BGI Cognitive Genomics Lab: Prxoposal for Gene-Trait Association Study of g

Christopher C. Chang[*], Stephen D.H. Hsu[†], James J. Lee[‡], Laurent C.A.M. Tellier[§], Rui Yang[¶], Bowen Zhao[‖]

Contents

1 Introduction

The human brain is one of the most complex objects in the universe, and human cognition is the result of billions of years of evolution. Human-level capabilities in perception, thought,

[*]PhD in Mathematics, University of California San Diego; Project Manager, BGI Cognitive Genomics Lab

[†]PhD in Physics, University of California Berkeley; Professor of Theoretical Physics, University of Oregon

[‡]PhD in Psychology, Harvard University

[§]Masters (expected) in Bioinformatics, University of Copenhagen; Chief Data Officer, BGI Cognitive Genomics Lab

[¶]PhD (expected) in Psychology, Brown University; Director, Psychology Arm, BGI Cognitive Genomics Lab

[‖]Director, Bioinformatics Arm, BGI Cognitive Genomics Lab

language, and motor control are still far beyond the power of any manmade machine. Thus, although psychology, neuroscience, computer science, and related fields concerned with cognition are advancing rapidly, we are still only beginning to penetrate the mysteries of the human mind.

The brain evolved to deal with a complex, information-rich environment. The blueprint for the brain is contained in our DNA, although brain development is a complicated process in which interactions with the environment play an important role. Nevertheless, in almost all cases a significant portion of cognitive or behavioral variability in humans is found to be heritable—i.e., attributable to genetic causes.

The goal of the BGI Cognitive Genomics Lab (CGL) is to investigate the genetic architecture of human cognition: the genomic locations, allele frequencies, and average effects of the precise DNA variants affecting variability in perceptual and cognitive processes. This document outlines the CGL's proposal to investigate one trait in particular: **general intelligence** or **general mental ability**, often referred to as "g."

1.1 Scientific Significance of the Proposed Research

People with higher scores on mental tests tend to live longer, even after adjusting for socioeconomic status [**?** **?** **?**]. Possible mediating variables along a causal pathway between cognitive ability and mortality include several well-established risk factors for poor health; higher scores are associated with lower prevalence of obesity [**?** **?** **?**], hypertension [**?** **?** **?**], physical inactivity [**?**], poor diet [**?** **?**], psychiatric disorder [**?** **?** **?**], and smoking [**?** **?** **?**].

The deeper mechanisms connecting cognitive ability to a healthy lifestyle are not known with certainty. One clue is afforded by the time trend in the association of g with smoking. Although there was no significant difference in childhood test scores between Scottish adults born in 1921 who had *ever* smoked and *never* smoked, each score increase of one standard deviation was accompanied by a 33% increased rate of *quitting* smoking [**?**]. Thus, ability level measured in childhood was not associated with the initiation of smoking when the causal link between smoking and cancer was not yet known, but became associated with the cessation of smoking once a complex body of evidence had established the link as an objective fact. This example suggests that a higher level of g enhances health and longevity because the proper care of the self depends on many decisions that can be informed by reasoning and knowledge.

The literature on the link between g and health suggests that interventions to increase cognitive functioning may dramatically improve lifespan and well-being. In order to realize this possibility, however, we need a better understanding of g's underlying biological substrate. g is correlated with a number of neural variables, including overall brain vol-

ume, connectivity of white matter, and concentrations of N-acetyl aspartate [**?** **?** **?**], and future investigations may elucidate the mechanisms by which these aspects of brain structure affect g. A natural complement to this approach is to pinpoint the precise DNA variants constituting the genetic architecture of g, thus revealing the biological pathways responsible for the variability in this trait.

The genetic study of g is important from both applied and fundamental perspectives. That is, we should expect any actionable discoveries derived from reductionistic research on g to provide basic insights into brain function and development. In addition, whereas an individual's genome provides a partial blueprint for the development of the phenome forward in time, our species' array of genomic data provides a partial record of our evolutionary history backward in time. Thus, knowledge of the genetic variants affecting g may shed light not only on proximate biological mechanisms, but also the ultimate evolutionary forces that have shaped the cognitive capacities of humankind. These powerful motivations compel extending gene-trait association studies to g and other ability factors.

1.2 Theoretical Framework

Since ancient times people have observed that some of their fellows are more clever than others. Psychologists who study individual differences have been reasonably successful in quantifying this variation with the use of standardized instruments that go under various names, including "IQ" or "scholastic aptitude" tests. Scores on such tests are highly reliable [**?** **?** **?**], stable over the lifespan [**?**], correlated with several neural variables [**?** **?** **?**] and predictive of many important life outcomes (educational achievement, earnings, occupational prestige, workplace performance, health and mortality) [**?** **?** **?** **?** **?**].

An IQ or aptitude test is an aggregate of items eliciting responses that can be unambiguously scored as *right* or *wrong*. It is a remarkable fact that the responses to almost all such items, regardless of the specific skills or knowledge required, are positively correlated [**?** **?** **?** **?**]. As a consequence a sample of items provides information about how the examinees would have performed on the much greater number of items that were not administered to them. This is one reason why an overall IQ score can be claimed to represent a valid measure of a person's intelligence: under certain mathematical conditions that are reasonably well satisfied by actual mental tests, an examinee's observed score on a test of increasing length approaches the score that would have obtained on an infinitely long test covering all subdomains of logical, factual, and semantic knowledge [**?** **?** **?**]. In order to generalize beyond the score obtained on a particular test to mastery of this idealized wider domain, differential psychologists refer to the random variable mapping into the latter as "the g factor" [**?** **?** **?** **?**].

The body of quantitative techniques used to construct and employ instruments for

measuring g (and other abstractive psychological traits) is known as **psychometrics** [**?** **?**]. Some texts on psychometrics treat the subject as a collection of unrelated topics. It can be shown, however, that the technique of **factor analysis** supplies a rigorous and unified treatment of the major concepts in psychometrics [**?** **?** **?**]. Factor-analytic models treat measured variables, such as the different items or tests in an IQ battery, as indicators of unmeasured quantitative variables called **common factors**. g is perhaps the prototypical example of a common factor. If the scores on a test could be regressed on the unobserved common factors, each regression coefficient would represent the sensitivity of the test as a measure of the corresponding factor. The regression coefficients in this linear model are called **factor loadings**. The application of a linear unidimensional factor model to individual test items yields **classical test theory** [**?** **?**], whereas a mild nonlinearization leads to **item response theory** [**?** **?** **?**]. Factor analysis thus provides the underpinnings of the psychological, educational, and military assessments taken by millions of examinees each year.

Some opponents of standardized testing have criticized psychometricians for "reifying" common factors [**?** **?**]. According to these opponents, psychometricians take a common factor to be a single physical "thing in the head," in principle directly observable given further advances in neuroscientific technique, manipulations of which will lead to changes in a person's observed test scores that are proportional to their factor loadings. In reality, however, no sober psychometrician believes this. Instead, common factors such as intelligence (g), extraversion, religiosity, political liberalism, and so on are regarded as powerful compressions of an enormously complex system.[1] They are abstractions of actual and potential behavior [**?** **?** **?**]—much like "sprinting ability," for example. No one supposes there to be single physical thing in the human body corresponding to sprinting ability; such a belief would amount to what philosophers call a **category mistake** [**?**]. Nevertheless it makes perfect sense to believe that there might be genetic variants affecting sprinting

[1]One psychometrician has responded to the allegation that his discipline interprets a common factor as an as-yet unobserved cause of the indicators (items or tests) used to measure it:

> [A] cause is defined independently of its effects. We could never "name"/"interpret," that is, guess the nature of an omitted cause and factor analysts do not do so. One way to say this is that causes do not "resemble" their effects, and so cannot be guessed from them. Electroshock has no common property with [galvanic skin response] GSR. Purina Dog Chow is not like canine saliva. (Excuse this recitation of the obvious.) We could not "infer" electroshock from GSR, or dog chow from salivation. It is not less obvious that if alcohol increases social extraversion we can say that alcohol can be observed to cause (but does not resemble) social extraversion, but cannot be inferred from socially extraverted behaviors as an unobserved cause, and it is not so much circular as simply muddled to say that social extraversion causes the indicators of social extraversion. [**?** , p. 670]

ability (and in fact there are such variants [**?** **?**]), to say that one position on an American football team requires more sprinting ability than another, and so forth. One purpose of GWAS is to open the "black box" of abstractions such as intelligence, revealing the causal networks underlying the phenomena that these abstractions subsume.

The construct of g formalizes the everyday observation that some people know more and learn more readily than others. It is also evident that this general mental ability tends to run in families. How much of this familial resemblance can be attributed to heredity, as opposed to environmental influences? This so-called "nature-nurture" issue can also be formalized. Let Y_j stand for the phenotype (g score) of the jth individual. We write the equation

$$Y_j = \mu + X_{j1}a_1 + X_{j2}a_2 + X_{j3}a_3 + \cdots + I_j + E_j. \tag{1}$$

At each genetic locus affecting the phenotype, we arbitrarily designate one allele to be counted. X_{ji} is then the count of the designated allele carried by the jth individual at the ith locus; this count can assume the values 0, 1, and 2. μ is a constant, and E_j is the jth individual's environmental deviation. Each a_i is the **average effect** of substituting one allele for another at the ith locus [**?** **?**]. Conceptually, we can equate each a_i to the partial regression coefficient of allele count at the ith locus in the regression of the phenotype on all loci in the genome.

We can imagine creating several clones of the jth individual and calculating the mean phenotype of these clones. Generally this mean will deviate from the value predicted by the regression just described; this **epistatic** deviation is given by I_j. For many purposes I_j is of little interest since the multilocus genotype of the jth individual may be unique. This genotype has probably never occurred in the evolution of the population and may never occur again.

In contrast, the value $\sum_i X_{ji}a_i = A_j$ is of the greatest interest. We can regard this value as the "genetically predictable" part of the jth individual's phenotype. The population variance in these values, σ_A^2, is called the **additive genetic variance**. The ratio of additive genetic to total phenotypic variance, σ_A^2/σ_Y^2, is the **heritability in the narrow sense** [**?** **?** **?**]. The ratio of total genetic to total phenotypic variance, $\sigma_{A+I}^2/\sigma_Y^2$, is called the **heritability in the broad sense**. If a phenotype has a high narrow-sense heritability, then most of the differences among individuals are caused by first-order genetic effects.

The correlations between the trait values of relatives are functions of variance components, including the narrow-sense heritability, and therefore enable the estimation of these underlying parameters [**?** **?** **?** **?** **?** **?**]. Studies of different kinships, including twins, parents-offspring, and adoptees, have provided impressively consistent results regarding the heritability of g: the broad-sense heritability lies between .70 and .80, and the additive genetic variance accounts for the majority of the genetic variation [**?** **?** **?** **?** **?** **?** **?** **?** **?**

].

In recent years the advent of advanced genotyping technology has allowed geneticists to move beyond macro-parameters such as σ^2_A/σ^2_Y for a given trait and estimate the individual a_i in **genome-wide association studies** (GWAS) [**?** **?**]. A recent spectacular example of this approach is the discovery of over 180 genomic regions associated with height in a sample of 180,000 individuals [**?**]. The results of GWAS confirm that the studied traits are indeed heritable and thus vindicate earlier studies of the correlations between relatives [**?** **?** **?**]. A typical GWAS examines associations between the focal phenotype and **single-nucleotide polymorphisms** (SNPs) scattered throughout the genome, and therefore any given signal probably does not arise from the causal variant itself but rather from a nearby marker. However, at least one recent study has followed up SNP genotyping with whole-genome sequencing, thereby tracing the original association signal to the causal effect of a particular missense variant [**?**].

Consider the major points in our theoretical framework: (1) the definition of the phenotype g, (2) the scientific and medical interest inhering in g as a result of its correlations with several important biological and social variables, (3) the substantial narrow-sense heritability of g, and (4) our growing capacity to identify the specific DNA variants contributing to the narrow-sense heritability of any given phenotype. These considerations jointly provide a powerful motivation to prosecute statistically well-powered GWAS of g.

1.3 Literature Review

Instead of assaying multiple markers throughout the genome, a **candidate gene study** tests for associations between the focal phenotype and markers within a single gene (or a few genes at most). Several positive findings from candidate gene studies of g have been reported [**?** **?** **?** **?** **?** **?** **?**]. However, given the consistent failures to replicate these findings [**?** **?** **?**], it appears that most or all of these reports are false positives.

The poor track record of candidate gene studies is not peculiar to research on g but rather is characteristic of research on a wide variety of traits. In retrospect this trend is not surprising. Researchers performing candidate gene studies have labored under the illusion that a lax statistical significance threshold is acceptable if the total number of tested hypotheses is small. As succinctly explained by the Wellcome Trust Case Control Consortium, however, the critical factor is not the number of tested hypotheses but rather the prior probability that any given hypothesis is correct [**?**]. Now consider the fact that there are more than 10^7 SNPs with a **minor allele frequency** (MAF) exceeding .01 in the human species [**?** **?**]. Any reasonable prior probability that one of these SNPs has a detectable effect on a particular phenotype must be extremely small. Since prior probabilities do not depend on the amount of data gathered, extremely strong evidence of

association is required to overcome a conservative prior probability, *regardless* of how many loci are examined in a particular study. Theoretical calculations and practical experience have shown that any associations clearing a significance threshold of 5×10^{-8} will attain a high posterior probability of being authentic [**?** **?** **?**]. Since candidate gene studies have not employed significance thresholds anywhere near this strict, statistical considerations provide a sufficient explanation for the inconsistent results under this approach.

More statistically sophisticated research teams have recently reported unbiased GWAS of g and other personality traits. Davis and colleagues administered a brief IQ test by telephone to a large sample of 7-year-old British children [**?**]. In the first stage of their study, two extreme groups were formed, each with 860 children. To qualify for the high-g (low-g) group, a child's score on the telephone IQ test had to be above (below) one positive (negative) standard deviation from the mean. The DNA of each group was pooled and typed on the Affymetrix GeneChip Human Mapping 500K chip. The 3,000 top-ranked SNPs from the first stage were taken forward to the second stage, in which a second sample of 1,000 children (500 high and 500 low) was selected and genotyped as in the previous stage. In the third and final stage, the 32 top-ranked SNPs from the second stage were selected for individual genotyping in a sample of 3,297 children. The final result was that no SNP showed convincing evidence of association with g. Evaluating the statistical power of the procedure leading to this result is not straightforward. Moreover, IQ measured at age 7 shows a rather weak correlation with g in adulthood [**?**]. But perhaps a fair conclusion to draw from this study is that no SNP accounts for more than 1 percent of the g variance in Europeans.

A paper in press by de Moor and colleagues reports a GWAS of the Big Five personality traits [**?**]. Although these personality traits do not include g, the results are nevertheless arguably relevant. The discovery sample consisted of 17,375 adults; five *in silico* replication samples totaling 3,294 adults were also employed. Genome-wide significance was obtained for Openness to Experience near the *RASA1* gene ($p = 2.8 \times 10^{-8}$) and for Conscientiousness in the brain-expressed *KATNAL2* gene ($p = 4.9 \times 10^{-8}$). However, the replication samples did not show significant associations between the top SNPs and the personality traits, although the direction of effect of the *KATNAL2* SNP on Conscientiousness was consistent in all replication samples.

The results of the de Moor study are sobering. Compare its number of hits (one at best) with those from the initial studies of height:

1. The GWAS of 13,664 individuals by Weedon and colleagues uncovered seven loci showing evidence of association at $p < 5 \times 10^{-8}$, all of which were later replicated [**?**]; and

2. The GWAS of 15,821 individuals by Lettre and colleagues also uncovered seven loci

showing evidence of association at $p < 5 \times 10^{-8}$, all of which were later replicated [**?**].

According to the comprehensive analysis of the GIANT Consortium, there are only ten common variants associated with height that account for more than 0.1% percent of the variance in that trait [**?**]. It appears, then, that the variants most strongly associated with the Big Five personality traits may not even account for 0.1% of trait variance.[2]

Schizophrenia is a complex cognitive disorder with a lifetime risk of ~ 1 percent. Assuming that schizophrenia represents the right tail of a continuous liability, a diagnosis of schizophrenia corresponds to a Z score of roughly $+2.33$. GWAS of schizophrenia may thus give some indication of the sample sizes necessary to discovering variants associated with g using case-control designs.

1. In a total comparison of 12,945 cases to 34,951 controls, deCODE Genetics found three genomic regions significantly associated with case status [**?**].

2. In a total comparison of 8,008 cases to 19,077 controls, the International Schizophrenia Consortium and the Molecular Genetics of Schizophrenia found one genomic region significantly associated with case status [**? ?**]. Interestingly, all three groups reported associations with SNPs in the major histocompatibility complex (MHC) region, suggesting that infectious agents play a role in the etiology of schizophrenia.

The study of deCODE Genetics was sufficiently well-powered to detect virtually all variants accounting for more than 0.1% percent of the variance in schizophrenia liability. From the fact that the study produced in fact only three hits, we can be confident that the common variants most strongly associated with schizophrenia typically account for much less than 0.1% percent of the variance. Before becoming overly pessimistic, however, we should realize that schizophrenia liability has almost certainly faced purifying selection in human evolution. Such selection tends to decrease the proportional variance of the leading loci and hence render them difficult to detect by GWAS. In contrast, g has probably faced positive directional selection at many points in human evolution, consistent with the steady expansion of brain volume in the lineage of *Homo sapiens* [**?**]. With a sufficiently large input of small and moderate mutations, such selection may well increase the proportional variance of the leading loci [**? ? ?**]. Therefore the available results on schizophrenia are

[2]The genetic architectures of quantitative traits other than height do seem to be more concentrated at loci of smaller effect. For example, the GIANT Consortium has reported that there are only three common variants associated with body mass index (BMI) that account for more than 0.1% percent of the variance in that trait [**?**].

reasonably interpreted as providing a very cautious lower bound on the likely effect sizes of the leading g loci.

What can be concluded from the existing literature? The failure of candidate gene studies is not at all in conflict with the substantial narrow-sense heritability derived from the correlations between relatives. Given the improperly calibrated significance thresholds used in candidate gene studies, false positives inevitably make up a preponderance of their reported findings. In contrast, the track record of unbiased and well-powered GWAS overwhelmingly shows that good study design produces replicable associations. After examining previous GWAS of g, personality, schizophrenia, and anthropometric traits, a reasonable lesson to take away is that studies of complex psychological traits should employ sample sizes large enough to detect a variant accounting for less than 0.1% percent of the trait variance. For a population association study, this implies a sample size exceeding 40,000 in order to attain the conventional power benchmark of .80 [**?**]. For a case-control study requiring a score of +2.5 standard deviations to qualify as a case, this level of power is attained with 10,000 cases and an equal number of controls [**?**].

2 Materials and Methods

We will pursue a series of case-control studies, in which allele frequencies are compared between a group of "normal" individuals (controls) and a group selected for exceptional intelligence (cases).

2.1 Participants

The source of the controls has not yet been determined.

Two samples of cases will be assembled. The first will consist of Chinese individuals living in China who have participated in the mathematics/science Olympiad training camps or who otherwise show evidence of mathematical precocity. We have reason to expect that the g scores of successful competitors will be very high [**?**], and our preliminary testing seems to bear out this expectation. Alternatively we can view the cases as being selected for "mathematical ability," a narrower trait than g that is nevertheless highly correlated with it. About 800 gifted youth participate in the training camps each year. Extending recruitment to past and future camps, we can expect to gather about 10,000 cases if the yield from this population is close to 100%.

The second sample of cases will consist of Americans who scored 760 or higher on the SAT Critical Reading (formerly Verbal) section and 800 on the SAT Mathematics section. Although lacking a component testing spatial ability, the SAT is otherwise an excellent

measure of g [**?** **?**]. Potential participants who took the SAT more than once may only use their most recent scores to qualify. Potential participants who do not meet the SAT thresholds may nevertheless qualify if they can provide satisfactory documentation of exceptional ability: a PhD in mathematics or physics from an elite university, participation in the mathematics/science Olympiads, a prize in the Putnam Mathematical Competition, or credentials of a similar caliber. Extending recruitment to past and future high school classes, we should aim to enroll at least 10,000 cases.

We are currently designing a website through which eligible individuals can enroll in the study. The site will provide information about BGI and the study, explain how individual data is kept private, and allow participants to arrange for a saliva collection kit to be sent to them.

2.2 DNA Collection and Phenotyping

Participants will provide DNA by spitting saliva into the OG-500 DNA Collection Kit. Those participants whose ability level requires confirmation by psychometric testing will be given either the appropriate Wechsler scale (translated into Chinese) or the appropriate version of Raven's Progressive Matrices. The latter test can be given over the web, although it is likely that online administration slightly alters its psychometric properties.

2.3 Genotyping and Analysis

The DNA will be extracted from the kits according to standard protocols. Initial genotyping of all Han Chinese individuals will be performed with BGI's SuperArray, a genotyping chip optimized to cover variants that are polymorphic in Han Chinese. Genotyping of individuals with different ancestral backgrounds will be performed either with one of the standard genotyping chips designed by Illumina or Affymetrix or another customized chip designed by BGI.

The extent to which cryptic relatedness and ancestral confounding are present in our samples will be estimated using the GCTA package [**?** **?**]. The principal components (PCs) of the successfully called genotypes will be computed using EIGENSTRAT and used as covariates in all association testing in order to control for ancestral confounding and genotyping artifacts [**?**].

Testing for association will be performed with PLINK [**?**]. Any genetic variants that show signal in all case-control comparisons, with an overall p-value less than 5×10^{-8}, will be declared hits.