67×10^{-5} (Model #18)	38628.00 2.98×10^{-5} (Model #19)

[a] Model #2 is part of Block #1 (see Table 3-3).

Table notes: AICcs are numbers greater than 30,000, whereas Akaike weights are proportions. "Age Moderation Effects" are those latent biometric factors the loadings of which were allowed to be moderated by age.

Table 3-5. Model-fitting results of Block #3: AICcs and Akaike weights.

		Age-Moderation Effects		
		ACE	AC	None
SES-Moderation Effects	ACE	38612.22 0.080 (Model #20)	38612.35 0.075 (Model #27)	38613.71 0.038 (Model #34)
	AC	38612.54 0.068 (Model #21)	38612.96 0.055 (Model #28)	38614.09 0.031 (Model #35)
	AE	38611.77 0.100 (Model #22)	38611.98 0.090 (Model #29)	38611.69 0.104 (Model #36)
	CE	38615.25 0.017 (Model #23)	38615.17 0.018 (Model #30)	38616.04 0.012 (Model #37)
	A	38611.97 0.090 (Model #24)	38612.44 0.071 (Model #31)	38612.14 0.083 (Model #38)
	C	38619.29 0.002 (Model #25)	38619.57 0.002 (Model #32)	38619.78 0.002 (Model #39)
	E	38621.08 0.001 (Model #26)	38620.95 0.001 (Model #33)	38620.39 0.001 (Model #40)
	None	38628.58 2.23×10^{-5} (Model #12)[a]	38628.90 1.90×10^{-5} (Model #13)[a]	38628.00 2.98×10^{-5} (Model #19)[a]

[a] Models #12, #13, and #19 are part of Block #2 (see Table 3-4).

Table notes: AICcs are numbers greater than 30,000, whereas Akaike weights are proportions. "Age Moderation Effects" are those latent biometric factors the loadings of which were allowed to be moderated by age. "SES Moderation Effects" are those latent biometric factors the loadings of which were allowed to be moderated by SES.

Table 3-6. Model-fitting results of Block #4.

Model Number (Free Interaction Parameters)	AICc	Akaike Weight
Model #41 (γ_{A3}, γ_{C3})	38615.52	0.015
Model #42 (γ_{A3})	38615.43	0.016
Model #43 (γ_{C3})	38614.27	0.029
Model #27[a] (none)	38612.35	0.075

[a] Model #27 is part of Block #3 (see Table 3-5).

Table 3-7. Multimodel inference from Blocks #1 through #4.

Parameter (Symbol)	Point Estimate	95% CI	P-value
Full-Sib Genetic Correlation (r_A)	0.523	(0.414, 0.631)	0.2932[a]
Twin Effects (γ_T)	-2.169	(-6.212, 1.875)	0.6818
Age-Moderation, A (γ_{A1})	0.355	(0.063, 0.647)	0.0171
Age-Moderation, C (γ_{C1})	-1.398	(-2.258, -0.538)	0.0014
Age-Moderation, E (γ_{E1})	-0.073	(-0.167, 0.020)	0.1250
SES-Moderation, A (γ_{A2})	2.994	(1.019, 4.969)	0.0030
SES-Moderation, C (γ_{C2})	2.138	(-2.618, 6.894)	0.3783
SES-Moderation, E (γ_{E2})	0.948	(-0.223, 2.119)	0.1127
SES × Age Interaction, A (γ_{A3})	0.163	(-0.784, 1.110)	0.7354
SES × Age Interaction, C (γ_{C3})	-1.630	(-3.847, 0.587)	0.1495
SES main effect, adoptees ($\beta_{SES,A}$)	6.920	(1.579, 12.725)	0.0111
SES main effect, bio offspring ($\beta_{SES,B}$)	9.863	(3.414, 16.313)	0.0027

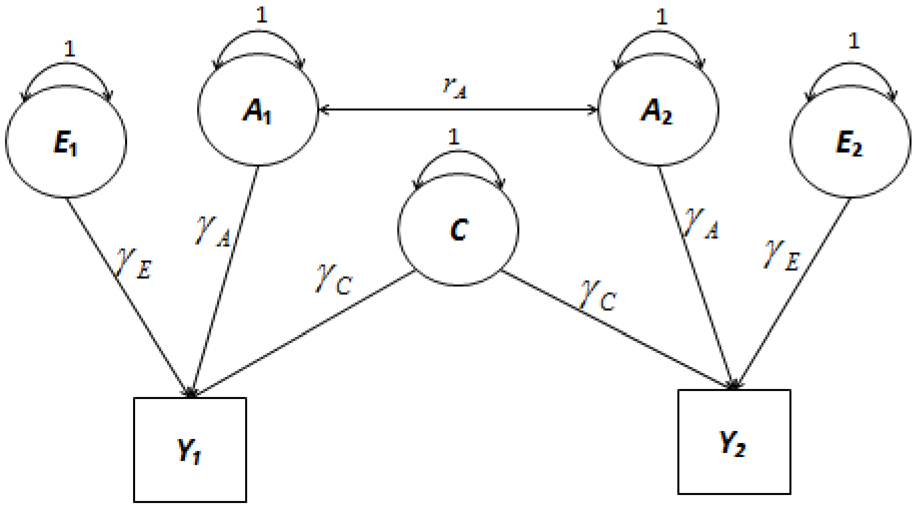
Table notes: Models #41, #42, and #43 are only included in calculating model-averaged inference for γ_{A3} and γ_{C3}, since the presence of those interaction terms and the mean-centering of age and SES changes the interpretation of the other moderation effects.

[a] Null parameter value for r_A is 0.5.

Table 3-8. Free Parameter Estimates and Standard Errors from Best-Approximating Model #36.

Parameter (Symbol)	Point Estimate	Standard Error
Intercept, Twins (β_0)	105.75	1.56
Intercept, Bio SIBS offspring (β_0)	109.86	1.84
Intercept, Adoptees (β_0)	114.19	2.70
Age, main effect (β_{Age})	-0.47	0.09
Sex, main effect (β_{Sex})	-3.91	0.44
SES, main effect, adoptees ($\beta_{SES,A}$)	7.08	2.83
SES, main effect, bio offspring ($\beta_{SES,B}$)	9.36	3.02
“Intercept” loading, A (γ_{A0})	8.65	0.63
“Intercept” loading, C (γ_{C0})	4.94	0.64
“Intercept” loading, E (γ_{E0})	6.08	0.33
SES moderation, A (γ_{A2})	2.81	0.86
SES moderation, E (γ_{E2})	0.91	0.58

Figure 3-1. Path diagram, classical twin model.



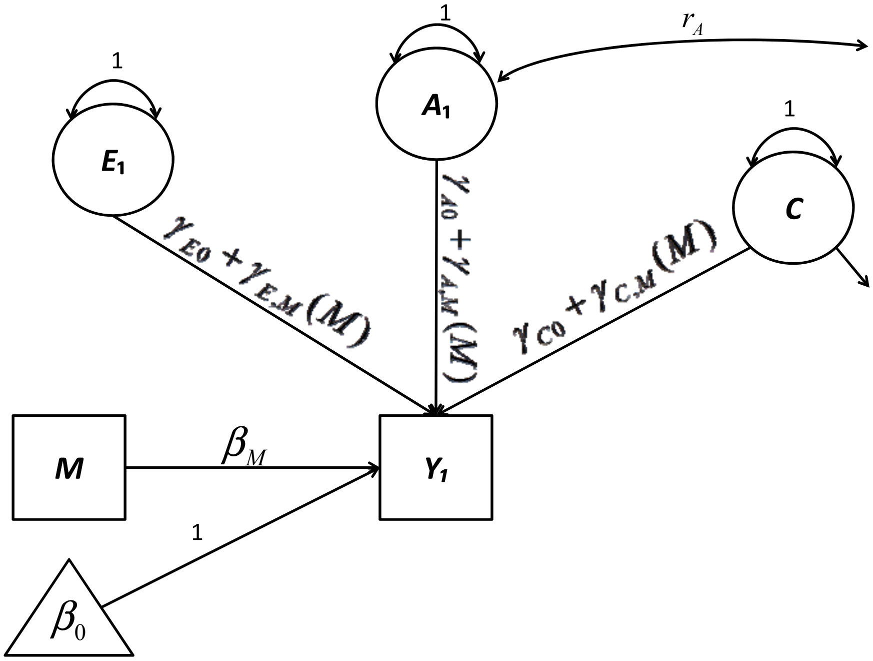
$$\mathrm{var}(Y_1) = \mathrm{var}(Y_2) = \gamma_A^2 + \gamma_C^2 + \gamma_E^2$$

$$\mathrm{cov}_{MZ}(Y_1, Y_2) = \gamma_A^2 + \gamma_C^2$$

$$\mathrm{cov}_{DZ}(Y_1, Y_2) = 0.5\gamma_A^2 + \gamma_C^2$$

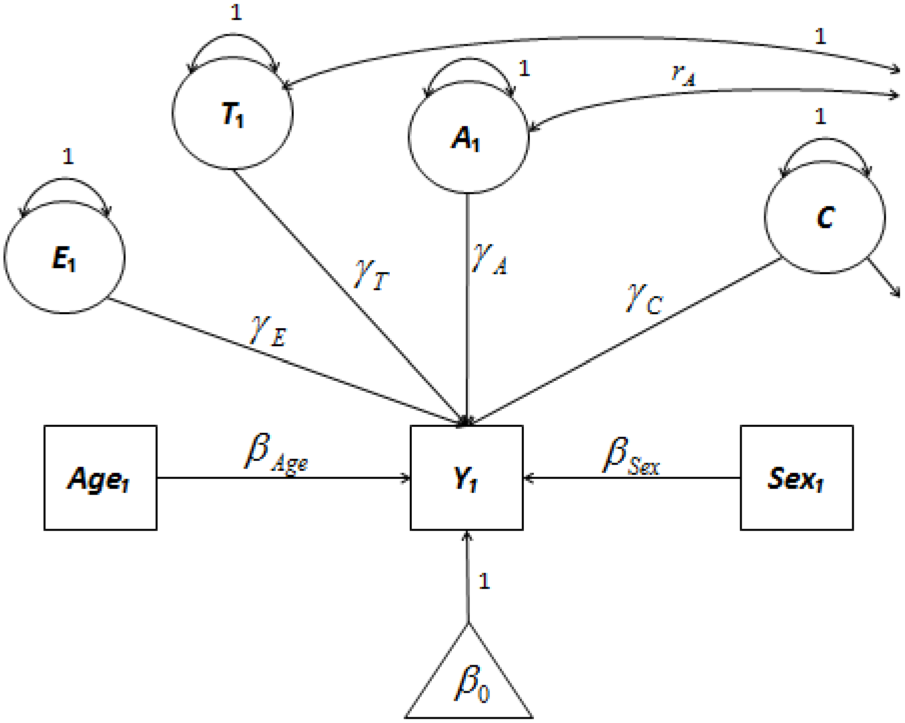
All variables are assumed to have zero mean, and latent variables (in circles) are further assumed to have unit variance. The covariance between A_1 and A_2, designated r_A, is fixed to 1 for MZ twins and to its expectation 0.5 for DZ twins, and thus the model is identified.

Figure 3-2. Purcell's (2002) continuous-moderator model.


1
1
r_A
A_1
1
E_1
C
$\gamma_{E0} + \gamma_{E,M}(M)$
$\gamma_{A0} + \gamma_{A,M}(M)$
$\gamma_{C0} + \gamma_{C,M}(M)$
M
β_M
Y_1
1
β_0

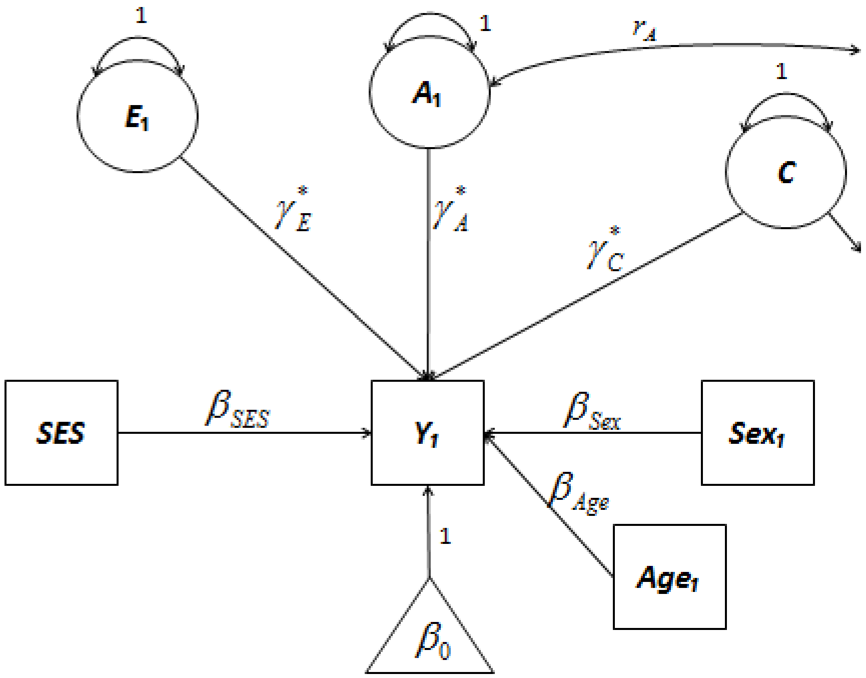
Only twin #1's side of the diagram is shown. The arrows that "lead to nowhere" from A_1 and *C* are assumed to connect to nodes in twin #2's side. For ease of presentation, the single-headed arrow representing variance in *M* is not shown. Now, Y_1 (and Y_2) are not assumed to be mean-centered; instead, its mean is estimated conditional on the moderator *M*.

Figure 3-3. Biometric ACE model with twin-specific effects.



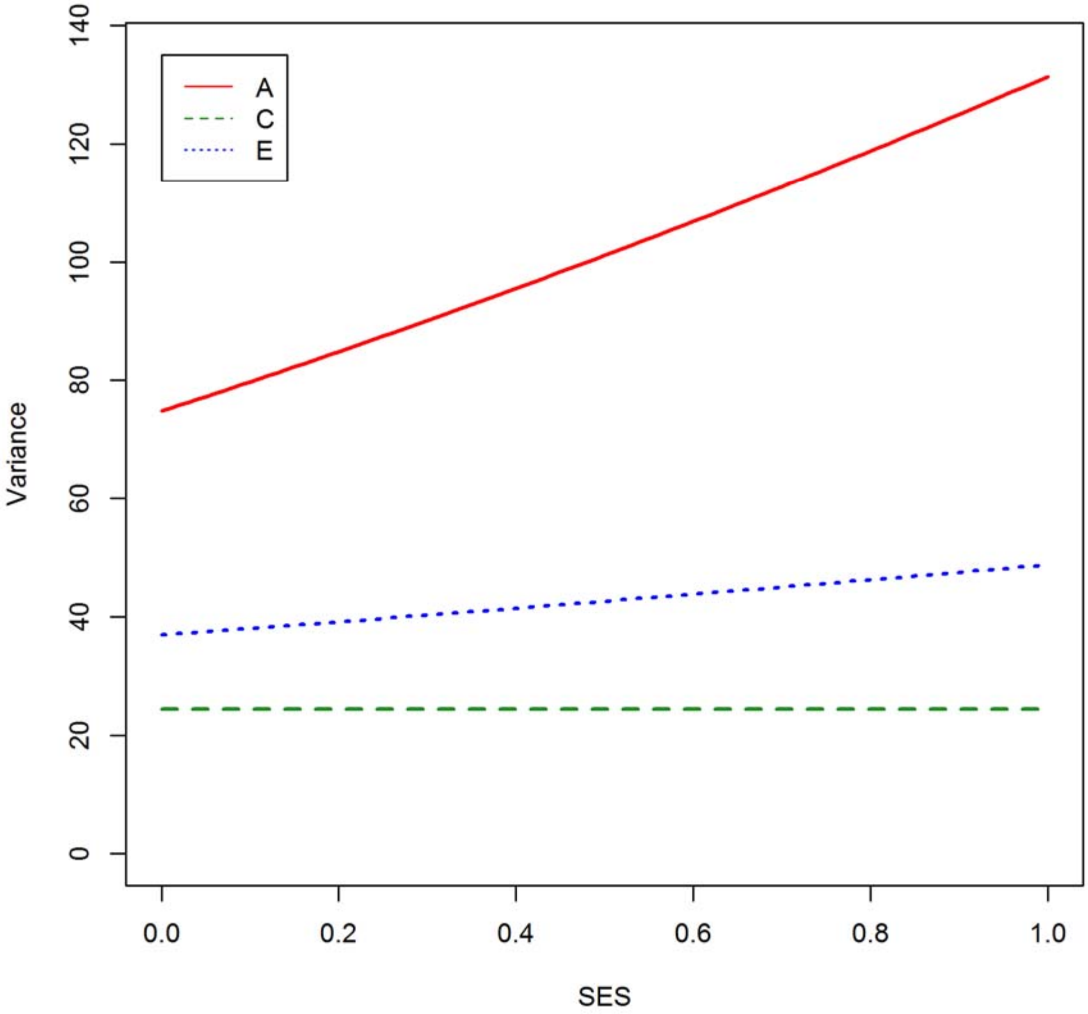
For ease of presentation, only twin #1's side of the diagram is shown, and the double-headed arrows representing variance in age, variance in sex, and their covariance are not included. Path coefficient γ_T is fixed to zero for non-twins.

Figure 3-4. Biometric moderation model—general form for Blocks #2 through #4.



For ease of presentation, only twin #1's side of the diagram is shown, and the double-headed arrows representing variances and covariances of age, sex, and SES are not included. $\gamma_A^* = \gamma_{A0} + \gamma_{A1}(Age_1) + \gamma_{A2}(SES) + \gamma_{A3}(SES \times Age_1)$, $\gamma_C^* = \gamma_{C0} + \gamma_{C1}(Age_1) + \gamma_{C2}(SES) + \gamma_{C3}(SES \times Age_1)$, and $\gamma_E^* = \gamma_{E0} + \gamma_{E1}(Age_1) + \gamma_{E2}(SES)$. Separate values of β_0 were estimated for twins, biological SIBS offspring, and adoptees. Separate values of β_{SES} were estimated for biological offspring of parents (including twins) and for adoptees.

Figure 3-5. Biometric variance components as function of SES, based on estimates from best-approximating Model #36.



Conclusion

This dissertation reports the results of three studies of the genetics of general cognitive ability. The first of these was a genome-wide association study, plus three other analyses that leverage genome-wide SNP data: gene-based association tests, polygenic scoring with cross-validation, and *GCTA*. No association signal from any SNP or gene reached genome-wide significance, though the GWAS SNP test statistics were overall slightly inflated relative to the null expectation. When it came to actually applying SNP weights to cross-validated prediction, the actual weights performed about as well as unit weights having the same signs. Overall predictive accuracy was poor—around 0.7% of variance at best, which occurred with very inclusive *p*-value thresholds for calculating the polygenic score. But, *GCTA* showed that with the GREML method, all the genotyped SNPs put together can account for at least 35% of phenotypic variance, and possibly as much as 77%, though the interpretation of the estimate becomes muddled as more-closely related individuals become included in analysis. The GWAS and polygenic scoring results hint at the presence of a great many SNP effects too small to be estimated reliably in samples of this size. But *GCTA* reveals that all the observed SNPs can be combined additively to explain a nontrivial portion of phenotypic variance. General cognitive ability is indeed heritable and highly polygenic.

The standardized variance component estimated with *GCTA* at low genetic-relatedness ceilings is of magnitude similar to those from other studies. Such estimates fall short of biometrically estimated heritability. CNVs, which are a different class of polymorphism from SNPs, are one potential molecular basis for the "missing heritability" (Maher, 2008). Study #2 is the largest study of CNVs and cognitive ability to date. We hypothesized that CNV mutational burden would be negatively associated with GCA, but none of our eight operationalizations of mutational burden showed significant association. Our power analysis showed that we had reasonable power to detect a specific CNV of moderate-to-large effect, so we also conducted three genome-wide association scans for CNVs, each using a different coding scheme for mutant state versus reference state at each locus. Though we observed a near-hit for a homozygous deletion on chromosome 14 starting at rs1950943, no association signal remained significant after correction for multiple testing. We conclude that CNVs, at least when called as is feasible from genome-wide SNP chips, are not likely to account for the missing heritability in general cognitive ability. We find it interesting that the average participant in our normal-range-IQ sample was homozygous for two deletion CNVs, and suggest that the phenotype may be more resistant to disruption by structural mutation than one might think.

Study #3 investigated how certain variables can moderate the biometric decomposition of general cognitive ability, in a sample of adolescent twins, full siblings, and adoptive siblings. It was first of all an attempt to replicate the Turkheimer effect, i.e. that with increasing family-of-origin SES, (primarily) genetic variance increases, and (secondarily) shared-environmental variance decreases. Our data allowed us to verify that certain sources of spurious gene × environment interaction were not present. We fit biometric models to our data and compared their sample-size-corrected AICs, first to resolve basic questions of specification, then to attempt to confirm well-replicated age-moderation of the biometric decomposition, then to attempt to replicate the Turkheimer effect, and finally to explore possible moderation effects due to an age × SES interaction. Once model-fitting was complete, we based our inference about parameters on model-averaged point estimates and standard errors. We found no evidence of SES-dependent assortative mating, twin-specific environmental effects, or differential heritability or shared-environmentality by trait level; we concluded that full siblings (including DZ twins) do indeed share about 50% of their trait-relevant alleles identical by state. We replicated both age-moderation effects, but only the primary element of the Turkheimer effect, i.e. increasing genetic variance with increasing SES. We found no evidence of age-dependent moderation effects of SES.

Study #3 provided standardized, age- and sex-corrected estimates of $a^2 = 0.61$ and $c^2 = 0.16$, which together sum to 0.77, the highest estimated value of h^2_{SNP} from GCTA. It would appear that Yang, Lee, Goddard, and Visscher (2011) were exactly right in their warning that including close relatives in a $GCTA$ analysis confounds the shared environment with aggregate SNP effects. But even at the genetic-relatedness ceiling that Yang et al. recommend, 0.025, we obtained an estimate of $h^2_{SNP} = 0.35$, in the same range as estimates reported in other studies. Again, we cannot overstate the theoretical gravity of *GCTA*'s results. Now, molecular genetics and quantitative genetics agree: there can be no doubt that cognitive ability is substantially heritable.

In spite of its nontrivial heritability, no SNP or gene-based set of SNPs has been found to be associated with the phenotype at genome-wide significance thresholds, even in large studies by multi-site consortia (e.g., Benyamin et al., 2013). It would seem that the trait-relevant SNPs must be Lilliputian in effect size but legion in number. General cognitive ability is also a massively polygenic trait.

Lower-bound heritability estimates computed from classically unrelated individuals via *GCTA* tend to be smaller than those computed biometrically. It is not clear what the molecular basis for this "missing heritability" might be. But, our results and those of recent studies (MacLeod et al., 2012; Bagshaw et al., 2013; McRae et al., 2013) show that CNVs (at least when called from genome-wide SNP data) are a dead end in this line of research.

When discussing the genetics of GCA, one must be careful not to neglect the environment. Our third study shows that the environment matters. The shared environment had a modest but nonzero influence on adolescent GCA, which declines as young people grow up. Further, an environmental variable, family SES, had a moderate, significant direct effect (in the statistical sense) among adoptees, indicating that the association is truly environmental in nature.

More interestingly, SES had a biometric moderation effect—specifically, genetic variance is greater at higher SES levels. Therefore, the overall conclusion of these three studies is that the heritability of general cognitive ability is substantial, is massively polygenic in nature, but is subject to moderation by age, SES, and other contextual variables.

References

Ackerman, P. L., & Lohman, D. F. (2003). Education and *g*. In H. Nyborg (Ed.), *The Scientific Study of General Intelligence: Tribute to Arthur R. Jensen* (pp. 275-292). New York: Pergamon.

American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.). Washington DC: Author.

Asbury, K., Wachs, T. D., & Plomin, R. (2005). Environmental moderators of genetic influence on verbal and nonverbal abilities in early childhood. *Intelligence, 33,* 643-661. doi:10.1016/j.intell.2005.03.008

Azen, R., & Budescu, D. V. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological Methods, 8*(2), 129-148. doi:10.1037/1082-989X.8.2.129

Bagshaw, A. T. M., Horwood, L. J., Liu, Y., Fergusson, D. M., Sullivan, P. F., & Kennedy, M. A. (2013). No effect of genome-wide copy number variation on measures of intelligence in a New Zealand birth cohort. *PLoS ONE, 8*(1), e55208. doi:10.1371/journal.pone.0055208

Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics, 7,* 781-791. doi:10.1038/nrg1916

Banks, G. C., Batchelor, J. H., & McDaniel, M. A. (2010). Smarter people are (a bit) more symmetrical: A meta-analysis of the relationship between intelligence and fluctuating asymmetry. *Intelligence, 38*, 393-401. doi:10.1016/j.intell.2010.04.003

Bartels, L. M. (1997). Specification uncertainty and model averaging. *American Journal of Political Science, 41*(2), 641-674.

Benjamin, D. J., Cesarini, D., van der Loos, M. J. H. M., Dawes, C. T., Koellinger, P. D., Magnusson, P. K. E., … Visscher, P. M. (2013). The genetic architecture of economic and political preferences. *PNAS, 109*(21), 8026-8031. doi: 10.1073/pnas.1120666109

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological), 57*(1), 289-300.

Benyamin, B., St Pourcaine, B., Davis, O. S., Davies, G., Hansell, N. K., Brion, M-J. A., … Visscher, P. M. (2013). Childhood intelligence is heritable, highly polygenic and associated with *FNBP1L*. *Molecular Psychiatry.* doi:10.1038/mp.2012.184

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., …Fox, J. (2011). OpenMx: An open source extended structural equation modeling framework. *Psychometrika, 76*(2), 306-317. doi: 10.1007/S11336-010-9200-6 . Software and documentation available at http://openmx.psyc.virginia.edu/ .

Boomsma, D. I., de Geus, E. J. C., van Baal, G. C. M., & Koopmans, J. R. (1999). A religious upbringing reduces the influence of genetic factors on disinhibition: Evidence for interaction between genotype and environment on personality. *Twin Research, 2,* 115-125.

Boring, E.G. (1923). Intelligence as the tests test it. *New Republic, 36*, 35-37.

Bouchard, T. J. (2004). Genetic influence on human psychological traits: A survey. *Current Directions in Psychological Science, 13*(4), 148-151.

Bouchard, T. J., & McGue, M. (1981). Familial studies of intelligence: A review. *Science, 212*(4498), 1055-1059.

Bouchard, T. J., & McGue, M. (2003). Genetic and environmental influences on human psychological differences. *Journal of Neurobiology, 54,* 4-45.

Braveman, P. A., Cubbin, C., Egerter, S., Chideya, S., Marchi, K. S., Metzler, M., & Posner, S. (2005). Socioeconomic status in health research: One size does not fit all. *Journal of the American Medical Association, 294*(22), 2879-2888.

Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *Journal of the American Statistical Association, 87*(419), 738-754.

Bronfenbrenner, U., & Ceci, S. J. (1994). Nature-nurture reconceptualized in developmental perspective: A bioecological model. *Psychological Review, 101*(4), 568-586.

Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology, 44,* 108-132. doi:10.1006_jmps.1999.1279

Browning, B. L., & Browning, S. R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *The American Journal of Human Genetics, 84,* 210-223. doi: 10.1016/j.ajhg.2009.01.005

Burnham, K. P., & Anderson, D. R. (2001). Kullback-Leibler information as a basis for strong inference in ecological studies. *Wildlife Research, 28,* 111-119.

Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach (2nd ed.).* New York: Springer.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research, 33*(2), 261-304. doi:10.1177/0049124104268644

Buse, A. (1973). Goodness of fit in generalized least squares estimation. *The American Statistician, 27*(3), 106-108.

Butcher, L. M., Davis, O. S. P., Craig, I. W., & Plomin, R. (2008). Genome-wide quantitative trait locus association scan of general cognitive ability using pooled DNA and 500K single nucleotide polymorphism microarrays. *Genes, Brain and Behavior, 7,* 435-446. doi: 10.1111/j.1601-183X.2007.00368.x

Campbell, C. D., Ogburn, E. L., Lunetta, K. L., Lyon, H. N., Freedman, M. L., Groop, L. C., … Hirschorn, J. N. (2005). Demonstrating stratification in a European American population. *Nature Genetics, 8*(37), 868-872. doi: 10.1038/ng1607

Carroll, J. B. (1993). *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. New York: Cambridge University Press.

Chabris, C. F., Hebert, B. M., Benjamin, D. J., Beauchamp, J., Cesarini, D., van der Loos, M., … Laibson, D. (2012). Most reported genetic associations with general intelligence are probably false positives. *Psychological Science,23*(11), 1314-1323. doi: 10.1177/0956797611435528

Cherny, S. S., Cardon, L. R., Fulker, D. W., & DeFries, J. C. (1992). Differential heritability across levels of cognitive ability. *Behavior Genetics, 22*(2), 153-162.

Coe, B. P., Girirajan, S., Eichler, E. E. (2012). The genetic variability and commonality of neurodevelopmental disease. *American Journal of Medical Genetics Part C (Seminars in Medical Genetics), 160C*, 118-129. doi:10.1002/ajmg.c.31327

Colella, S., Yau, C., Taylor, J. M., Mirza, G., Butler, H., Clouston, P., … Ragoussis, J. (2007). QuantiSNP: An Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Research, 35*(6), 2013-2025. doi:10.1093/nar/gkm076

Cooper, G. M., & Mefford, H. C. (2011). Detection of copy number variation using SNP genotyping. In Schwarz, P. H., & Wesselschmidt, R. L., editors. *Human Pluripotent Stem Cells: Methods and Protocols.* New York: Springer Science+Business Media, LLC. p. 243-252. doi:10.1007/978-1-61779-201-4_18

Coward, W. M., & Sackett, P. R. (1990). Linearity of ability-performance relationships: A reconfirmation. *Journal of Applied Psychology, 75*(3), 297-300.

Cullen, M. J., Hardison, C. M., & Sackett, P. R. (2004). Using SAT-grade and ability-job performance relationships to test predictions derived from stereotype threat theory. *Journal of Applied Psychology, 89*(2), 220-230. doi: 10.1037/0021-9010.89.2.220

Dana, J., & Dawes, R. M. (2004). The superiority of simple alternatives to regression for social science predictions. *Journal of Educational and Behavioral Statistics, 29*(3), 317-331.

Daniels, J., Holmans, P., Williams, N., Turic, D., McGuffin, P., Plomin, R., & Owen, M. J., et al. (1998). A simple method for analyzing microsatellite allele image patters generated from DNA pools and its application to allelic association studies. *American Journal of Human Genetics, 62,* 1189-1197.

Davies, G., Tenesa, A., Payton, A., Yang, J., Harris, S. E., Liewald, D., … Deary, I. J. (2011). Genome-wide association studies establish that human intelligence is highly heritable and polygenic. *Molecular Psychiatry 16*(10), 996-1005. doi:10.1038/mp.2011.85

Davis, O. S. P., Butcher, L. M., Docherty, S. J., Meaburn, E. L., Curtis, C. J. C., … Plomin, R. (2010). A three-stage genome-wide association study of general cognitive ability: Hunting the small effects. *Behavior Genetics, 40,* 759-767. doi:10.1007/s10519-010-9350-4

Davis, O. S. P., Haworth, C. M. A., & Plomin, R. (2009). Dramatic increase in heritability of cognitive development from early to middle childhood: An 8-year longitudinal study of 8,700 pairs of twins. *Psychological Science, 20,* 1301-1308.

Deary, I. J. (2012). Intelligence. *Annual Review of Psychology, 63*, 453-482. doi: 10.1146/annurev-psych-120710-100353

Deary, I. J., Spinath, F. M., & Bates, T. C. (2006). Genetics of intelligence. *European Journal of Human Genetics, 14*, 690-700. doi:10.1038/sj.ejhg.5201588

Deary, I. J., Johnson, W., & Houlihan, L. M. (2009). Genetic foundations of human intelligence. *Human Genetics, 126,* 215-232. doi: 10.1007/s00439-009-0655-4

DeFries, J. C., & Fulker, D. W. (1985). Multiple regression analysis of twin data. *Behavior Genetics, 15*(5), 467-473.

DeFries, J. C., & Fulker, D. W. (1988). Multiple regression analysis of twin data: Etiology of deviant scores versus individual differences. *Acta Geneticae Medicae et Gemellologiae, 37*, 205-216.

Detterman, D. K., Thompson, L. A., & Plomin, R. (1990). Differences in heritability across groups differing in ability. *Behavior Genetics, 20*(3), 369-384.

Dick, D. M., Aliev, F., Bierut, L., Goate, A., Rice, J., et al. (2006). Linkage analyses of IQ in the Collaborative Study on the Genetics of Alcoholism (COGA) sample. *Behavior Genetics, 36*(1), 77-86. doi: 10.1007/s10519-005-9009-8

Ellis, L., & Walsh, A. (2003). Crime, delinquency and intelligence: A review of the worldwide literature. In H. Nyborg (Ed.), *The Scientific Study of General Intelligence: Tribute to Arthur R. Jensen* (pp. 343-365). New York: Pergamon.

Evans, G. W. (2004). The environment of childhood poverty. *American Psychologist, 59*(2), 77-92. doi:10.1037/0003-066X.59.2.77

Fischbein, S. (1980). IQ and social class. *Intelligence, 4,* 51-63.

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh, 52,* 399-433.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95*(1), 29-51.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101*(2), 171-191.

Galton, F. (1869). *Hereditary Genius: An Inquiry into its Laws and Consequences.* London: MacMillan & Co. Retrieved from http://galton.org/ .

Gangestad, S. W. (2010). Evolutionary biology looks at behavior genetics. *Personality and Individual Differences, 49*, 289-295. doi:10.1016/j.paid.2010.03.005

Gangestad, S. W, & Thornhill, R. (1999). Individual differences in developmental precision and fluctuating asymmetry: a model and its implications. *Journal of Evolutionary Biology, 12*, 402-416.

Gauderman, W. J, & Morrison, J. M. (2006). QUANTO 1.1: A computer program for power and sample size calculations for genetic-epidemiology studies [software and manual]. Available at http://hydra.usc.edu/gxe/ .

Gläscher, J., Rudrauf, D., Colom, R., Paul, L. K., Tranel, D., et al. (2010). Distributed neural system for general intelligence revealed by lesion mapping. *PNAS, 107*(10), 4705-4709. doi:10.1073/pnas.0910397107

Glessner, J. T., Connolly, J. J. M., & Hakonarson, H. (2012). Rare genomic deletions and duplications and their role in neurodevelopmental disorders. *Current Topics in Behavioral Neuroscience, 12*, 345-360. doi:10.1007/7854_2011_179

Gottfredson, L.S. (1997a). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence, 24*(1), 13-23. (Original work published 1994).

Gottfredson, L. S. (1997b). Why *g* matters: The complexity of everyday life. *Intelligence, 24*(1), 79-132.

Gottfredson, L. S. (2003). *g*, Jobs, and Life. In H. Nyborg (Ed.), *The Scientific Study of General Intelligence: Tribute to Arthur R. Jensen* (pp. 293-342). New York: Pergamon.

Gottfredson, L. S., & Deary, I. J. (2004). Intelligence predicts health and longevity, but why? *Current Directions in Psychological Science, 13*(1), 1-4.

Gould, S. J. (1996). *The Mismeasure of Man*. New York: W. W. Norton & Company, Inc.

Grant, M. D., Kremen, W. S., Jacobson, K. C., Franz, C., Xian, H., Eisen, S. A., … Lyons, M. J. (2010). Does parental education have a moderating effect on the genetic and environmental influences of general cognitive ability in early adulthood? *Behavior Genetics, 40,* 438-446. doi:10.1007/s10519-010-9351-3

Hanscombe, K. B., Trzaskowski, M., Haworth, C. M. A., Davis, O. S. P., Dale, P. S., & Plomin, R. (2012). Socioeconomic status (SES) and children's intelligence (IQ): In a UK-representative sample SES moderates the environmental, not genetic, effect on IQ. *PLoS ONE, 7*(2), e30320. doi:10.1371/journal.pone.0030320

Harden, K. P., Turkheimer, E., & Loehlin, J. C. (2007). Genotype by environment interaction in adolescents' cognitive aptitude. *Behavior Genetics, 37,* 273-283. doi:10.1007/s10519-006-9113-4

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.).* New York: Springer Science+Business Media, LLC. doi:10.1007/b94608

Haworth, C. M. A., Wright, M. J., Luciano, M., Martin, N. G., de Geus, E. J. C., van Beijsterveldt, C. E. M., … Plomin, R. (2010). The heritability of general cognitive ability increases linearly from childhood to young adulthood. *Molecular Psychiatry, 15,* 1112-1120. doi:10.1038/mp.2009.55

Hawrylycz, M. J., Lein, E. S., Guillozet-Bongaarts, A. L., Shen, E. H., Ng, L., Miller, J. A., … Jones, A. R. (2012). An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature, 489,* 391-399. doi:10.1038/nature11405

Herrnstein, R. J., & Murray, C. (1994). *The Bell Curve: Intelligence and Class Structure in American Life*. New York: Simon & Schuster, Inc.

Hirschorn, J. N., Lohmueller, K., Byrne, E., & Hirschorn, K. (2002). A comprehensive review of genetic association studies. *Genetics in Medicine, 4*(2), 45-61.

Hollingshead, A. B. (1957). *Two Factor Index of Social Position*. New Haven, CN: August B. Hollingshead.

Hurvich, C. M., & Tsai, C-L. (1989). Regression and time series model selection in small samples. *Biometrika, 76*(2), 297-307.

Iacono, W. G., Carlson, S. R., Taylor, J., Elkins, I. J., & McGue, M. (1999). Behavioral disinhibition and the development of substance-use disorders: Findings from the Minnesota Twin Family Study. *Development and Psychopathology*, *11*, 869-900.

Iacono, W. G,, McGue, M. (2002). Minnesota Twin Family Study. *Twin Research, 5*(5), 482-487.

The International HapMap Consortium. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature, 449,* 851-862. doi: 10.1038/nature06258

The International Schizophrenia Consortium. (2008). Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature, 455,* 237-241. doi:10.1038/nature07239

The International Schizophrenia Consortium. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature, 460,* 748-752. doi: 10.1038/nature08185

Ioannidis, J. P. A., Ntzani, E. E., Trikalinos, T. A., & Contopoulos-Ioannidis, D. G. (2001). Replication validity of genetic association studies. *Nature Genetics, 29,* 306-309. doi: 10.1038/ng749

Jensen, A. R. (1980). *Bias in Mental Testing*. New York: The Free Press.

Jensen, A. R. (1998). *The* g *Factor: The Science of Mental Ability*. London: Praeger.

Jinks, J. L., & Fulker, D. W. (1970). Comparison of the biometrical genetical, MAVA, and classical approaches to the analysis of human behavior. *Psychological Bulletin, 73*(5), 311-349.

Johnson, W., & Bouchard, T. J., Jr. (2005). The structure of human intelligence: It is verbal, perceptual, and image rotation (VPR), not fluid and crystallized. *Intelligence, 33,* 393-416. doi:10.1016/j.intell.2004.12.002

Johnson, W., Bouchard, T. J., Jr., Krueger, R. F., McGue, M., & Gottesman, I. I. (2004). Just one *g*: Consistent results from three test batteries. *Intelligence, 32*, 95-107. doi:10.1016/S0160-2896(03)00062-X

Johnson, W., te Nijenhuis, J., & Bouchard, T. J. (2007). Replication of the hierarchical visual-perceptual-image rotation model in de Wolff and Buiten's (1963) battery of 46 tests of mental ability. *Intelligence, 35,* 69-81. doi:10.1016/j.intell.2006.05.002

Johnson, W., te Nijenhuis, J., & Bouchard, T. J. (2008). Still just 1 *g*: Consistent results from five test batteries. *Intelligence, 36*, 81-95. doi:10.1016/j.intell.2007.06.001

Jung, R. E., & Haier, R. J. (2007). The parieto-frontal integration theory of intelligence: Converging neuroimaging evidence. *Behavioral & Brain Sciences, 30,* 135-187. doi:10.1017/S0140525X07001185

Kapetanios, G., Labhard, V., & Price, S. (2008). Forecasting using Bayesian and information-theoretic model-averaging: An application to U.K. inflation. *Journal of Business & Economic Statistics, 26*(1), 33-41. doi:10.1198/073500107000000232

Keyes, M. A., Malone, S. M., Elkins, I. J., Legrand, L. N., McGue, M., & Iacono, W. G. (2009). The Enrichment Study of the Minnesota Twin Family Study: Increasing the yield of twin families at high risk for externalizing psychopathology. *Twin Research and Human Genetics, 12*(5), 489-501.

Kirkpatrick, R. M., McGue, M., & Iacono, W. G. (2009). Shared-environmental contributions to high cognitive ability. *Behavior Genetics, 39*, 406-416. doi:10.1007/s10519-009-9265-0

Kohler, H.P., & Rodgers, J.L. (2001). DF-analyses of heritability with double-entry twin data: Asymptotic standard errors and efficient estimation. *Behavior Genetics, 31*(2), 179-191.

Kremen, W. S., Jacobson, K. C., Xian, H., Eisen, S. A., Waterman, B., Toomey, R., … Lyons, M. J. (2005). Heritability of word recognition in middle-aged men varies as a function of parental education. *Behavior Genetics, 35*(4), 417-433. doi:10.1007/s10519-004-3876-2

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics, 22*(1), 79-86.

Kuncel, N. R., & Hezlett, S. A. (2007). Standardized tests predict graduate students' success. *Science, 315*, 1080-1081. doi: 10.1126/science.1136618

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology, 86*(1), 148-161. doi: 10.1037/0022-3514.86.1.148

Lango Allen, H., Estrada, K., Lettre, G., Berndt, S. I.,Weedon, M. N., Rivadeneira, F., et al., on behalf of the GIANT Consortium. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature, 467,* 832-838. doi: 10.1038/nature09410.

Lee, T., Henry, J. D., Trollor, J. N., & Sachdev, P. S. (2010). Genetic influences on cognitive functions in the elderly: A selective review of twin studies. *Brain Research Reviews, 64,* 1-13.

Li, X., Basu, S., Miller, M. B., Iacono, W. G., & McGue, M. (2011). A rapid generalized least squares model for genome-wide quantitative trait association analysis. *Human Heredity, 71*, 67-82. Package and manual available at http://www.cran.r-project.org/web/packages/RFGLS/ .

Li, Y., Willer, C. J., Ding, J., Scheet, P., & Abecasis, G. R. (2010). *MaCH*: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology, 34,*816-834. doi: 10.1002/gepi.20533. *Minimac* software and documentation available at http://genome.sph.umich.edu/wiki/Minimac).

Liu, J. Z., Mcrae, A. F., Nyholt, D. R., Medland, S. E., Wray, N. R., Brown, K. M., … Macgregor, S.. (2010). A versatile gene-based test for genome-wide association studies. *The American Journal of Human Genetics, 87,* 139-145. doi:10.1016/j.ajhg.2010.06.009

Loehlin, J. C., Harden, K. P., & Turkheimer, E. (2009). The effect of assumptions about parental assortative mating and genotype-income correlation on estimates of genotype-environment interaction in the National Merit Twin Study. *Behavior Genetics, 39,* 165-169. doi:10.1007/s10519-008-9253-9

Luciano, M., Wright, M. J., Duffy, D. L., Wainwright, M. A., Zhu, G., Evans, D. M., et al. (2006). Genome-wide scan of IQ finds significant linkage to quantitative trait locus on 2q. *Behavior Genetics, 36*(1), 45-55. doi: 10.1007/s10519-005-9003-1

Lukacs, P. M., Burnham, K. P., & Anderson, D. R. (2009). Model selection bias and Freedman's paradox. *Annals of the Institute of Statistical Mathematics, 62,* 117-125. doi:10.1007/s10463-009-0234-4

MacLeod, A. K., Davies, G., Payton, A., Tenesa, A., Harris, S. E., Liewald, D., … Deary, I. J. (2012). Genetic copy number variation and general cognitive ability. *PLoS One, 7*(12), e37385. doi:10.1371/journal.pone.0037385

Maher, B. (2008). The case of the missing heritability. *Nature, 456*(6), 18-21.

Maraun, M. D. (1996). Metaphor taken as math: Indeterminacy in the factor analysis model. *Multivariate Behavioral Research, 31*(4), 517-538.

Marchini, J., Cardon, L. R., Phillips, M. S., & Donnelly, P. (2004). The effects of human population structure on large genetic association studies. *Nature Genetics, 36*(5), 512-517. doi: 10.1038/ng1337

McCallum, R. C., & Mar, C. M. (1995). Distinguishing between moderator and quadratic effects in multiple regression. *Psychological Bulletin, 118*(3), 405-421.

McDaniel, M. A. (2005). Big-brain people are smarter: A meta-analysis of the relationship between in vivo brain volume and intelligence. *Intelligence, 33,* 337-346. doi:10.1016/j.intell.2004.11.005

McGue, M., & Bouchard, T. J. (1984). Adjustment of twin data for the effects of age and sex. *Behavior Genetics, 14*(4), 325-343.

McGue, M., Bouchard, T. J., Iacono, W. G., & Lykken, D. T. (1993). Behavioral genetics of cognitive ability: A life-span perspective. In Plomin, R., McClearn, G. E. (Eds.), *Nature, nurture and psychology* (pp. 59-76). Washington: American Psychological Association.

McGue, M., Keyes, M., Sharma, A., Elkins, I., Legrand, L., Johnson, W., & Iacono, W. G. (2007). The environments of adopted and non-adopted youth: Evidence on range restriction from the Sibling Interaction and Behavior Study (SIBS). *Behavior Genetics, 37*, 449-462. doi:0.1007/s10519-007-9142-7

McRae, A. F., Wright, M. J., Hanselle, N. K., Montgomery, G. W., & Martin, N. G. (2013). No association between general cognitive ability and rare copy number variation. *Behavior Genetics.* Advance online publication. doi:10.1007/s10519-013-9587-9

Meehl, P. E. (2006). The power of quantitative thinking. In N. G. Waller, L. J. Yonce, W. M. Grove, D. Faust, & M. F. Lenzenweger (Eds.), *A Paul Meehl Reader: Essays on the Practice of Scientific Psychology (p. 433-444).* Mahwah, NJ: Lawrence Erlbaum Associates. (Original address delivered May 23, 1998.)

Mefford, H. C., Batshaw, M. L., & Hoffman, E. P. (2012). Genomics, intellectual disability, and autism. *New England Journal of Medicine, 366*, 733-743.

Miller, M. B., Basu, S., Cunningham, J., Eskin, E., Malone, S. M., Oetting, W. S., … McGue, M. (2012). The Minnesota Center for Twin and Family Research Genome-Wide Association Study. *Twin Research & Human Genetics, 15*(6), 767-774.

Morgan, T. H. (1911). Random segregation versus coupling in Mendelian inheritance. *Science, 34*(873), 384.

Morton, N. E. (1955). Sequential tests for the detection of linkage. *American Journal of Human Genetics, 7,* 277-318.

Myrianthopolous, N. C., & French, K. S. (1968). An application of the U.S. Bureau of the Census socioeconomic index to a large, diversified patient population. *Social Science & Medicine, 2,* 283-299.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78*(3), 691-692.

Neale, M. C. (2009). Biometrical models in behavioral genetics. In Y-K Kim (Ed.), *Handbook of Behavior Genetics* (p. 15-33). New York: Springer Science+Business Media. doi:10.1007/978-0-387-76727-7_2

Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H. (2003). *Mx: Statistical Modeling.* VCU Box 900126, Richmond, VA 23298: Department of Psychiatry. 6th Edition. Software and manual available at http://www.vcu.edu/mx/ .

Neale, M. C., & Maes, H. H. M. (2004). Methodology for genetic studies of twins and families. Available from http://www.vipbg.vcu.edu/~vipbg/mx/book2004a.pdf .

Need, A. C., Ge, D., Weale, M. E., Maia, J., Feng, S., Heinzen, E. L., … Goldstein, D. B. (2009). A genome-wide investigation of SNPs and CNVs in schizophrenia. *PLoS Genetics, 5*(2), e1000373. doi:10.1371/journal.pgen.1000373

Neubauer, A. C., & Fink, A. (2009). Intelligence and neural efficiency. *Neuroscience and Biobehavioral Reviews, 33,* 1004-1023. doi:10.1016/j.neubiorev.2009.04.001

Nyholt, D. R. (2004). A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *American Journal of Human Genetics, 74*, 765-769.

Ott, J. (1999). *Analysis of Human Genetic Linkage.* Baltimore: The Johns Hopkins University Press.

Pan, Y., Wang, K-S., & Aragam, N. (2011). NTM and NR3C2 polymorphisms influencing intelligence: Family-based association studies. *Progress in Neuro-Psychopharmacology & Biological Psychiatry, 35,* 154-160. doi:10.1016/j.pnpbp.2010.10.016

Pankratz, N., Dumitriu, A., Hetrick, K. N., Sun, M., Latourelle, J. C., Wilk, J. B., … the PSG–PROGENI and GenePD Investigators, Coordinators and Molecular Genetic Laboratories. (2011). Copy number variation in familial Parkinson disease. *PLoS ONE, 6*(8), e20988. doi:10.1371/journal.pone.0020988

Payton, A. (2006). Investigating cognitive genetics and its implications for the treatment of cognitive deficit. *Genes, Brain and Behavior, 5*(Suppl. 1), 44-53. doi:10.1111/j.1601-183X.2006.00194.x

Petrill, S. A., Saudino, K., Cherny, S. S., Emde, R. N., Fulker, D. W., et al. (1998). Exploring the genetic and environmental etiology of high general cognitive ability in fourteen- to thirty-six-month-old twins. *Child Development, 69*(1), 68-74.

Pinto, D., Pagnamenta, A. T., Klei, L., Anney, R., Merico, D., Regan, R., … Betancur, C. (2010). Functional impact of global rare copy number variation in autism spectrum disorders. *Nature, 466,* 368-372. doi:10.1038/nature09146

Purcell, S. (2002). Variance components models for gene-environment interaction in twin analysis. *Twin Research, 5*(6), 554-571.

Plomin, R. (2003). Molecular genetics and *g*. In H. Nyborg (Ed.), *The Scientific Study of General Intelligence: Tribute to Arthur R. Jensen* (pp. 275-292). New York: Pergamon.

Plomin, R., DeFries, J. C., & Loehlin, J. C. (1977). Genotype-environment interaction and correlation in the analysis of human behavior. *Psychological Bulletin, 84*(2), 309-322.

Plomin, R., Fulker, D. W., Corley, R., & DeFries, J. C. (1997). Nature, nurture, and cognitive development from 1 to 16 years: A parent-offspring adoption study. *Psychological Science, 8*(6), 442-447.

Posthuma, D. (2009). Multivariate genetic analysis. In Y-K Kim (Ed.), *Handbook of Behavior Genetics* (p. 47-59). New York: Springer Science+Business Media. doi:10.1007/978-0-387-76727-7_4

Posthuma, D., Luciano, M., de Geus, E. J. C., Wright, M. J., Slagboom, P. E., et al. (2005). A genomewide scan for intelligence identifies quantitative trait loci on 2q and 6p. *American Journal of Human Genetics, 77,* 318-326.

Price, A. L, Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics, 38*(8), 904-909. doi:10.1038/ng1847

Price, T. S., & Jaffee, S. R. (2008). Effects of the family environment: Gene-environment interaction and passive gene-environment correlation. *Developmental Psychology, 44*(2), 305-315. doi:10.1037/0012-1649.44.2.305

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007). *PLINK*: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics, 81*, 559-575. doi:10.1086/519795. Software and documentation available at http://pngu.mgh.harvard.edu/~purcell/plink/ .

Rathouz, P. J., Van Hulle, C. A., Rodgers, J. L., Waldman, I. D., & Lahey, B. B. (2008). Specification, testing, and interpretation of gene-by-measured environment interaction models in the presence of gene-environment correlation. *Behavior Genetics, 38,* 301-315. doi:10.1007/s10519-008-9193-4

Risch, N., & Merikangas, M. (1996). The future of genetic studies of complex human diseases. *Science, 273*(5281), 1516-1517.

Rijsdijk, F. V., Vernon, P. A., & Boomsma, D. I. (2002). Application of hierarchical genetic models to Raven and WAIS subtests: A Dutch twin study. *Behavior Genetics, 32*(3), 199-210.

Rodgers, J. L., & Kohler, H. P. (2005). Reformulating and simplifying the DF analysis model. *Behavior Genetics, 35*(2), 211-217.

Rodgers, J. L., & McGue, M. (1994). A simple algebraic demonstration of the validity of Defries-Fulker analysis in unselected samples with multiple kinship levels. *Behavior Genetics, 24*(3), 259-262.

Rowe, D. C., Jacobson, K. C., & van den Oord, E. J. C. G. (1999). Genetic and environmental influences on vocabulary IQ: Parental educational level as moderator. *Child Development, 70*(5), 1151-1162.

Rushton, J. P., & Ankney, C. D. (2009). Whole brain size and general mental ability: A review. *International Journal of Neuroscience, 119,* 692-732. doi:10.1080/00207450802325843

Rubin, D. B. (1976). Inference and missing data. (1976). *Biometrika, 63*(3), 581-592.

Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2008). High-stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist, 63*(4), 215-227. doi: 10.1037/0003-066X.63.4.215

Salthouse, T. A. (2004). What and when of cognitive aging. *Current Directions in Psychological Science, 13*(4), 140-144.

Sattler JM. (1974). *Assessment of Children (Revised).* Philadelphia: W. B. Saunders Company.

Saudino, K. J., Plomin, R., Pedersen, N. L., & McClearn, G. E. (1994). The etiology of high and low cognitive ability during the second half of the life span. *Intelligence, 19,* 359-371.

Scarr, S. (1992). Developmental theories for the 1990s: Development and individual differences. *Child Development, 63,* 1-19.

Scarr-Salapatek, S. (1971). Race, social class, and IQ. *Science, 174*(4016), 1285-1295.

Scarr, S., & McCartney, K. (1983). How people make their own environments: A theory of genotype → environment effects. *Child Development, 54*, 424-443.

Scarr, S., & Weinberg, R. A. (1978). The influence of "family background" on intellectual attainment. *American Sociological Review, 43*(5), 674-692.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147-177. doi:10.1037//1082-989X.7.2.147

Schaie, K. W. (1994). The course of adult intellectual development. *American Psychologist, 49*(4), 304-313.

Scherer, S. W., Lee, C., Birney, E., Altshuler, D. M., Eichler E. E., … Feuk, L. (2007). Challenges and standards in integrating surveys of structural variation. *Nature Genetics, 39,* S7-S15.

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., … Wigler, M. (2007). Strong association of de novo copy number mutations with autism. *Science, 316*, 445-449. doi:10.1126/science.1138659

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., … Wigler, M. (2004). Large-scale copy number polymorphism in the human genome. *Science, 305*, 525-528. doi: 10.1126/science.1098918. Supporting online material: http://www.sciencemag.org/content/suppl/2004/07/22/305.5683.525.DC1.html.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica, 7,* 221-264.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society, Series B (Methodological), 39*(1), 44-47.

Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. *Intelligence, 35*, 401-426. doi:10.1016/j.intell.2006.09.004

Spearman, C. (1904). "General intelligence," objectively determined and measured. *The American Journal of Psychology, 15*(2), 201-292.

Spearman, C. (1927). *The Abilities of Man: Their Nature and Measurement.* London: MacMillan and Co.

Tucker-Drob, E. M., & Harden, K. P. (2012). Learning motivation mediates gene-by-socioeconomic status interaction on mathematics achievement in early childhood. *Learning and Individual Differences, 22,* 37-45.

Tucker-Drob, E. M., Harden, K. P., & Turkheimer, E. (2009). Combining nonlinear biometric and psychometric models of cognitive abilities. *Behavior Genetics, 39,* 461-471. doi:10.1007/s10519-009-9288-6

Tucker-Drob, E. M., Rhemtulla, M., Harden, K. P., Turkheimer, E., & Fask, D. (2011). Emergence of a gene × socioeconomic status interaction on infant mental ability between 10 months and 2 years. *Psychological Science, 22*(1), 125-133. doi:10.1177/0956797610392926

Turkheimer, E. (2011). Commentary: Variation and causation in the environment and genome. *International Journal of Epidemiology, 40,* 598-601.

Turkheimer, E., Haley, A., Waldron, M., D'Onofrio, B., & Gottesman, I. I. (2003). Socioeconomic status modifies heritability of IQ in young children. *Psychological Science, 14*(6), 623-628.

Uher, R., Dragomirecka, E., & Papezova, H. (2006). Use of socioeconomic status in health research. *Journal of the American Medical Association, 295*(15), 1770.

Van den Ooord, E. J. C. G., & Rowe, D. C. (1998). An examination of genotype-environment interactions for academic achievement in an U.S. National Longitudinal Survey. *Intelligence, 25*(3), 205-228.

Van der Sluis, S., Willemsen, G., de Geus, E. J. C., Boomsma, D. I., & Posthuma, D. (2008). Gene-environment interaction in adults' IQ scores: Measures of past and present environment. *Behavior Genetics, 38,* 348-360. doi:10.1007/s10519-008-9212-5

Van Soelen, I. L. C., Brouwer, R. M., van Leeuwen, M., Kahn, R. S., Hulshoff Pol, H. E., & Boomsma, D. I. (2011). Heritability of verbal and performance intelligence in a pediatric longitudinal sample. *Twin Research and Human Genetics, 14*(2), 119-128.

Visscher, P. M. (2008). Sizing up human height variation. *Nature Genetics, 40*(5), 488-489.

Visscher, P. M., Yang, J., & Goddard, M. E. (2010). A commentary on 'Common SNPs Explain a Large Proportion of the Heritability for Human Height' by Yang et al. (2010). *Twin Research and Human Genetics, 13*(6), 517-524.

Wang K, Li M, Hadley D, Liu R, Glessner J, et al. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research, 17,* 1665-1674. doi:10.1101/gr.6861907. Software and manual available at: http://www.openbioinformatics.org/penncnv/ .

Wechsler, D. (1939). *The Measurement of Adult Intelligence.* Baltimore: The Williams & Wilkins Company.

Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A, … Rothberg, J. M. (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature, 452*, 872-878. doi:10.1038/nature06884

Wickett, J. C., Vernon, P. A., & Lee, D. H. (2000). Relationships between factors of intelligence and brain volume. *Personality and Individual Differences, 29*, 1095-1122.

Wilson, R. S. (1983). The Louisville Twin Study: Developmental synchronies in behavior. *Child Development, 54*, 298-316.

Woodberry, K. A., Giuliano, A. J., Seidman, L. J. (2008). Premorbid IQ in schizophrenia: A meta-analytic review. *American Journal of Psychiatry, 165*, 579-587.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., … Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics, 42*(7), 565-569.

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: A tool for genome-wide complex trait analysis. *The American Journal of Human Genetics, 88,* 76-82. doi:10.1016/j.ajhg.2010.11.011

Yang, J., Weedon, M. H., Purcell, S., Lettre, G., Estrada, K., Willer, C. J., … the GIANT Consortium. (2011). Genomic inflation factors under polygenic inheritance. *European Journal of Human Genetics, 19*, 807-812. doi:10.1038/ejhg.2011.39

Yeo, R. A., Gangestad, S. W., Liu, J., Calhoun, V. D., & Hutchinson, K. E. (2011). Rare copy number deletions predict individual variation in intelligence. *PLoS ONE 6*(1): e16339. doi:10.1371/journal.pone.0016339

_

_