IS THIS THE "GOLDEN AGE" OF BEHAVIORAL GENETICS?

EVAN CHARNEY

THE SAMUEL DUBOIS COOK CENTER ON SOCIAL EQUITY

TABLE OF CONTENTS

Executive Summary

Decades of twin studies have given rise to the conviction that all, or almost all, human behavior is heritable, no matter how embedded in culture, language, history, complex human interactions, and social institutions. With virtually all human behavior deemed heritable, the next step has been to identify the underlying genetic variants or genetic risk factors. The enduring hope has been to be able to determine on the basis of, for example, a newborn's genotype her "genetic risk" for engaging in criminal behavior or for doing poorly in school. This search has preceded in two phases.

The first phase of this search, the heyday of which took place from approximately 1990-2010, involved candidate gene association (CGA) studies of a small number of polymorphisms of a handful of genes. Despite the hype and promise, they failed for a variety of reasons. From their inception, CGA studies were plagued by failures of replication. For example, for every study in which researchers reported that specific polymorphisms of the MAOA gene predicted "anti-social" behavior, there was a study displaying no association. In addition, CGA studies frequently exhibited data mining or multiple hypothesis testing through, for example, the inclusion of numerous interaction terms with no p-value adjustment.

It is now the consensus view in behavior genetics that all of these studies' significant findings were "false positives." In an article titled, "A Waste of 1000 Research Papers," psychiatric geneticist Matthew Keller, reflecting on CGA studies in general, asked, "How on Earth could we have spent 20 years and hundreds of millions of dollars studying pure noise?" This report asks if the lessons of the cautionary of CGA studies have been learned.

The second phase of the search for genetic variants underlying heritability centers on genome wide association studies (GWASs), polygenic scores, and the resurrection of Fisher's "infinite infinitesimal" allele model, first proposed 100 years ago. According to this model, hundreds or thousands or millions of polymorphisms, each of miniscule effect, act as genetic risk factors for heritable behavior. Hence, massive sample sizes are required to find alleles that each exercise a tiny effect.

GWASs involve the search for differences in the frequencies of a type of genetic variant known as a single nucleotide polymorphism (SNP) between cases and controls. A million or more of these polymorphisms are examined in the genomes of large numbers of persons, with sample sizes in the hundreds of thousands or even millions. A GWAS is an acknowledged form of data mining: For this reason, a Bonferroni correction is typically used to adjust for a million or more individual tests. Any SNP that is significant at the stringent p-value of $\leq 5.0 \times 10^{-8}$ is said to have genome wide significance and to be a "lead" SNP. As sample sizes have grown, the number of lead SNPs also have grown.

However, like the CGA era, these lead SNPs have not been consistently replicated across studies. Six large metanalyses of "intelligence"/"cognitive ability" of hundreds of thousands of individuals identified 1906 SNPs across the six studies at a stringent level of statistical significance; of these, no SNP was replicated in all of the studies, 11 percent were replicated in more than one study and among those 76 percent were replicated in only one additional study. Four metanalyses of educational attainment displayed a similarly poor record of replication.

Nevertheless, the results of these studies are now commonly used to construct polygenic scores. In theory, a polygenic score (PGS) provides a single numerical measure of genetic risk. It is constructed by adding up the estimated effect sizes of all of the SNPs that show an association with a particular phenotype (which can be well over a million SNPs). PGSs are often characterized as predictors of individual risk, or one's likelihood to attain a certain outcome.

In fact, polygenic scores for educational attainment or intelligence or income have no individual predictive value whatsoever. It is an open question as to whether any individual polygenic scores for any phenotype have predictive value.

The success of a PGS is measured by its R-squared: the higher the R-squared, the greater the amount of phenotypic variance predicted in a given population. Researchers have an enormous amount of freedom in constructing a PGS with the goal of achieving the highest R-squared possible. This includes the freedom to try different p-values for inclusion of the estimated effect sizes of individual SNPs in the score, including values up to p=1.

To be clear, this approach abandons any statistical correction for data mining. Researchers claim that this is justified by an increase in predictive power (as measured by a higher R-squared). There are strong reasons to conclude that all of this freedom results in model overfitting—a prediction based on sample noise rather than a reflection of a real relationship—despite the claims of researchers to have introduced safeguards against this.

All of the studies alluded to, as well as most of the major published studies involving GWASs and PGSs, are intentionally limited to "whites of European ancestry" (WoEA). When non-WoEA are included in any part of the study, the R-squared of the polygenic score plummets. The standard reason given for this concerns population differences in allele frequencies due to different migratory histories, known as population "structure."

The notion that researchers can effectively account for population "structure" within WoEA—but not between WoEA and other ancestral groups—reinforces the idea of intra-group (relative) genetic homogeneity, inter-group genetic heterogeneity, and the significance of the system of classification based upon these presumed genetic differences. Certainly, this is a dangerous belief. Moreover, there is increasing evidence that the usual techniques for dealing with population structure among WoEA (principal components analysis and linear mixed models) are inadequate and have led to biased findings.

In sum, the lessons of the candidate gene association study era have not been learned in the subsequent era of genome wide association studies and polygenic scores. While the methodologies are different, they exhibit many of the same infirmities that doomed CGA studies, and have introduced a host of new ones. The search for genetic foundations of complex human behaviors has been a diversion. The costs of this research have been immense in terms of time, energy, and financial resources, and the research itself has distracted scholars from more promising routes toward understanding such things as differences in academic performance or income.

Introduction

Decades of twin studies have given rise to the conviction that all, or almost all human behavior, no matter how entwined with culture, language, history, complex human interactions, and social institutions, is heritable. With so much deemed heritable, the next step has been to identify the underlying genetic variants. This process has unfolded in two phases. The first phase, which extended from the early 1990s to approximately 2012, was characterized by the use of candidate gene association (CGA) studies.

During this time, researchers reported, literally, in thousands of publications that specific polymorphisms of specific genes could predict everything from income to educational attainment to attitudes toward nuclear power. CGA studies have fallen largely out of favor, so much so that psychiatric geneticist Matthew Keller, reflecting on CGA studies in general, recently asked, "How on Earth could we have spent 20 years and hundreds of millions of dollars studying pure noise?" (quoted in Yong 2019).

This brings us to the second phase, which has recently been referred to as a "golden age" of genomics: "Genomic technology has ushered in a golden age of new tools to address enduring questions about how genes and environments combine to create unique human lives" (Harden 2020). These "new tools" include genome wide association studies (GWAS), single nucleotide polymorphism heritability estimates, and polygenic scores. The claims are the same as in the era of CGA studies: that complex social behaviors, including income and educational attainment, can be predicted on the basis of an individual's genotype.

The rise of the new golden age from the ashes of CGA studies has occurred at a precipitous rate. The failure of CGA studies has been called a cautionary tale (Keller, quoted in Yong 2019). Reflecting back, researchers warned (Chabris et al. 2012, 8) that associations of genes, "with psychological and other social science traits should be viewed as tentative until they have been replicated in multiple large samples," because "[d]oing otherwise may hamper scientific progress by proliferating potentially false positive results." Have the lessons of the cautionary tale of CGA studies been learned in the age of GWAS and polygenic scores? And has the "true" golden age of behavior genetics arrived at last? As argued here, the answer to both of these questions is "no." The current iteration of the search for genetic variants underlying heritability is subject to many of the same infirmities that doomed its predecessor.

I. Heritability

A heritability estimate for a specific trait or attribute[1] is a measure of the amount of variation in that attribute among the members of a given population, at a given time, that can be correlated with members' genetic variation. The easiest way to conceptualize a heritability estimate and what it represents is to think of heritability in relation to the population of parents and children. A heritability estimate is intended to provide an answer to the following question: How much of the variation (0 [0%] to 1 [100%]) in a given trait of a child is correlated with, and presumably caused by, the DNA sequences she inherits from her parents, and how much is due to her "environment" (generally construed as everything other than DNA sequences)?

However, this definition needs to be amended as follows. For the most part, heritability is *not* taken to refer the percentage of variation in an attribute (in a given population at a given time) that is correlated with variation in the inherited DNA (of the members of that population), but rather to the amount of variation in the *risk* of an attribute that is correlated with genetic variation. This is because for most complex traits (traits that are not caused by single gene mutations and inherited in a straightforward Mendelian manner), inherited DNA sequences are not, by themselves, deemed sufficient to be causal. Rather, they contribute a certain amount of risk and the environment contributes a certain amount of risk. In this context, "risk" has no connotation of adverse consequences. Rather, it refers simply to the increased probability of an occurrence of something, be it good, bad, or indifferent.

Individuals typically inherit 22 pairs of numbered (homologous) chromosomes, one from each parent, called autosomes. In addition, females inherit a pair of homologous X sex-chromosomes, and males inherit a single Y sex-chromosome. Each of these chromosomes contains a long DNA molecule. Segments of the corresponding nucleotide sequences on each pair of chromosomes are typically the same but can also differ.

For example, at a specific location on chromosome 13, one might inherit an A nucleotide from her father and a G nucleotide from her mother (or she might inherit two A's, or two G's). Or, at a particular position on chromosome 4 she might inherit three copies of the repeating DNA sequence "ACCA" from her mother, and four copies from her father. Each of these copies of DNA sequences at a given position on a chromosome is called an allele.

Twin Studies

Twin studies are based upon a comparison of pairs of monozygotic (MZ) twins, derived from a single egg and sperm and dizygotic (DZ) twins, derived from two separate egg and sperm. Because MZ twin are derived from a single egg and sperm, they are presumed to share 100% of their inherited DNA sequences. DZ twins, derived from two separate egg and sperm, differ from each other in approximately 50% of their inherited DNA sequences.

In a twin study, researchers measure the concordance rate—the probability that a pair of individuals will both have a certain trait—for pairs of MZ and DZ twins for a trait of interest. If MZ twin pairs have a greater concordance rate for that trait than DZ twin pairs, this greater concordance is ascribed to MZ twins' greater genetic concordance (1 for MZ twins and .5 for DZ twins) and is used to derive a heritability estimate. According to the standard behavior genetics model (Posthuma et al. 2003), which measures so-called "narrow sense" heritability (h^2), the genetic heritability of an attribute is twice the difference of the correlations between MZ and DZ twins: $h^2 = 2(R_{(MZ)} - R_{(DZ)})$.

One important assumption of this model is *additivity*: If, for example, one

[1] " Attribute" is my preferred expression for any observable characteristic. For the most part, I use it instead of "trait," which has connotations that are inappropriate in a number of contexts. I also employ the term "phenotype"—which implies genotype—in certain contexts without necessarily endorsing the assumption of a corollary genotype..

variant form of a gene (an allele) increases the risk of being 1 cm taller, then have having two copies of that allele (i.e., being homozygous) increases the risk of being 2 cm taller. This model assumes no dominant or recessive effects, no gene x gene (G x G) interactions, and no gene x environment (G x E) interactions. It is also the model that is assumed in all of the studies that will be considered throughout this paper.

Over the years, twin studies appear to have shown the heritability of any and every human "behavior," from attitudes toward nuclear power (Alford et al. 2005) to religious affiliation (i.e., whether one is an Episcopalian, a Hassidic Jew, a Sunni Muslim, or Zoroastrian), despite the apparent absurdity of such findings.

Such findings, and countless others, were the impetus for the so-called "first law" of behavior genetics (Turkheimer 2000): "All human behavior is heritable," which was, apparently, not intended as a "universal, mechanistic truth," but as "a robust empirical reality" (Chabris et al. 2015).

It is not my intent here to undertake a general critique of the twin study methodology, which has been vigorously challenged elsewhere (Joseph 2014; Richardson and Norgate 2005; Kamin and Goldberger 2002; Wood 2020; Pam et al. 1996; Charney 2012). I will, however, have something to say later concerning the additive model that has been used to derive the vast majority of twin study heritability estimates.

II. The Search for Differences in Polymorphism Frequencies I: Candidate Gene Association Studies

Heritability concerns inter-individual differences. Attributes that all persons share with the rarest of exceptions, like a brain, are not heritable (although differences in brain size, for example, are said to be heritable). Individual human genomes contain approximately 3 billion nucleotide base pairs. The sequence of these bases is roughly 99.9 percent alike in all humans. The search for the genetic basis of heritability, or genetic risk factors, is a search across the 0.1% of the human genome in which humans differ in their DNA sequences.

When genetic variants are "common" in a given population, which is defined as occurring in more than 1% of the population, they are referred to as polymorphisms. The search for genetic variants' underlying heritability, or for genetic "risk factors," is a search for differences in *allele frequencies*, i.e. differences in frequencies between cases and controls in dichotomous traits, or between those with different trait values in a quantitative trait . For ease of exposition, generally, I refer to attributes/traits as if they were dichotomous. In short,the search for genetic variants is an attempt to ascertain whether those with a given attribute have a higher frequency of certain alleles in their genomes compared against those who do not have the attribute.

i. CGA Studies

The first phase of the search for genetic risk factors in behavior genetics began in the 1990s. Prior to the sequencing of the human genome and dramatic advances in technology and reductions in cost that now enable the sequencing of millions of base pairs of the genomes of millions of individuals, the hunt for genetic variants focused on a small number of polymorphisms of a small number of genes (roughly 5 to 10).

In particular frequencies of polymorphisms of the genes MAOA, 5HTT, DRD2, and DRD4, which are involved in the synthesis of proteins that play a role in regulating various neurotransmitters, were examined for correlations with a wide array of behaviors. A relationship between these genes and human behavior was first made in the context of pharmacology. For example, the enzyme monoamine oxidase, synthesized from the MAOA gene is involved, among other things, in the breakdown of monoamines (the neurotransmitters dopamine, noradrenaline and serotonin); in the 1950s, monoamine oxidase inhibitors (MAOIs) were introduced for the treatment of depression and later a host of other conditions including panic disorder, social phobia, ADHD, migraines, and Parkinson's disease.

Polymorphisms of these genes were typically classified as being associated with differences in "transcriptional efficiency," the amount of protein synthesized from a gene in a given period of time, or "protein efficiency," the speed with which a protein acts or its endurance. For example, a polymorphism in the promoter region of the MAOA gene is characterized by a repeating sequence 30 bp long that can occur with different numbers of repeats (2, 3, 3.5, 4, or 5).

On the basis of *in vitro* analysis, the 3.5 and 4 repeat variants were classified as "high" (H-MAOA), for "high transcriptional efficiency," and the 3 and 5 repeat variants as "low" (L-MAOA), for "low" "transcriptional efficiency" (Sabol et al. 1998). Similar classifications were made for polymorphisms of the other genes typically studied at this time (e.g., repeat polymorphisms of the serotonin transporter gene [5HTT] were classified as "long" [l] or "short" [s] on the basis of presumed differences in transcriptional efficiency).

[2]MZ twins and DZ twins have all inherited 100% of their DNA sequences from their parents. When DZ twins are said to differ in about 50% of their inherited DNA sequences, the difference consists in which alleles each DZ twin has inherited from the mother and which from the father. [3]"Religious affiliation" is frequently cited by researchers as an example of a non-heritable "behavior" (Olson et al. 2001; Eaves et al. 1990; D'Onofrio et al. 1999). This is taken as a confirmation of the legitimacy of the twin study methodology, namely, that it does not lead to results that "defy common sense" (Flint et al. 2020, p. 15). What is heritable is "religious attitudes—not specific religious affiliations but general views about the value of religion" (Olson et al. 2001). The supposed distinction between "religious attitudes" and "religious affiliation" is unsustainable, inasmuch as religious "attitudes" (which are, presumably, equivalent to religious beliefs) are set forth by the religion to which one belongs. That said, and disavowals notwithstanding, "religious affiliation" has been shown to be heritable, although researchers have bent over backwards to deny, minimize, or obfuscate such findings. Loehlin and Nichols (1976), who conducted a study of 850 twin pairs and published all of their raw data in a subsequent book. Among the questions twins were asked was, "What is your present religious preference?" and respondents were asked to circle one of the following: "Protestant; Roman Catholic; Jewish; other; None." While Loehlin and Nichols never indicated whether they found "religious preference" to be heritable, Schönemann (1997) calculated that on the basis of the published intraclass correlations, religious affiliation was 85% heritable for males and 21% for females. D'Onofrio et al. (1999), commenting on an earlier study of 3810 pairs of twins (Eaves et al. 1990), indicate that (1999, p. 967): "There is a slight (and statistically significant) reduction in the DZ resemblance [for religious affiliation] in women [emphasis added]." Despite the correlation being statistically significant, they do not calculate a heritability estimate, but do provide the twin correlations, which indicate a heritability of 24% for women. This figure (24%) is higher than the heritability of other attributes that they highlight, including "institutional conservatism" (12%) and "religious occupational interests" (19%). Bradshaw and Ellison (2008) report a "high" heritability (65%) for responding "yes" to the following question: "Have you been 'born-again,' that is, had a turning point in your life when you committed yourself to Jesus Christ?," which indicates whether one is a Baptist. In a recent book on behavior genetics, the authors refer to the heritability of religious affiliation with revealing wording (Flint et al. 2020, p. 15): "Twins strongly resemble one other in their religion, but the degree of resemblance is *virtually* the same for MZ and DZ twins [emphasis added]." This is clearly not the case. One senses an effort to minimize uncomfortable findings (when has one ever encountered researchers playing down the significance of something that they acknowledge to be "statistically significant"?). That the twin study methodology leads to what, according to those who employ the methodology, are absurd results is just one of its many problems.

Polymorphisms of these genes were associated with behaviors via candidate gene association (CGA) studies.

A standard CGA study is hypothesis-driven: a researcher proposes, on the basis of the presumed biological effect of, for example H-MAOA v. L-MAOA, that those with L-MAOA are more likely to engage in a certain type of behavior. This hypothesis is then tested in a data set.

The years 1990 to 2010 were what might be called the "golden age" of CGA studies. Aided by the proliferation of large data sets that included an array of behavioral data (usually in the form of self-reporting), and genotypic data, often limited to the same polymorphisms of the same handful of genes, researchers published thousands of studies reporting statistically significant correlations between polymorphisms of the same handful of genes and every conceivable behavior.[4] For example, using data from The National Longitudinal Study of Adolescent Health (Add Health)[5],researchers hypothesized that inasmuch as L-MAOA had been associated with anti-social behavior (Caspi et al. 2002), H-MAOA was likely to be associated with "pro-social" behavior, and since voting is a form of "pro-social" behavior, H-MAOA might be associated with "voting behavior" (Fowler and Dawes 2008). Based on responses in the Add Health data set to the question, "Did you vote in the last Presidential election?", they reported that H-MAOA was significantly associated with increased voter turnout ($p = 0.03$), and that the odds of those with the H-MAOA genotype voting were 1.26 times greater than those with the L-MAOA genotype.

On an almost weekly basis, media reports heralded the findings of the latest CGA study with sensationalist headlines (e.g., "Procrastination is in your genes" [CNN, April 8, 2014]; "The urge to infidelity ... it's in her genes" [*The Guardian*, Nov. 21, 2004]); Researchers suggested that the results of CGA studies might contribute to "individualized preventive psychiatry" (Muller-Spahn 2008) and that early intervention services for the families of L-MAOA children might be a means to reduce violent crime (Brooks-Crozier 2011). Legal scholars debated whether L-MAOA could count as a defense in a criminal trial (McSwiggan et al. 2017); and medical ethicists suggested that we might have a moral obligation to avoid having children with the L-MAOA genotype (Savulescu 2014).

ii. Replication I

What was missing in all of the hullabaloo surrounding CGA studies was that, from their inception, they were plagued by failures of replication. Consistent replication of findings across studies remains one of the most important tools for verification in the empirical sciences. For every study, or multiple studies, reporting an association between, for example, L-MAOA and "anti-social" behavior (Caspi et al. 2002; Kim-Cohen et al. 2006; Widom and Brzustowicz 2006), a study (or studies) reported no association (Prichard et al. 2008; Haberstick et al. 2005; Huizinga et al. 2006). The situation was precisely the same for all of the other notable associations between polymorphisms of MAOA, 5HTT, DRD2, DRD4, and well as several other genes.

In 2012, the editor of *Behavior Genetics,* the premier behavioral genetics publication noted, in setting forth strict criteria for publication of CGA studies (Hewitt 2012, 1):

> The literature on candidate gene associations is full of reports that have not stood up to rigorous replication. This is the case both for straightforward main effects and for candidate gene-by-environment interactions. As a result, the psychiatric and behavior genetics literature has become confusing and it now seems likely that many of the published findings of the last decade are wrong or misleading and have not contributed to real advances in knowledge [citations omitted].

In adopting a similar editorial policy for the publication of CGA studies, the editor of the *Journal of Abnormal Psychology* noted, regarding publication bias (Johnston et al. 2013, 512):

> The tendency for novel findings to subsequently fail replication may be particularly great in new and "hot" areas of research, such as candidate gene associations and gene-environment interactions. The existence of a strong publication bias towards positive findings is partly due to incentives both for authors and editors to publish positive reports. Other things being equal, reviewers and editors may be more likely to agree that exciting and novel findings should be published than research on more established topics.

In an article titled, "Most Reported Genetics Associations with General Intelligence are Likely False Positives" (Chabris et al. 2012), researchers reported the failure to replicate, using large (for the time) sample sizes (5571, 1759, and 2441), any of the previously reported candidate gene associations between "g" (a measure of intelligence) and polymorphisms of 12 different genes. In their conclusion, they cautioned (Chabris et al. 2012, 8):

> Associations of candidate genes with psychological and other social science traits should be viewed as tentative until they have been replicated in multiple large samples. Doing otherwise may hamper scientific progress by proliferating potentially false positive results, which may then influence the research agendas of other scientists who do not appreciate that the associations they take as a starting point for their efforts may not be real. And the dissemination of false results to the public risks creating an incorrect perception about the state of knowledge in the field, especially the existence of genes described as being "for" traits on the basis of unintentionally inflated estimates of effect size and statistical significance.

Similar large disconfirming studies were also published regarding schizophrenia (Farrell et al. 2015) and depression (Border et al. 2019). In an article titled, "It is Time to Abandon the Candidate-Gene Approach to Depression" (Border et al. 2019), researchers reported that candidate genes for depression were no more correlated with depression than any gene chosen at random. As noted in the introduction, Keller, reflecting on CGA studies in general, recently said of the heyday of CGA studies (quoted in Yong 2019):

> This should be a real cautionary tale. How on Earth could we have spent 20 years and hundreds of millions of dollars studying *pure noise*?

[4]For a list of some of the phenotypes reported to be associated with polymorphisms of MAOA, 5HTT, DRD2, and DRD4, see "Four Genes Predict Everything": https://sites.duke.edu/evancharney/files/2014/12/4-GENE-PREDICT-EVERYTHING-12-1.pdf
[5]https://addhealth.cpc.unc.edu/
[6]See "Four Genes Predict Everything": https://sites.duke.edu/evancharney/files/2014/12/4-GENE-PREDICT-EVERYTHING-12-1.pdf

Of CGA studies, Flint et al. comment (2020, p. 60): "There are literally thousands of papers reporting the reults of [CGA studies], but it's not too harsh to say simply that these studies have taught us nothing useful about the genetic basis of psychiatric disease."

The fact that 20 years and hundreds of millions of dollars were spent studying "pure noise" is now the consensus view in behavior genetics (Border and Keller 2017; Duncan et al. 2019), although it is by no means universally accepted inasmuch as studies involving polymorphisms of the same handful of genes continue to be published to this day. The question posed by Keller never has been addressed in any serious and sustained manner by the behavior genetics community.

Keep in mind that all of this "noise" was generated by thousands of studies published in prestigious science and social science journals with the appearance of being based on rigorous statistical analyses. These studies were conducted by faculty at high cachet research universities and funded by millions of research dollars from organizations like the National Institutes of Health. An extensive analysis as to how such a thing could have happened will not be presented here (although we are certainly in need of one). Rather, discussion is limited to two central themes in what follows immediately.

iii. Data Mining and Multiple Hypothesis Testing

Null hypothesis significance testing is the most widely used data analysis method in most scientific disciplines. According to the null hypothesis, there is no relationship between the two variables being studied and results showing a relationship are due to chance alone. The alternative hypothesis is the one you would believe if the null hypothesis isdetermined to be untrue.

The p-value represents the probability of finding a relationship between the two variables when the null hypothesis is true. This is typically expressed as a level of statistical significance between 0 and 1. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis. For testing a hypothesis, the commonly employed p-value is ≤ 0.05. A type 1 error, or a false positive, occurs when a true null hypothesis is incorrectly rejected, and a type II error, or false negative, occurs when a false null hypothesis is not rejected.

A p-value ≤ 0.05 is a statistical threshold for testing a *single* hypothesis. When more than one test is run without any kind of correction in the form of a more stringent (that is, lower) p-value threshold, the overall type I error rate is much greater than 5%. For example, suppose one is using a behavior data set that has genotypic information for participants MAOA genotype (L or H) and subject responses to 1000 behavioral questions. A comparison of the frequencies of L v. H-MAOA for each of these "traits" (or responses) constitutes a hypothesis. Each association test is essentially a χ^2 test, if the trait is categorical, or a linear regression test if the trait is continuous (and follows a normal distribution).

Thus, the testing of 1000 different traits amounts to a 1000 χ^2 tests (or linear regressions), each with its own null hypothesis. If the null hypothesis was true, an a level of 0.05 could theoretically produce 50 "significant" correlations by change alone. The most straightforward way to deal with multiple hypothesis testing is the Bonferroni correction, in which the a level is divided by the number of tests performed (i.e., the more tests performed, the more stringent the level of statistical significance). Dividing .05 by 1000, for example, yields a p-value of .00005.

In tCGA studies, there are numerous examples of multiple hypothesis testing in the absences of any p-value adjustment. At one extreme, we have what might be termed a "statistical felony," which does not actually concern any kind of hypothesis testing (as the term "hypothesis" is traditionally understood), multiple or otherwise. Rather, it involves "data mining," that is, searching for correlations between a given gene and any and all of the (hundreds or thousands) of behavioral variables in a behavioral dataset with no statistical correction for running hundreds or thousands of tests (i.e., employing the standard p-value for testing a single hypothesis of p ≤ 0.05). Then, when a correlation is found, to construct a post hoc "hypothesis" to explain the finding (constructing a "hypothesis after the results are known," or what is commonly referred to as HARKing [Kerr 1998]).

Researchers also can engage in, if not pure data mining, a more limited form of multiple hypothesis testing (although there is no way to tell which is occurring in a given instance). For example, researchers might test multiple different polymorphisms or combinations of polymorphisms of the same gene and choose those that show a statistically significant correlation with a given behavior. For example, they can select those who have at least one copy of the number 3 (out of 9) repeat (R) allele of a polymorphism of the DRD4 gene have been reported to be more likely to have sexual intercourse at a younger age (Guo and Tong 2006).

Or, associations may be reported between any combination of alleles and a given phenotype; for example, those with the genotype DRD4 2R/2R are more likely to exhibit depression (Guo and Tillman 2009b). Or associations may be reported between any combination of alleles of a given gene and the alleles of another gene (a G × G interaction); for example, those with DRD2 polymorphisms A1/A2 or A1/A1 and at least one 7R copy of the DRD4 gene are more likely to exhibit conduct disorders (Beaver et al. 2007a).

Furthermore, associations may be reported between any of these alleles and specific age groups, or genders, or ethnicities, or specific ethnicities and genders, and these associations may involve any conceivable gene x environment (G × E) interaction. An example is Vaske et al.'s (2009b) finding that African American females who use marijuana and have the short/short 5-HTTLPR genotype are more likely to engage in "property offending" (Vaske et al. 2009b). CGA studies are supposed to be a form of hypothesis testing in which the hypothesis is formed prior to undertaking the statistical analysis, not afterwards. On what basis would one propose, a priori, that, for example, African American females who use marijuana and have a short/short 5-HTT genotype are more likely to engage in "property offending"?

The manipulation of data in order to produce a desired p-value is sometimes referred to as "p-hacking." P-hacking is typically accomplished through manipulation of "researcher degrees of freedom," or the decisions made by the investigator. These include when to stop collecting data, whether or not the data will be transformed, which statistical tests (and parameters) will be used, how many and which variables and interaction terms will be used,

and so on. It has been estimated that by simply manipulating a researcher's degrees of freedom, even absolutely negative data—data that shows no statistical correlation—can produce a p-value under 0.05 61% of the time (Simmons et al. 2011). Excessive degrees of freedom can lead to model overfitting, which will be explained in greater detail below.

Antother type of multiple hypothesis testing involves the widespread use of the same data sets by researchers engaged in a global hunt for correlations between the same handful of polymorphisms and every conceivable human behavior. For example, the DRD2 Taq1A polymorphisms have been associated, by researchers using the same data set (The National Longitudinal Study of Adolescent Health) with at least 19 different behaviors in 19 different studies. Each of these studies constitutes (ideally) the testing of a single hypothesis, and collectively the testing of 19.

Population stratification is often cited as a reason for the failure of CGA studies (Thomas and Witte 2002; Cardon and Palmer 2003). This phenomenon will be considered in the discussion of genome wide association studies and polygenic scores below.

III. The Search for Differences in Allele Frequencies II: Genome Wide Association Studies

The main methodology used in the search for polymorphisms underlying the presumed heritability of complex behaviors gradually shifted from CGA studies to genome wide associations studies (GWAS), a shift that was enabled by advances in DNA sequencing technology and dramatic reductions in cost. GWAS now constitute the new workhorse of behavior genetics.

While GWAS can be used to examine a variety of different kinds of genetic variants, the focus of most current GWAS is a particular type of polymorphism known as a single nucleotide polymorphism (SNP). A SNP is characterized by the substitution of a particular nucleotide at a given position or locus on a DNA molecule. SNPs are largely, but by no means exclusively, diallelic, meaning they come in two possible forms (e.g., A or G). SNPs can occur anywhere in human genomes—within genes or in intergenic regions. They are the most common form of genetic variation in human populations (numerically, although not in terms of the total size of the region of the genome implicated).

It is estimated that the genome of any individual contains approximately 8-11 million SNPs. The less frequently occurring SNP in a given population is called the minor allele; the more frequently occurring is the major allele. As with other forms of genetic variation, a SNP is considered "common" if it occurs in more than 1% of a population. The search for SNPs associated with behavioral traits typically focuses on common SNPs, the assumption being that "common" traits, or relatively common traits, will be associated with common alleles.

Like a CGA study, a GWAS can be used to test the frequency differences of alleles known in advance in cases versus controls. However, the most common type of GWAS is somewhat confusingly referred to as "hypothesis free." It does not involve a specific polymorphism or polymorphisms that researchers hypothesize to be associated with a trait on the basis of its presumed physiological effect. Rather, large numbers of SNPs (a million or more in some studies) in a large number of cases and controls (a million or more in some studies) are compared to ascertain whether there is any difference in the frequency of SNP-alleles between cases and controls. There are no a priori assumptions as to what these genetic variants might be.

It is, in effect, a "blind search" for correlations deemed statistically significant between an attribute of interest and SNPs. However, while GWASs do not test pre-existing hypotheses of the sort, "we hypothesize that polymorphism x will predict behavior y," they are by no means "hypothesis free." A comparison of the frequencies of each of the million SNPs is essentially a χ^2 test, if the trait is categorical, or a linear regression test if the trait is continuous (and follows a normal distribution). Thus, the testing of a million SNPs amounts to a million χ^2 tests (or linear regressions), each with its own null hypothesis.

If the null hypothesis was true, an a level of 0.05, theoretically, could produce 50,000 "significant" SNPs. If we employ a Bonferroni correction and divide .05 by 1 million (the number of tests performed), the resulting p-value is 5.0×10^{-8}. This is the threshold of statistical significance that is most commonly employed in GWASs. A SNP that achieves this significance level is said to have genome wide significance.

i. Linkage Disequilibrium

The search for differences in allele frequencies among cases versus controls is not a search for alleles that have not been previously identified. The alleles are that investigated have already been catalogued and mapped across the genome, and they are the SNPs that are investigated on a DNA microarray (or DNA chip). These SNPs are referred to as "tag" or "marker" SNPs. Hence the search for SNPs in a GWAS is not actually a search for just any SNPs, but rather, in the first instance, a search for previously identified marker-SNPs. To understand how this works, it is necessary to say something about the phenomenon of linkage disequilibrium.

---

[7]Associations claimed between the DRD2 Taq1A polymorphism and behaviors using the National Longitudinal Study of Adolescent Health data set include: Violent delinquency among males (Guo et al. 2007a); "homophily," a desire for friends who also have the DRD2 Taq1A polymorphism (Fowler et al. 2011); the number of vaginal sexual partners in the previous year among males (Halpern et al. 2007); whether or not an offender has been violently victimized (Vaske et al. 2009); depression (Guo and Tillman 2009a); victimization among white males who have delinquent peers (Beaver et al. 2007b); partisanship (Dawes and Fowler 2009); contraceptive use (Daw and Guo 2011); the intergenerational transmission of parenting (Beaver and Belsky 2012); resiliency to victimization (Beaver et al. 2010c); polydrug use among males who exhibit maternal withdrawal (Vaughn et al. 2009); continuation of education beyond secondary school among males who have mentors who are teachers (Shanahan et al. 2007); frequency of alcohol consumption among young adults excluding adolescents (Guo et al. 2007b); five antisocial phenotypes among African American females who have a criminal father (DeLisi et al. 2009); academic achievement during middle and high school (Beaver et al. 2010b); smoking among young adults who report at least six inattentive symptoms (McClernon et al. 2008); verbal skills among whites (Beaver et al. 2010a); continuation of education beyond secondary school among males who have high parental socioeconomic status, high parental involvement in school, or attend high-quality schools (Shanahan et al. 2008); conduct disorder among males who also possess the DRD4 Exon 3 VNTR polymorphism (Beaver et al. 2007c); and the depressive effects of violent victimization on African American females (Vaske et al. 2009). Each one of these studies is the testing of a hypothesis..

[8]A recent analysis of data from the 1000 Genomes Project has identified 271,934 tri-allelic SNPs, or approximately 0.32 % of all listed SNPs. The researchers note that multiple allele SNPs, including tetrallelic and triallelic, now make up a significant proportion of total SNP variation (Phillips et al. 2020). For tetrallelic SNPs, see (Phillips et al. 2015).

According to Mendel's law of independent assortment, any given allele will be inherited independently of any other allele. If someone inherits the SNP-allele A at a specific position on a specific chromosome, this tells us nothing about the allele that she will inherit on the same chromosome a hundred base pairs away.

In reality, SNPs that occur "close" to each other tend to be inherited together because the entire segment of DNA on which they are located is inherited as a single piece. This segment of DNA is referred to as a *haplotype bloc* or simply a *haplotype*. The genotypes of different SNPs within a haplotype tend to be inherited together. Suppose that two SNPs are located at different loci in a haplotype. In SNP-1, the nucleotides G or T occur, and in SNP-2, the nucleotides C or G occur. Suppose we know that the genotype of SNP-1 is T; on this basis we can infer that the genotype of SNP-2 is G (with varying degrees of probability).

This non-random association of alleles at two different loci in a haplotype is known as linkage disequilibrium (LD). *Linkage disequilibrium* plays a critical role in GWAS. Haplotypes have been mapped out across large segments of the genome. A current genome assay typically checks a person's DNA sequences for a million marker-SNPs. However, researchers can infer, on the basis of the genotype of the marker-SNP, the genotype of many more SNPs known (or presumed) to be in linkage disequilibrium with it. This allows researchers to examine large segments of the genome without having to genotype each of the three billion pairs of DNA nucleotides.

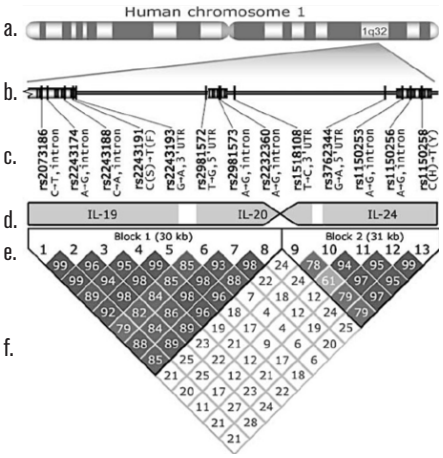
If researchers find a "hit" or correlation for a marker-SNP, the assumption, generally,is made that the marker-SNP itself is not "causal." Rather, the causal SNP is assumed to be an unknown SNP in LD with the marker-SNP. In other words, the marker-SNP serves as a proxy for the presumed, or hypothesized, causal SNP. The importance of this fact cannot be understated.

One complication regarding LD is that multiple marker-SNPs that achieve genome wide significance could be in LD with each other. None of these marker-SNPs are considered to be causal. Rather, they are all assumed, in addition to being in LD with each other, to be in LD with, and to serve as proxies for, the same unknown causal SNP. Hence, if all of the SNPs in LD with each other were counted as having genome wide significance, this would be a form of "overcounting." The simplest way to deal with this is referred to as "clumping." Marker-SNPs in LD with each other in a given locus are "thinned," retaining only the marker-SNP with the lowest p-value. After this is done for every region of LD, the remaining SNPs are assumed to be "independent" of each other and hence, not in LD. These are typically referred to as "lead" SNPs.

Whether any two SNPs are in LD is a matter of probability, not certainty. Researchers use as a measure of the strength of correlation between any two marker-SNPs (i.e., in a pairwise comparison) either a coefficient of linkage disequilibrium (*D'*) or, more commonly, an r^2, which is equivalent to the Pearson correlation coefficient. LD can range from 0 (no correlation) to 1 (perfect LD). Inasmuch as most marker-SNPs are not in perfect LD, it is up to the researcher to choose the threshold level beneath which SNPs will be considered not to be in LD. This creates a potentially significant problem for clumping inasmuch as, in the absence of any agreed upon threshold, most researchers select an arbitrarily chosen one (Wray et al. 2013). What is more, the variation in techniques for dealing with LD extends beyond choosing different clumping thresholds (Choi et al. 2020). Some algorithms do not assign all SNPs whose pairwise r^2 with the lead SNP exceeds the user-specified cutoff to the lead SNP's "clump," but only a subset of these SNPs whose distance to the lead SNP is below some cutoff (e.g. 250 kb). Other algorithms involve a second stage in which lead SNPs that are physically close to each other are merged and considered to be a single "locus."

Figure 1. Illustration of Linkage Disequilibrium

Representation of two haplotypes or LD blocks on chromosome 1.



a. Chromosome 1. The region 1q32 on the chromosome (to the right) is amplied in b.
b. Amplication of region 1q32 of chromosome 1.
c. SNPs (or marker SNPs) in region 1q32 of hromosome 1. Beneath each SNP is shown its two forms and where it is located. For example, SNP rs2073186 occurs as either a C or a T. It is located in an "intron" or non-coding part of a gene. SNP rs2243191 occurs as a C or T. It is located in a protein coding region of a gene and causes a change in a single amino acid (the building block of proteins) from S (serine) to F (phenylalanine). UTR stand for an "untranslated region" that is located at the beginning and end of a gene.
d. Three genes in which the SNPs are located – IL-19, IL-20, and IL-24 - as well as intergenic regions (in white).
e. Two LD blocs or haplotypes in region 1q32 Bloc 1 is 30 kb long and bloc 2 is 31 kb. The numbers 1-13 correspond to the SNPs shown in c.
f. Chart showing degree of LD between any two of the 13 SNPs. The numbers in the boxes are a coefcient of linkage disequilibrium (D) which ranges from 0 to 1. For example, the coefcient of LD between SNP 1 (rs2073136) and SNP 3 (rs2243188) is .99, indicating high LD. By contrast, the LD score between SNP 6 (rs2981572) and SNP 9 (rs1518108) is only .7.

ii. Infinite Infinitesimal Alleles

All of the studies considered henceforth involving GWAS, SNP-heritability estimates, and polygenic scores, as well as the vast majority of such studies in general (and certainly all of the most well-known studies) are limited to "whites of European ancestry." The reason for this extraordinary limitation will be considered below (and the reader's forbearance is requested). In the meantime, what is typically mentioned as an aside in the published studies themselves and excluded altogether from accounts of the studies, will be highlight by indicating that the study is limited to whites of European ancestry (WoEA).

In 2010, researchers reported the results of a GWAS of "childhood general cognitive ability" using an array of more than 350,000 SNPs and a population sample of 7900 7-year-old WoEA (Davis et al. 2010). They summed up their findings as follows (2010, 760):

> Despite our large sample size and three-stage design, the genes associated with childhood g ["general intelligence"] remain tantalizingly beyond our current reach providing further evidence for the small effect sizes of individual loci. Larger samples, denser arrays and multiple replications will be necessary in the hunt for the genetic variants that influence human cognitive ability.

While it is odd to claim that their study provided evidence for an association that they failed to find, the assumption that alleles of small effect size underlie heritability, and that larger samples and larger SNP arrays would be needed to find them, is based in part on the results of GWASs of height. Height is considered a highly heritable trait. The first GWAS of height in 2007 examined 365,000 marker-SNPs in an initial sample of nearly 5000 WoEA and identified a single SNP in the "non-translating" region of the HMGA2 gene as being "strongly" associated with height (Weedon et al. 2007). A subsequent study of WoEA identified two additional SNPs (Sanna et al. 2008). As McEvoy and Visscher wrote of these findings at the time (2009, 298):

> [The effect sizes of these alleles] were surprisingly small and possibly bitter to the hopes of many GWAS investigators. The "tall" allele of the SNP in the HMGA2 gene most strongly associated with height (compared to the other allele which is relatively 'short') increases a person's height by just about 0.4 cm and explains only 0.3% of the total phenotypic variation in normal height across the population [of WoEA].

Findings such as these—the purported identification of alleles that explained only a fraction of attribute variation—were given theoretical support by the resurrection of, or reemphasis on, the "infinite infinitesimal allele" model, first proposed by Fisher more than 100 years ago (Fisher 1990 [1918]). Fisher was concerned with variation in what we would call "complex traits": common traits that were not caused by the inheritance of one or two alleles (for example, as seen in so-called monogenic disorders caused by mutations on a single gene). He proposed that the heritability of complex traits involved the inheritance of an indefinitely large ("infinite") number of alleles, each allele contributing a miniscule ("infinitesimal") amount to trait heritability. In modern terms, this would be described as "massive polygenicity."

The contribution of alleles to heritability was primarily additive (the same assumption that underlies the additive model of heritability). That is, the average mathematical effect of two or more alleles on trait heritability in a population is equivalent to the sum of their average individual effects in that population. While the average effect of any individual allele might be insignificant, what was not insignificant was the combined average effect of all of those alleles. The infinite infinitesimal allele model has become a "new" old dogma in behavior genetics.

Crucially, this dogma serves as a justification for the apparently limitless expansion of sample sizes under the assumption that massive samples will be needed to have sufficient power to find many alleles of tiny effect size. To achieve ever larger samples sizes, researchers have taken to combining the data from multiple different GWASs, usually in the form of summary statistics, and performing a metanalysis on the combined data. Combined sample sizes can now exceed one million and involve the testing of 10 million SNPs (Lee et al. 2018a).

Recall that the Bonferroni correction involves dividing the a level (.05) by the number of tests performed. Ten million tests should, theoretically, involve an a level $\leq 5 \times 10^{-9}$, not 5×10^{-8} (the level for 1 million tests). The latter, however, is still commonly employed as the threshold for genome wide significance. In fact, as we shall see, the most highly touted application of GWAS data—polygenic scores—involves setting an a level ≤ 1, which amounts to discarding any statistical correction for multiple hypothesis testing.

Psychologist Eric Turkheimer, also known for the first law of behavior genetics, has commented (2018) that the modus operandi "of increasing sample size endlessly until some unpredicted correlation reaches an arbitrary level of significance sounds a lot like p-hacking." He notes:

> Suppose Brian Wansink, the nutrition researcher who was brought down by revelations of p-hacking and other questionable research practices, had adopted the following strategy in response to criticism of his experiments. Instead of designing individual studies with hypotheses that were always susceptible to hacking, he got funding for an enormous nationwide program that monitored pizza restaurants across the country. The behaviors of hundreds of thousands of pizza eaters were recorded, as were many thousands of tiny characteristics of the environments they ate in. Then they searched for correlations between characteristics of restaurants and eating behaviors, at stringent levels of significance. At first, nothing is significant, so the samples were pushed up from hundreds of thousands to nearly a million pizza eaters, and finally, some significant "hits" emerge. It turns out that eating in a pale green restaurant is associated with a 1 milligram increase in pizza consumption, R^2=.0004, $p < 10^{-8}$.

However, in the eyes of many in the behavior genetics community, Fisher has been proved right, and the practice Turkheimer refers to as "p-hacking" has been vindicated (Visscher and Goddard 2019). Such a conclusion is reached because, unsurprisingly, the larger the sample sizes have grown and the more SNPs tested, the greater the number of SNPs that are said to have genome wide significant correlations with behavioral attributes.

For example, consider several metanalyses of "educational attainment," measured as the number of years of schooling that individuals reported having completed (EduYears). In 2013, researchers performed a metanalysis of the combined data from 64 studies (or "cohorts") across Europe and the United States totaling 126,559 WoEA. Using a standard GWAS significance level

of p=5 x 10^{-8}, they reported one significant SNP for educational attainment (Rietveld et al. 2013). By 2016, using a sample of 293,723 WoEA, researchers reported 74 GWAS significant marker-SNPs (Okbay et al. 2016).

And in 2018, with a combined sample of more than 1 million WoEA from 71 different GWASs and combined microarray data for more than approximately 10 million SNPs, researchers reported the identification of 1,271 GWAS significant marker-SNPs (Lee et al. 2018a). A similar pattern has been observed for a number of other behavioral attributes, such as "general cognitive function" (Davies et al. 2018). In line with the infinite infinitesimal allele model, each of the lead SNPs that have been identified thus far for educational attainment—or intelligence or income—accounts for only a tiny fraction of the supposed heritability of these attributes. For example, Lee et al. (2018a) report that of the 1,271 lead SNPs that they identified for EduYears, the median effect size of each SNP accounts for only 1.7 weeks of schooling.

Given that the postulated causal SNPs could lie anywhere in a region of LD to the lead SNPs, one would assume that without somehow tracking down the "actual" causal alleles, nothing more could be said concerning biological causation. This obstacle, however, has not stopped researchers from proposing elaborate schemas involving the interactions of multiple specific genes. For example, in their study of EduYears, Lee et al. (2018a, p. 1114) note the following:

> [W]e applied the bioinformatic tool DEPICT and found that, relative to other genes, genes near our lead SNPs were overwhelmingly enriched for expression in the central nervous system. The SNPs implicate genes involved in brain-development processes and neuron-to-neuron communication[.]

It is not entirely clear what the authors mean here by "near," "enriched," or "implicate." That said, there is nothing particularly distinctive about a gene being expressed in the brain. Brain tissue is characterized by a high level of gene expression and at least 30–50% of approximately 20,000 protein-coding genes are expressed across all parts of the human brain (Naumova et al. 2013).

The authors also note that they used a "fine-mapping" statistical software program (CAVIARBF) to identify likely causal alleles within 50 kb of their 1274 lead SNPs. One of these was a nonsynonymous SNP (rs61734410) in the gene CACNA1H (meaning that the alternate form of the SNP results in the synthesis of a different amino acid in the resulting protein). As they note, CACNA1H is used to synthesize a calcium channel protein that plays a role in neuronal excitability. However, any involvement of CACNA1H is simply speculation or, perhaps, wishful thinking.

The authors fail to mention other genes "prioritized" by their fine-mapping statistical software program (Lee et al. 2018, Supplementary Table 2), such as PPIP5K2, mutations of which have been associated with various forms of deafness; PRKAG2, transcribed to synthesize a protein involved in responding to energy demands in cardiac muscle and associated with various forms of heart disease; EPB41L3, transcribed to synthesize erythrocyte membrane protein and implicated in several forms of cancer; and many genes of unknown function.

Among the SNPs in LD with their lead SNPs, the authors are cherry picking and highlighting SNPs in genes whose functions they believe accords with the nature of EduYears (i.e., completing high school involves the brain and neurons). This is now a widespread practice (see, for example, the discussion of the neurobiology of income in Hill et al. [2019]). A more formal and perhaps more informative expression for "cherry picking" is the fallacy of incomplete evidence, which consists of pointing to individual cases or data that seem to confirm a particular position and simultaneously skipping the many cases and data that, while perhaps not necessarily contradicting one's position, do not support it.

What is perhaps most misleading about highlighting SNPs in specific genes and using them to construct diagrams of genetic etiology (of risk) is that, according to researchers' own assumptions, the effect of any one of these genes on risk is miniscule (e.g., according to Lee at al. [2018] the SNP rs10189857 has an effect size on EduYears of 0.0158). These sizes are so small that it is inconceivable that the role of any of the unknown SNPs with which the lead SNPs are presumed to be in LD could ever be demonstrated or analyzed in an experimental manner. Of course, one could "knock-out" the proposed gene in a mouse (i.e., engineer a mouse that lacks the gene), or find a family that lacks the gene, which would be a way of distorting the findings by transforming what is insignificant– and one effect among millions–into the sole main effect or cause. Effect sizes this small only make sense in the statistical realm not in the biological realm.

iii. LD, CNVs, and Somatic Mosaicism

When researchers discuss lead "loci" characterized by LD, they generally mean contiguous stretches of DNA extending, on average, several hundred kb upstream and downstream from lead SNPs (250 kb is a commonly cited number). However, LD can span regions of a chromosome more than 1Mb apart in what is known as long range linkage disequilibrium (LRLD), which occurs throughout human genomes. Perhaps the most familiar example of LRLD concerns the short arm of chromosome 6 in a region known as the major histocompatibility complex (MHC), which plays a key role in the immune system, where thousands of SNPs exhibit LRLD across a region in excess of 1.5Mb.

However, recent analysis has revealed that LRLD in excess of 5Mb is common across the human genome (Park 2019; Nelson et al. 2020; Price et al. 2008; Koch et al. 2013). To the extent that LRLD is not taken into account, the result can be a significant overcounting of marker-SNPs presumed to be "independent" (i.e., not in LD with each other). This is not an idle concern. Park et al. (2019) have shown that significant sites from GWAS catalogues mostly overlap regions of LRLD, and that because GWASs typically do not evaluate whether an SNP correlation arises from LDLR, they will incorporate "false signals."

There remains one final, but crucial, point to make concerning SNPs and LD. Copy number variations (CNVs) are a common type of DNA variation, ranging from 50 bp to 10 Mb and involving DNA deletions, duplications, higher order amplifications (e.g., triplications, quadruplications), insertions, and inversions, as well as more complex rearrangements. In human genomes, CNVs involve more DNA sequences than SNPs. CNVs less than 500 kb in size cover 12% (approximately 360 Mb) of the human genome

(Torres et al. 2020), and the number is likely much higher inasmuch as the full extent of CNVs in human DNA is still unknown.

CNVs can be inherited and are responsible for more than ten times the heritable sequence differences in general populations (Pang et al. 2010), but they can also arise somatically during cell divisions at any point in the life course. CNV deletions can result in the loss of an entire gene and duplications can result in multiple copies of a single gene. CNVs are known to play a causal role in a number of disorders including cancer and to affect certain physiological responses (Hu et al. 2018) . For example, the protein cytochrome P450 monooxygenase, synthesized from the CYP2D6 gene, plays an important role in drug metabolism. Because of CNVs, the "normal" number of CYP2D6 genes can range from zero to ten, leading to decreased or increased drug metabolism (Jarvis et al. 2019).

Identifying CNVs still presents a number of technical challenges (Jenko Bizjan et al. 2020) and with the exception of a few very common CNVs (such as those involving the CYP2D6 gene), genomic reference data used with SNPs does not include CNVs. The coexistence of SNPs with unrecognized or unaccounted for CNVs can result in significant distortions. Thus far, about 10% of SNPs across the genome have been found to map to the same genomic regions as common SNPs (Liu et al. 2018).

To give just one example of the kind of distortions this could produce, consider it is common practice to infer the identity of SNPs that have not actually been genotyped on the basis of their presumed LD with marker-SNPs that have been genotyped. CNV losses, for example, could entail that in a segment of a study population, certain SNPs presumed to be in LD do not exist because the segment of DNA on which they are presumed to lie has been deleted. Alternatively, CNV gains could entail that certain SNPs presumed to be in LD exist in multiple copies because the genes in which they are located exist in multiple copies.

iv. A Second Replication Crisis?

Are the marker-SNPs that are identified as "lead" for a given attribute consistently replicated across studies? This question is not easily answered insofar as different researchers use different algorithms for dealing with LD and for identifying lead marker-SNPs, including using different arbitrary thresholds for determining which marker-SNPs are in LD in the first place (see III.i, above). When researchers list their lead marker-SNPs, it is reasonable to ask whether they are replicated across studies.

Let us assume that, in this context, "lead" at the very least denotes a genome wide significant marker-SNP that has a p-value lower than all of the other marker-SNPs with which it is assumed to be in LD (according to whatever criteria are used in this determination). With that as a preliminary, the answer is that lead SNPs are not consistently replicated across studies.

Consider, for example, four large metanalyses of EduYears of WoEA: Lee at al. (2018a), Davies et al. (2018), Lam et al. (2017), and Okbay et al. (2016).[9] Limiting ourselves to the unique marker-SNPs found to be genome wide significant in each study at $p \leq 5x10^{-8}$ yields 4819 SNPs across the four studies. Of these, no SNP was replicated in all of the studies, 3.5% were replicated in more than one study and of these, 97% were replicated in only one other study.

Or consider six large metanalyses of "intelligence"/"cognitive ability" of WoEA: Coleman et al. (2019), Hill et al. (2019), Savage et al. (2018), Davies et al (2018), Lam et al. (2017), and Sniekers et al. (2017). Limiting ourselves to the unique marker-SNPs found to be genome wide significant at $p \leq 5x10^{-8}$ yields 1906 SNPs in total across the six studies. Of these, no SNP was replicated in all of the studies, 11% were replicated in more than one study and of these, 76% were replicated in only one other study.

The poor record of replication for purportedly "independent" and genome wide significant "lead" marker-SNPs goes largely unnoticed - or if it has been noticed it has not been publicly addressed, despite the fact that these same SNPs are often used to construct elaborate stories of genetic causation of risk. What is more, these marker-SNPs, along with SNPs that fail to achieve genome wide significance, are used in the construction of polygenic scores.

IV. Polygenic Scores

The poor record of replication for purportedly "independent" and genome wide significant "lead" marker-SNPs goes largely unnoticed. This is likely because the primary focus today is not on identifying individual genome wide significant SNPs—despite the fact that they are often used to construct elaborate stories of genetic causation (of risk)—but in constructing so-called polygenic scores.

A polygenic score is intended to be a single value estimate of an individual's genetic risk for an attribute, whether Type I diabetes, educational attainment, or anything in-between. Take the example of constructing a polygenic score for a "trait" treated as dichotomous, such as "graduated high school." Suppose that a given GWAS, or a combination of them, contain information on 1 million SNPs for 100,000 persons as well as self-reported information as to whether each respondent graduated high school. For each of the 100,000 members of the sample, the data would show whether the respondent completed high school and which SNP allele she possessed for each of 1 million marker-SNPs on a chip array.

All of the individual data could then be combined into "summary statistics," which present the results of the GWAS in the form of population averages. For each SNP, the summary statistics will identify the "effect allele" (A1), the marker-SNP allele that shows a correlation with high school completion, which can be either positive or negative, and the other allele (A2), the frequency of A1 in the population, and the average effect size of A1 on college completion (in the form of an odds ratio or a a for a quantitative trait); the standard error (SE); and the p-value of the association (see Table 1).

[9] All of the data for all of the lead SNPs for all of these studies is available at GWAS Catalog (https://www.ebi.ac.uk/gwas/).

Suppose that we wanted to construct a newborn's polygenic score for completing college and suppose that her genotype for marker-SNP rs4686944 in Table 1 was GG, which is two "effect," or "risk" alleles. Based on our summary statistics, we would multiply the effect size of this allele (0.029) by the number of risk alleles she has (2). We would repeat this process for each of her 1 million marker-SNPs and sum the results. After assorted tweaking and testing, in theory this sum would tell us the newborn's "genetic risk" for completing high school.

However, in reality it does no such thing. Polygenic scores for educational attainment or intelligence or income have no individual predictive value whatsoever, and it is an open question as to whether any individual polygenic scores for any phenotype have predictive value. This might come as a surprise given all of the hype surrounding the implications of being able to predict someone's genetic risk for doing well in school when they are still in the womb.

What then, do polygenic scores for educational attainment, intelligence, and income actually "predict"? The answer is that they predict a percentage of variance of attribute risk in a population. This is tantamount to constructing a heritability estimate of that population on the basis of differences in allele frequencies.

Still, insofar as polygenic scores predict far less trait variance than SNP-heritability estimates (not to mention twin studies), they are referred to simply as explanations for a certain amount of variance of trait risk in a population. On the population level, a polygenic score is said to be predictive if the average polygenic score in the case group is higher than the average polygenic score in the control group, or, for instance, if the average polygenic score of those in the lowest decile of educational attainment is lower than those in the highest decile.

Consider a polygenic score of "educational attainment" (Plomin and von Stumm 2018). This polygenic score was constructed to predict students' performance on the General Certificate of Secondary Education (GCSE) examination. According to the authors (2018, p. 156), a scatter plot between a polygenic score of educational attainment and GCSE scores (Figure 2a.) indicates, "the difficulty of predicting individual outcomes when the correlation is modest (0.30 in this example)." Squaring this correlation, they estimate that the polygenic score predicts 9% of the variance in risk in their study population, while noting that (2018, p. 156),

> Although higher [polygenic scores] can be seen to predict higher GCSE scores on average, there is great variability between individuals…[I]ndividuals within the lowest and highest [polygenic score] deciles vary widely in school achievement…The overlap in the two distributions is 61%. [see Figure 2b]

However, on a more optimistic note they comment,

> Despite this variability, powerful predictions can be made at the extremes. For example, when the sample was divided into ten equal-sized groups (deciles) on the basis of their EA2 GPS [EA2 is their study population and GPS stands for "genetic polygenic score"], a strong relationship between average EA2 GPS and average GCSE scores emerged that was most evident at the extremes [see the Figure 2c]. Specifically, the average school achievement of individuals in the lowest EA2 GPS decile is at the 28th percentile. For the highest EA2 GPS decile, the average school achievement is at the 68th percentile.

It is common practice to divide a polygenic score into lowest and highest deciles or quintiles—and then note what seems like an impressive difference in the mean prevalence between the lowest and highest decile/quintile. For example, Lee et al. (2018b, p. 129) point out that, "Comparing the 1 and 5 quintiles, there is a 45.4-percentage-point difference in college completion in Add Health and a 35.5-percentage-point difference in the HRS [Add Health and HRS are their two training samples]."

But what at first glance seems like a notable differentiation is simply a general property of small correlations: Given large enough sample sizes, looking at the extremes will produce what appear to be large differences, even though the magnitude of the relationship is small. Finally, Plomin and von Stumm (2018, p. 156), reflecting on the poor individual predictive power of their polygenic score, have recourse to a familiar refrain: "As bigger and better GPSs [genetic polygenic scores] emerge, the predictive power will increase."

Although the focus has been on behavioral attributes, a word here concerning claims of polygenic prediction with more direct medical implications is in order. In two highly publicized studies (Khera et al. 2018; Inouye et al. 2018), researchers reported that polygenic scores for several common disorders including coronary artery disease, could be used as the basis for medical intervention. These scores were derived from, and intended to be predictive for, WoEA. For example, according to Khera et al. (2018), their polygenic score for coronary artery disease could identify 5% of the individuals at highest risk with an odds ratio of 3.34.

However, odds ratios compare odds in the tail ends of a single distribution, which ignores the majority of persons who will or will not develop the disease but fall in the region between the tails of the distribution. If this odds ratio is converted into a measure of predictive performance, it shows a coronary artery disease detection rate of 15% with a 5% false positive rate—and would thus miss 85% of affected individuals (Wald and Old 2019). This is not much better than identifying people at random. At present, polygenic scores are not accepted by the medical community for any kind of medical intervention.

Table 1: Simplied summary statistics for a single SNP-allele for a hypothetical study of college completion.

Marker Name (designation of a particular marker-SNP)	A1 (effect allele)	A2	Frequency A1	Effect size	SE	Pval
rs4686944	G	T	.02041	0.029	0.01087	4.41E-02

i. Training the Score

The construction of a polygenic score begins, as noted, with the summary statistics from various GWAS. This is referred to as the discovery sample. The score is then further developed on a training sample, which must be entirely separate from the discovery sample.

A good deal of confusion is generated by the lack of a consistent terminology for these samples. What this paper refers to as the training sample is also called, variously, the target, validation, prediction, and replication sample.

Apparently, the objective of researchers is to modify the manner in which the polygenic score is constructed to achieve as high an R-squared in the training sample as possible. R-squared, or the incremental R-squared statistic, is a statistical measure that represents the proportion of the variance for a dependent variable (in this case, risk of a phenotype of interest) that is explained by an independent variable or variables (marker-SNPs) in a regression model. The higher the R-squared, the more of the variation in the data that the polygenic score explains.

Researchers have an enormous amount of freedom in determining how to construct their polygenic score so as to achieve the largest possible R-squared. For example, one might assume that in the construction of a polygenic score, only the values associated with marker-SNPs that achieve genome wide significance with a Bonferroni correction of $p=5 \times 10^{-8}$ would be included.

It is now an accepted practice to try out different p-values for inclusion of marker-SNPs in the polygenic score, ranging from $p=5 \times 10^{-8}$, which would entail the inclusion only of genome wide significant marker-SNPs, all the way up to p=1. P=1 is the absence of any statistical correction for (massive) multiple hypothesis testing and results in the inclusion of estimated effect sizes for a million or more marker-SNPs in the construction of a polygenic score.

The freedom to choose p-values in this way, including p=1, so as to achieve the largest R-squared possible, might seem a willful disregard for the consequences of multiple hypothesis testing and data mining. But as the saying goes, the proof is in the pudding. For example, Lee et al. (2018a), in constructing their polygenic score of EduYears, tried out four different p-value thresholds: $p<5 \times 10^{-8}$, 5×10^{-5}, 5×10^{-3} , and 1. They note that their R-squared increased from 3.2% at $p<5 \times 10^{-8}$ to 9.4% at $p \leq 1$. Trying out different p-values and choosing the one that leads to the construction of a polygenic score with the highest R-squared (which is invariably p=1) is now common practice.

Therefore, choosing a significance threshold of $p \leq 1$ is justified by its appearing to work. It enables researchers to achieve a higher R-squared. Moreover, it also accords with the narrative of infinite infinitesimal alleles. Because the effects of the SNPs are infinitesimal, the Bonferroni correction—or any correction—is too stringent. Furthermore, what is statistically significant is not the effect of any individual SNP but their combined effect. Doubtless any polygenic score constructed in this manner will contain many marker-SNPs that have no effect whatsoever on phenotypic risk, but by containing many more that do and otherwise would have been excluded, a threshold of $p \leq 1$ is justified.

Figure 2. The use of genome-wide polygenic scores to predict individual outcomes

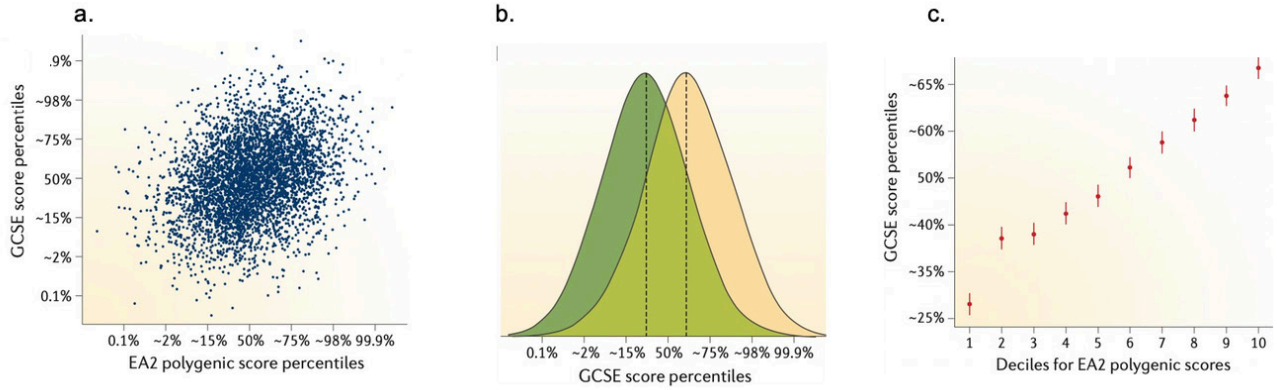


Figure 1. Adapted from Plomin and von Stumm (2018).

a. Scatterplot between EA2 polygenic score percentiles and GCSE score percentiles; b. Population distribution of polygenic scores (normally distributed). Shows that the average polygenic score for those who score higher on the GCSE is slightly higher. On this basis, the authors predict that their polygenic score predicts 9% of the variance in genetic risk. The overlap in the two distributions is 61%; c. Division of the sample into ten equal-sized groups (deciles) on the basis of their EA2 polygenic score. Shows the relationship between average EA2 polygenic scores and average GCSE scores.

There are many, many other decisions that researchers can make in their quest for the highest R-squared including trying different algorithms embodied in different software programs to determine the weighting of individual SNPs, or determining how many principal components to use in trying to deal with population stratification (see below), or how to account for "winner's curse," or which cutoffs to use for minor allele frequency. In addition to setting the statistical threshold at $p \leq 1$, Lee et al. (2018a, p. 1115) also note that they were able to achieve a yet higher R-squared of 11.4%—the figure they ended up using in the end—when, in addition to using $p \leq 1$, they switched from removing SNPs in linkage disequilibrium with each other to using the software LDpred, "a Bayesian method that weights each SNP by (an approximation to) the posterior mean of its conditional effect, given other SNPs." In this study, when the authors controlled for household income and the educational attainment of the mother or father, the score's incremental R-squared dropped to 4.6%.

All of this "freedom" on the part of researchers is a recipe for *model overfitting*. When a researcher has so many proverbial "degrees of freedom," that is, when one is free to try so many different analytic alternatives and value thresholds to achieve a preferred result, including setting significance thresholds, the likelihood of creating a manufactured rather than real statistical correlation is quite high.

Defenders of this approach might argue that they employ all sorts of significance tests every step of the way, but the tests used, and the values deemed confirmatory, are themselves largely a matter of researcher discretion. All data sets have random quirks; an overfit model will incorporate these quirks to such an extent that that the model ends up explaining the random error present in the data. Hence, an overfit model will not be *generalizable* because it describes the random error in the data rather than the relationships between variables.

Ultimately, the regression coefficients represent noise rather than genuine relationships in the population. Inflated R-squared values are a symptom of overfit models, and overfit models are a common occurrence when researchers chase a high R-squared.

The problem of overfitting is exacerbated by the fact that in constructing polygenic scores, there are more predictors (p), in the form of individual SNPs, than number of persons (n) in the sample. For example, the study of Lee et al. (2018) has a sample size of 1.1 million and 7.1 million predictors (i.e., SNPs). Each person (i.e. one observation) has millions of possible gene combinations. When there are more predictors than samples in the dataset, the researcher is confronted with a problem of "big-p, little-n" ($p >> n$), sometimes referred to as the "curse of dimensionality" (Altman and Krzywinski 2018).

In short, it describes what happens when you add more and more variables or "dimensions" (in this case, SNPs) to a multivariate model. The more dimensions added to a data set, the more difficult prediction becomes. While one might assume that more is better, when it comes to adding variables, the opposite is true. Each added variable results in an exponential decrease in predictive power. Such dimensionality can cause models to behave one way when devised in one sample, and in a completely different and unpredictable way when applied to a different sample (Hastie and Tibshirani 2003).

It is for this reason that, according to the "gold standard," a polygenic score developed on a training sample should be tested, prior to any claims that it actually predicts anything, on a third sample entirely separate from both the discovery *and* training samples (Choi et al. 2020). For the sake of clarity, let us refer to this as the "*independent validation data set.*" *Frequently, this is not done.*

Sometimes, researchers will remove one or two study "cohorts" from their discovery sample (which consists of numerous cohorts), use these excluded cohorts as their training sample, and publish results derived from the training sample. Given the impetus to combine ever larger numbers of cohorts in an attempt to create ever larger discovery samples for metanalyses, finding a sufficiently large third independent sample is becoming increasingly difficult.

To be sure, there are a number of statistical algorithms designed to deal with the problem of model overfitting in the absence of a third independent sample and to estimate how accurately a predictive model will perform in practice, such as "cross-validation." Cross-validation uses a single training sample that is then split into smaller "training" and "validation" subsets. In discussing a recent study of the performance of cross-validation based on real-world fraud experiments using 2013 to 2016 Medicare claims data, the authors note (Bauder et al. 2019, p. 19):

> In real-world production applications, it is critical to establish a model's usefulness by validating it on completely new input data, and not just using the cross-validation results on a single historical dataset. In this paper, we present results for both evaluation methods, to include performance comparisons…Using this Medicare case study, we assess the fraud detection performance, across three learners, for both model evaluation methods. We find that using the separate training and [validation] sets generally outperforms cross-validation, indicating a better real-world model performance evaluation.

Simply put, estimating predictive performance is inferior to observing it in action in the real world.

There is growing evidence that polygenic scores consistently are *overfit* models. For example, Mostafavi et al. (2020, p. 6) demonstrated that "the portability of a polygenic score [for WoEA] can vary markedly depending on sample characteristics of both the original GWAS and the prediction set, and that this variation in prediction accuracy can be substantial."

Variation in the samples for such things as percentage of male versus female participants, or the percentage of persons in various age or socio-economic categories, were all shown to have a substantial impact on polygenic score accuracy. In fact, all published polygenic scores are overfit models to the extent that they are limited to WoEA and cannot be applied to persons of any other ancestry. They are not "portable" outside this category; in other words, they are not generalizable. However, as we shall see, limiting studies to WoEA does not solve the problem it was intended to.

V. Population Stratification

All of the studies cited, as well as the vast majority of such studies today, are performed exclusively on what are referred to as "persons of European descent," referred to here as "whites of European ancestry" (WoEA). Non-WoEA are typically viewed in terms of other racial categories, which include, at the very least, Africans, Asians, Native Americans, and Oceanians.

Persons considered non-WoEA are intentionally excluded from every study population. The reason is that when non-WoEA are included in either the discovery, training, or third, independent replication sample (assuming that there is one), the predictive power—that ever-important R-squared—goes way down. As Lee et al. note of their study (2018a, p. 1115):

> Because the discovery sample used to construct the score consisted of individuals of European ancestry, we would not expect the predictive power of our score to be as high in other ancestry groups. Indeed, when. . .used to predict EduYears in a sample of African-Americans. . .the score only has an incremental R^2 of 1.6%, implying an attenuation of 85%.

Of course, their discovery sample was limited to WoEA by design.

Why then, must a polygenic score be constructed exclusively from a discovery sample of WoEA and be applied, for purposes of prediction, only to WoEA? Why, otherwise, will the R-squared go way down? The explanation comprises two elements: first, the well-known phenomenon of population differences in such things as allele frequencies, haplotypes, degree of linkage disequilibrium, and degree of genetic diversity, and second, the fact that marker-SNPs and maps of linkage disequilibrium upon which both GWAS and polygenic scores depend have been developed almost exclusively from studies of WoEA (Popejoy and Fullerton 2016).

Differences in what can be called genetic population characteristics—which are often characterized as differences in allele frequencies—are the result of different population histories and can be influenced by such things as differences in migratory patterns, founder events and population bottlenecks (loss of genetic variation that occurs when a new population is established by a very small number of individuals from a larger population), population expansions, relative population isolation, endogamy, inbreeding, adaptive pressures, and genetic drift (a random change in allele frequencies). With the exception of genetic drift, which is stochastic, all of these processes are environmental and often cultural, as in the case of endogamy, for example.10 Genetic population characteristics are highly correlated with geography although the sharing of certain allele frequencies between populations need not entail a shared ancestry or geographic origin.

The assertion that a study has been limited to WoEA generates the impression that WoEA constitutes a single, clearly delineated population whose members are defined by certain shared population genetic characteristics. While researchers are aware of the existence of population genetic differences *within* WoEA, they believe that these can effectively be dealt with in a straightforward manner. Genetic difference of a sufficient magnitude to disrupt the construction of polygenic scores only occurs via the inclusion of non-WoEA.

This belief—that researchers can effectively deal with what is known as population "structure" within WoEA, but not between WoEA and other ancestral groups—reinforces the idea of intra-group (relative) genetic homogeneity, inter-group genetic heterogeneity, and the significance of the system of classification (e.g. folk racial categories) based upon these presumed genetic differences.

The problem with dividing all populations into five or six categories—Europeans, Africans, Asians, Native Americans, Oceanians, as well as various "admixed populations" (such as Hispanic)—is that it presupposes clear genetic boundaries between these groups and ignores the significant amount of genetic heterogeneity within these groups. Both of these errors are important and of course, not entirely separate, but my focus here is on the latter.

A population is said to be *structured* when it contains subpopulations that are often, but not exclusively, distinguished by geographic location and that exhibit systematic differences in population genetic characteristics (or allele frequencies). WoEA is clearly a structured population at every level, and the more fine-grained one's analysis, the more nested levels of population structure appear.

At the broadest level, one can speak of all Europeans as being descendants of three original migratory populations from the Near East, North Asia, and (geographic) West Europe, and as bearing different mixes of genetic population characteristics that can be traced back to these groups (Lazaridis et al. 2014). As one moves away from this level, which is itself a heterogenous mix, and toward greater specificity, a good deal of additional genetic heterogeneity arises.

To illustrate, the Finns experienced two major population bottlenecks over their history which reduced the population size as well as the genetic di-

---

[10] As an aside, it is worth noting that in the worldview in which all beliefs (or differences in all beliefs) that one would call cultural, such as the propriety of endogamy, are influenced by differences in allele frequencies, the practice of endogamy itself can be a cause of differences in allele frequencies.

[11]For examples of non-geographic/aancestral sharing of allele frequencies, consider the following. Tay-Sachs is a rare somatic recessive disorder caused by mutations in the HEXA gene. It is estimated that among "Europeans," 1 in 300 persons is a carrier of the Tay-Sachs associated allele (a heterozygote) and 1 in 360,000 is an affected homozygote. By contrast, among Ashkenazi Jews (or persons of Ashkenazi Jewish descent) 1 in 27 are carrier and 1 in 2900 have the disorder; among French Canadians of South Eastern Quebec, 1 in 30 is a carrier and roughly 1 in 3600 have the disorder; among the Cajun of Southern Louisiana 1 in 30 is a carrier and approximately 1 in 3600 have the disorder; and there are similarly high rates among certain Amish communities. With the exception of the French Canadians of South Eastern Quebec and the Cajuns of Southern Louisiana, who share a common ancestry, none of the differences in allele frequencies among the remaining population can be due to common ancestry. Alternatively, consider blood type. The ABO blood group antigens are transcribed from the ABO gene, which has three alternative allelic forms—A, B, and O. About 21% of all humans have at least one A allele, with the highest frequencies found in small, unrelated populations, and in particular the Blackfoot Indians of Montana (30-35%), the Australian Aborigines (many groups are 40-53%), and the Lapps, or Saami people, of Northern Scandinavia (50-90%).

versity. As a result (and perhaps also, as a result of cultural practices) Finns have a degree of genetic homogeneity that distinguishes them from other Europeans (Locke et al. 2019). But for all of their "genetic homogeneity," Finns themselves exhibit regional population genetic differences such as that between eastern and western Finland (Lappalainen et al. 2006).

Consider an analysis of the genetic population structure of the United Kingdom. The ancestral population of the United Kingdom, which includes Great Britain, Northern Ireland, and many smaller islands (including the Hebrides, Shetland, and Orkneys), is structured, as are populations throughout the rest of Europe (Leslie et al. 2015). Genetic analysis has revealed a fine-scaled genetic structure of the British population composed of 17 distinct "clusters" that are highly localized (Leslie et al. 2015, p. 312):

> Examples of fine-scale differentiation include the separation of: islands within Orkney; Devon from Cornwall; and the Welsh/English borders from surrounding areas. The edges between clusters follow natural geographical boundaries in some instances, for example, between Devon and Cornwall (boundaries the Tamar Estuary and Bodmin Moor), and Orkney is separated by sea from Scotland. However, in many instances clusters span geographic boundaries; for example, the clusters in Northern Ireland span the sea to Scotland.

Structured populations, which are most populations, are considered an omnipresent threat to the validity of genetic association studies due to *population stratification*. Population stratification arises when differences in allele frequencies between cases and controls, ascribed to genetic risk factors, are actually due to ancestry related population genetic differences.

For example, suppose our study population is composed of persons of ancestry A and ancestry B, who as a result of differences in migratory histories and/or patterns of endogamy and the like, exhibit differences in the frequencies of certain alleles. At the same time, there are structural social differences between A and B: Population A has been historically discriminated against and denied social opportunities, while population B has enjoyed all manner of social advantage. Suppose our "phenotype" was income or years of education. In constructing a polygenic score, the allelic differences we might ascribe to differences in "income predisposing alleles" might in fact be due to allelic differences between A and B arising from different ancestral population histories, and these population-level allelic differences would be associated with different social environments.

The two most widely used statistical methods for dealing with population structure are principal components analysis (PCA) and linear mixed models (LLM). PCA is used to identify patterns, or "axes of variation" that explain the greatest amount of variance in SNP allele frequency in the sample. In a GWAS, what typically accounts for the greatest amount of variation in SNP allele frequency (the first principal component) is not the attribute of interest, but the ancestries of the participants which often corresponds with a particular geographical region.

Genotypes and phenotypes are adjusted by amounts attributable to ancestry along each axis. It is fairly common for researchers to do a PCA on the entire set of genotype data and then use the first 5, 10, or 20 principal components as covariates in the association model. An alternative widely used approach is LLM, which incorporates both fixed and random effects, population structure being treated as a random effect.

Do these methods effectively deal with the problem of population stratification? There is strong evidence that they do not.

In several studies, researchers reported that the average polygenic score for height increased from south-to-north across Europe (i.e., exhibited a "latitudinal cline"), paralleling average population differences in height from Italy to the Netherlands (Turchin et al. 2012; Berg and Coop 2014; Robinson et al. 2015; Zoledziewska et al. 2015; Berg et al. 2019b; Racimo et al. 2018; Guo et al. 2018). The results of these studies were highly touted not only as multiply-replicated polygenic scores for height but also as an example of polygenic adaptation. Most of these studies were based on data assembled by an international collaborative effort known as the GIANT Consortium (The Genetic Investigation of ANthropometric Traits), consisting of the combined summary statistics from 79 individual GWASs totaling 253,288 persons of European ancestry from across Europe.

In subsequent studies (Berg et al. 2019a; Sohail et al. 2019) researchers, including some of the same individuals who had earlier published studies reporting evidence of polygenic adaptation for height, attempted to replicate these findings using a larger sample from the UK Biobank. The UK Biobank contains GWAS and "health and well-being" data of 500,000 volunteer participants from the UK. Researchers limited themselves to participants who self-identified as being of "white British ancestry" (N=336,474). This study population was both larger and more homogeneous in terms of ancestry than the population that comprised the GIANT Consortium from which the polygenic scores for height had been derived. They failed to replicate the original findings. As Berg et al. (2019a, p. 14) noted of their results, "[W]hat once appeared an ironclad example of population genetic evidence for polygenic adaptation now lacks any strong support."

What both Berg et al. (2019a) and Sohail et al. (2019) concluded is that the differences in the polygenic scores for height were picking up ancestral population differences in allele frequencies between groups (such as the Italians and the Swedes) that had nothing to do with height. And they knew this because these scores did not identify differences in height in a more homogeneous population (or at the very least, the differences were significantly attenuated). Of their findings, Sohail et al. (2019, p. 1) commented, "[M]ethods for correcting for population stratification in GWAS may not always be sufficient for polygenic trait analyses[.]"

Doubts concerning the ability of current methodologies to deal with population stratification have continued. The UK Biobank, a large segment of which has been assumed to represent a single relatively homogeneous population—the population of persons of "white British ancestry"—itself exhibits population structure as indicated by principal components analysis (Cook et al. 2020; Haworth et al. 2019). This is not surprising given that, as noted above, "white British ancestry" does not constitute a structure-free population but rather a population composed of 17 distinct "clusters" that are highly geographically localized (Leslie et al. 2015).

Both Cook et al. (2020) and Haworth et al. (2019), showed that polygenic scores for traits including education, income, body mass index, hypertension, smoking, and alcohol consumption were associated with birth location. These associations with geography persisted even after the use of principal component analysis (involving up to 100 principal components [Zaidi and Mathieson 2020]) and a linear mixed model. Differences in all of these attributes—education, income, body mass index, hypertension, smoking, and alcohol consumption—and many more are known to vary by geographic region throughout Great Britain.

At the same time, allele frequencies are known to differ throughout Great Britain by geographic region due to differences in ancestral populations, providing a source of covariance between genotype and attribute that will lead to population stratification. As Haworth et al. note (2019, p. 6), "[T]his phenomenon is important, both as a source of ecological-level covariance between genotypes and geographically heterogeneous complex traits, and because of its apparent persistence across different analytical contexts and modes of statistical adjustment."

It is common for researchers to assume that siblings from the same parents are immune to population stratification because the genetic differences between them will be due to the random partitioning of parental genomes. Hence, it is common for researchers to attempt to replicate GWASs and polygenic scores using family-based studies.

Recall that according to two separate studies, the once apparently iron-clad example of polygenic adaptation for height was shown to be confounded by population stratification. Prior to this, one of the confirmatory studies involved using SNPs ascertained from a sibling-based GWAS involving roughly 17,500 sibling pairs from European cohorts (Robinson et al. 2015). They reported that the north-south frequency gradient (i.e., polygenic adaptation) replicated using the sibling data. These results were consistent with, and in some cases even stronger than, the associations found using the GIANT data, but were inconsistent with the results obtained using the UK Biobank data. As Berg et al. note (2019a, pp. 3-4):

> [M]ultiple lines of evidence suggest that population-structure confounding in GIANT and R15 sibs [the data set of approximately 17,500 sibling pairs] is the main driver of the discrepancy with UKB [UK Biobank]-based analyses… [B]oth the GIANT and R15 sibs GWAS are confounded due to stratification along the North-South gradient where signals of selection were previously reported.

Similar doubts concerning the absence of stratification in sibling data, also in relation to variation in height, are expressed by Cox et al. (2019). Zaidi and Mathieson (2020) have recently demonstrated that when SNPs are ascertained using a standard (non-sibling) GWAS discovery sample and the effect sizes are re-estimated in siblings, stratification in the polygenic score persists.

VI. Genetic Estimate Breeding Values

The end of a discussion of population structure is an appropriate place to mention genetic estimate breeding values (GEBV). Polygenic scores are based upon GEBV which are now used extensively in plant and animal breeding. They constitute the closest approximation to an empirical demonstration of the validity of polygenic scores and the assumption that genetic risk is influenced by an "infinite" number of infinitesimal alleles acting in an additive manner. Estimated breeding values (EBVs) are estimates of the likelihood that the offspring of an animal will possess a given trait considered valuable.

Breeding for milk yield is probably the most well-known example. Predicting the EBV value for a dam is easy: The dams with the greatest milk yield will be chosen as breeders. Traditionally, the bulls that sired daughters with the greatest milk yield would be selected to breed with these dams. The selection of sires is now guided by genomic estimated breeding values that have been derived from the genotyping of bulls that have shown a high rate of success in producing female offspring with high milk yields (Calus 2010). These bulls are generally mated with high milk producing dams.

As noted in an article comparing polygenic scores and GEBV (Wray et al. 2019, p. 1132), "As in human genetics, although the goal for PRS [polygenic risk scores or polygenic scores] is prediction of the phenotype, the accuracy of prediction for an individual is low; hence, the value of polygenic scores is, like in livestock genetics, best interpreted at the group level." To be clear, what GEBV is predicting in this example is which bulls, when mated with high milk yield dams, will produce female offspring with a high yield (a trait the bulls themselves, of course, do not possess). In human terms, this would be like attempting to develop a polygenic score for fathers to predict the quality of their daughters' breast milk.

How do livestock differ from humans genetically (other than that humans have a human genome and cows have a bovine genome)? In most breeds of livestock there is a significant lack of genetic diversity. In dairy cattle, due to artificial insemination, bulls that have been selected for their "genetic merit" for siring offspring with high milk production traits can sire hundreds of thousands of offspring.

Cattle breeds have a very small effective population size, meaning that the number of individuals in the population contributing genes to the next population is small. The international black and white Holstein dairy cattle population is 25 million but the current effective population size (Ne) is estimated to be only 50 (Kim and Kirkpatrick 2009). It estimated that the ancestry of 99% of male bulls alive today can be traced back to two bulls, both born in the 1960s (Yue et al. 2015). It is for this reason that in cattle breeds, as in most livestock, there is a significant lack of genetic diversity and deleterious health effects related to inbreeding are considered a constant threat (Bjelland et al. 2013).

As a result, haplotype blocks in dairy cattle are about double the length of human linkage disequilibrium and generate linkage disequilibrium across chromosomes. There is no population stratification within a herd of livestock. At the same time, every aspect of the environment of the breeding stock of a herd is carefully regulated, and all animals share identical environments. When the methods of GEBV have been applied to natural populations, even the most basic predictions have failed (Charmantier et al. 2014). According to the additive model that underlies polygenic scores, each SNP acts in isolation from every other SNP: There is neither epistasis (genotype x genotype, G x G) interaction nor genotype x environment

(G x E) interaction. Each SNP contributes a fractional amount to filling the "space" occupied by genetic risk, and the environment fills the remainder of the space. According to this additive model, the overall risk (or liability) of persons with different levels of "genetic risk" will increase exactly the same amount as they move from "low risk" to "high risk" environments (Figure 3a.). Using GEBV, however, animal breeders tell a different story (Souza et al. 2016, p. 207):

> The genetic merit of an animal can be significantly influenced by changes in the breeding environment, and the progenies of a sire may not repeat the performance of their progenitors if they are raised in different micro-regions or farms, denoting the need for care when buying sires or semen due to the presence of genotype-environment interactions (GEI).

Figure 3b. shows the performance of bulls with different breeding values for milk yield as measured by the milk yield of their daughters, in environments characterized by increasingly negative conditions (as measured by a variety of environmental factors—climatic, sanitary, alimentary—known to adversely affect milk yield). It shows unpredictable G x E interactions, where G is the GEBV of the sire, and E is the daughters' environmental gradient. Bulls with the highest GEBV, based on the performance of daughters raised in the most positive environment, exhibit the lowest GEBV when they are raised in the most negative environment and vice versa, a phenomenon known as "cross-over." G x E interaction results in a change in GEBV itself.

G x E occurs because environmental changes bring about changes in *genotype expression,* that is, whether, when, where (in which cells or tissues), and in what manner a gene is transcribed and whether, when, where, and in what manner the product of gene transcription is synthesized into a non-coding RNA or potentially a multitude of different proteins.

Genotype expression is not an all-or-nothing matter (i.e., in a certain tissue, a gene either can or cannot be transcribed) but rather a state of continual interchange between DNA sequences and the cellular and extracellular environments. The result of this is that a polygenic score, assuming that it is really based upon differences in gene frequencies that bear some causal relation to genetic risk, will not be generalizable across different environments.

And in this case, the non-generalizability would result not from model-overfitting or population stratification but from the nature of genetics itself.

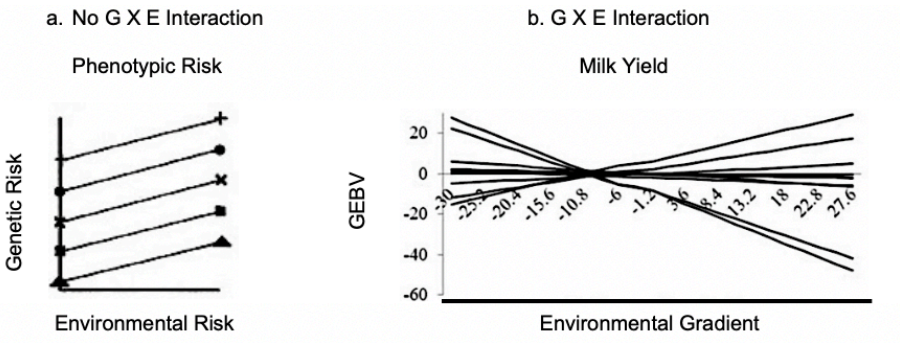
VII. Genetic Heterogeneity

Researchers in behavior genetics have the tendency to treat complex social behaviors as if they are single, well defined phenotypes, and to assume that each of these phenotypes must have a single set (or "core" set) of risk alleles. In other words, they largely deny genetic heterogeneity, which suggests that different polymorphisms can be risk factors for the same phenotype (Maier et al. 2018).

On the one hand, genetic heterogeneity could mean that there are different genetic risk factors for the same disorder/phenotype (e.g., a form of thyroid disease known as thyroid dyshormonogenesis can result from mutations in one of several different genes, including DUOX2, SLC5A5, TG, and TPO). On the other hand, genetic heterogeneity can indicate something important about the nature of the phenotype we are considering.

Consider chronic nausea (CN). How does CN differ from EduYears? Even asking this question seriously is an indication that something has gone very wrong with the study of human behavior. And indeed, it has, but what is important in the current context is that if researchers were to characterize CN in the same manner as EduYears is characterized, they would

Figure 3. G x E interaction



a. Additive model of polygenic risk in which there is no G x E. The reaction to "environmental risk" (x- axis) is exactly the same, regardless of the level of genetic risk (y-axis).

b. Extensive G X E interaction graphed in what is known as a "reaction norm." Different GEBVs (=genetic risk) respond in signicantly different ways in different environmental gradients (= environmental risk). There is signicant cross-over and "re-ranking" of the animals.

1b is adapted from Souza et al. (2019)

[12] For more on GEBV and G x E, see Cheruiyot et al. (2020), Santos et al. (2020), Mulim et al. (2020).

seriously *mischaracterize* it. EduYears is treated as a distinct phenotype with a distinct set of genetic risk factors (ignoring here the fact that the "phenotype" EduYears is going to mean different things [i.e., be a different phenotype] in different educational systems).

In contrast, while CN may be a phenotype, it is a phenotype that can be caused by numerous other phenotypes. Put another way, it is a *symptom* that can have innumerable *causes/risk factors* (e.g. gastroesophageal reflux disease, peptic ulcer disease, gastroparesis, bowel obstruction, migraines, postural hypotension [abnormal change in heart rate when changing posture], etc.). Each of these is a different phenotype (i.e., a different disorder) which may or may not have its own set of genetic and/or environmental risk factors. At the same time, CN could have multiple *different interacting* risk factors: CN risk could be due to peptic ulcer and reflux disease, or gastroparesis and migraines, or extreme emotional distress and postural hypotension.

Each of these phenotypes—peptic ulcer, reflux disease, migraines—can itself be due to different phenotypes with different genetic and/or environmental risk factors, and so on. Thus, CN risk might be due to migraines (one out of many possible "risk phenotypes" for CN) and migraines might be due to emotional stress (one out of many possible "risk phenotypes" for migraines). At the same time, the causal structure of risk is not unidirectional (e.g., stress could be a risk factor for migraines, but migraines could also be a risk factor for stress). Why is EduYears treated as having a genetic risk structure so much simpler than chronic nausea?

Which brings us to the following question: When researchers claim that polygenic scores can predict EduYears or income, are they actually claiming that these are phenotypes with their own distinct set of genetic risk factors? At times, the answer is unclear. For example, in a paper titled, "Genome-wide analysis identifies molecular systems and 149 genetic loci associated with income," the authors make the following comments concerning their phenotype (Hill et al. 2019a, p. 11):

> A further limitation is that molecular genetic analyses of phenotypes, such as intelligence, income or SEP [socioeconomic position], appear prone to being misinterpreted. Such misunderstandings include describing associated variants as, genes for income…[G]enetic variants do not act directly on income; instead, genetic variants are associated with partly heritable traits (such as intelligence, conscientiousness, health, etc.), which have their own complex gene-to-phenotype paths (including neural variables) and are ultimately associated with income.

This assertion runs contrary to much of what the authors say elsewhere in this paper. They note that (Hill et al. 2019a, p. 6) "genetic correlations [a measure of the percentage of SNPs shared between two attributes] were calculated between household income and a set of 27 data sets covering psychological traits, mental health, health and well-being, anthropometric traits, metabolic traits and reproduction."

However, if income is not "directly" heritable but rather influenced by other traits that are heritable, then it has no genetic risk-SNPs of its own to be compared with the risk-SNPs of other phenotypes. What, precisely, are they correlating with what? To be sure, illness, including mental illness, can adversely affect someone's income. So, are we predicting heritable health problems, which have an indirect effect on income, which is itself non-heritable? Furthermore, what is the difference between indirect causation of the kind the authors discuss here and *confounding?*

Consider the following well-known example, adapted to the present. If we were constructing SNP heritability estimates and polygenic risk scores for the phenotype "earring-wearing" using data from the UK Biobank, and did not take into account sex differences, we could predict earring wearing on the basis of a polygenic score that was actually predicting sex differences. Would it be legitimate to call sex a heritable phenotype that indirectly affects earring wearing? Sex is, of course, a heritable phenotype, but any study that resulted in a polygenic score that was predicting sex as opposed to "earring wearing" *per se* would be accused of confounding.

Another example: If we did not take into account the possibility of population stratification, we might construct a polygenic score that predicted being a member of an ethnic group that was discriminated against. While the generalizability of scores might be limited, in the population of interest it would in fact (let us assume) predict average differences in income. Therefore, we could say that our risk score predicts being of a certain ethnicity and being a member of this ethnicity indirectly (via social discrimination) predicts income. Why is this kind of indirect causation considered illegitimate (i.e., a form of population stratification)? And how is this different from a score predicting, say, schizophrenia, bipolar disorder, and depression, and indirectly predicting income and SES because serious mental illness is associated with poverty (Sylvestre et al. 2018)?

This question applies not just to studies of the genetics of income, but to the genetics of all complex social behaviors because if we assume, *arguendo,* that polygenic scores are actually predicting a certain amount of trait variance in a population, there is a lingering question as to what, precisely, the genetic risk factors are risk factors for.

Conclusion

Researchers in behavior genetics, reflecting on candidate gene association studies, warned that results should be deemed tentative until they have been replicated in multiple large samples, that the failure to exercise caution could hamper scientific progress, that extra caution should be exercised in new and "hot" areas of research, and that the failure of CGA studies was a cautionary tale.

Have the lessons of the cautionary tale been learnt? The answer is quite clear. We are still seeing the same incautiousness, the same exaggerated claims, and the same hype. Just as there were thousands of published CGA studies in which researchers heralded the ability to predict such things as school performance, income, and intelligence on the basis of differences in allele frequencies of a handful of polymorphisms, we are now confronted with an ever-growing number of studies in which researchers herald the ability to predict exactly the same things on the basis of millions of polymorphisms. Most importantly, this research is plagued by many of the same methodological problems that brought down CGA studies.

We are beginning to see published studies pointing to significant problems with the current methodologies, problems related to model overfitting, population stratification, and long-range linkage disequilibrium, to name a few. One can only hope that they are taken more seriously than were the early warnings concerning CGA studies. The failures of that era have never been given a proper post-mortem. One can only hope that, decades from now, the same question that was asked of that era will not have to be repeated: How on earth could we have spent 20 years and hundreds of millions of dollars studying *pure noise?*

Glossary

Allele: One of two or more alternate forms of a gene or any segment of DNA. Persons inherit two copies of each allele, one from each parent.

Base pair (bp): A pair of complimentary nucleotides on a DNA strand. It is used as a relative unit of measure, (e.g., two polymorphisms are 100 bp apart).

"Big-p, little-n" (p >> n): Refers to problems that arise when there are more predictors than samples (n) in a multivariate model. Also called "the curse of dimensionality."

Bonferroni correction: A statistical technique for dealing with the problem of multiple hypothesis testing—that is, as one performs more tests, the likelihood of false positives (Type I errors) increases. In a Bonferroni correction, the p-value of .05 is divided by the number of tests performed.

Candidate gene association study: A type of study in which a researcher hypothesizes a relationship between a particular polymorphism and a particular phenotype on the basis of the presumed physiological effect of that polymorphism. The hypothesis is then tested on a study sample.

Copy number variation (CNV): A common type of DNA variation, ranging from 50 bp – 10 Mb and involving DNA deletions, duplications, higher order amplifications (e.g., triplications, quadruplications), insertions, and inversions, as well as more complex rearrangements. In human genomes, CNVs involve more DNA sequences than SNPs.

Diallelic: A single nucleotide polymorphism (SNP) that has two possible variant forms.

Discovery sample: The sample (or samples) on which a GWAS (or GWASs) is performed.

G x E interaction: Occurs when the effect of the environmental exposure on a certain outcome is strongly influenced or contingent upon genotype and vice versa (gene effect on the outcome is contingent on exposure).

G x G interaction: Also referred to as epistasis. Occurs when the effect of one polymorphism on a phenotype is modified by another polymorphism or polymorphisms. In biological epistasis, the gene-gene interaction has a biological basis. Statistical epistasis describes deviation from additivity in a linear statistical model.

GEBV: Genetic estimate breeding values. A GEBV is used animal (as well as plant) breeding to predict whether offspring will have a trait deemed valuable on the basis of the genotype of one of the parents. Polygenic scores are based on GEBV.

Genetic heterogeneity: The phenomenon is which different polymorphisms or mutations of different genes act as risk factors for the same phenotype.

Genome wide association study: A study that involves investigating DNA markers across large sections of the genomes of a large number of persons to find genetic variations associated with a particular phenotype.

Haplotype: A group of two or more alleles that are inherited together (and are conventionally thought of as lying in close proximity).

Heritability: A measure of the amount of variation in a phenotype (or phenotypic risk) among the members of a given population, at a given time, that can be correlated with members' genetic variation.

Independent validation sample: A sample separate from both the discovery and training samples on which the results of a polygenic score can be tested.

Kilobase (kb): 1 kilobase = 1000 bp,

Linkage Disequilibrium (LD): Alleles that tend to be inherited together (i.e., form a haplotype) are said to be in linkage disequilibrium.

MAOA: A gene for the enzyme monoamine oxidase A, which plays a role in the regulation of several different neurotransmitters. Polymorphisms of MAOA were often associated with a wide variety of behaviors via candidate gene association studies.

Marker-SNP: Any of the million or more SNPs across the human genome that are examined in a GWAS. Marker SNPs are not themselves thought to be causal. Rather, they are used to locate regions in a genome where (unknown) causal alleles are thought to be located.

Megabase (Mb): 1 Mb = 1,000,000 bp/1000 kb.

Model overfitting: The phenomenon in which a model refers to quirks in the data (or noise) rather than real relationships between the variables. Can occur when researchers have too much freedom to manipulate the data.

Null hypothesis: When the null hypothesis is true, there is no relationship between the two variables being studied; results showing a relationship are due to chance alone.

P-hacking: The manipulation of data in a way that produces a desired p-value. P-hacking is typically done through manipulation of "researcher degrees of freedom," or the decisions made by the investigator. These include when to stop collecting data, whether or not the data will be transformed, which statistical tests (and parameters) will be used, and so on.

P-value: Represents the probability of finding a relationship between the two variables when the null hypothesis is true. This is typically expressed as a level of statistical significance between 0 and 1. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis. For testing a hypothesis, the commonly employed p-value is ≤ 0.05.

Polygenic score: A numerical score that is intended to be a measure of genetic risk for a given phenotype. Polygenic scores are said to predict a certain amount of trait variance in a given population.

Polymorphism: A form of a genetic variant that occurs with a certain frequency in a given population (typically defined as greater than 1% in a given population).

Population stratification: The phenomenon is which differences in allele frequencies between cases and controls is actually due to (or is confounded by) ethnic/ancestral differences in allele frequencies.

R-squared: A statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

Single nucleotide polymorphism (SNP): A substitution of a single nucleotide at a specific position in the genome with a different nucleotide, that occurs in a sufficiently large fraction of the population.

Somatic mosaicism: Having two or more genetically distinct populations of cells within the same individual.

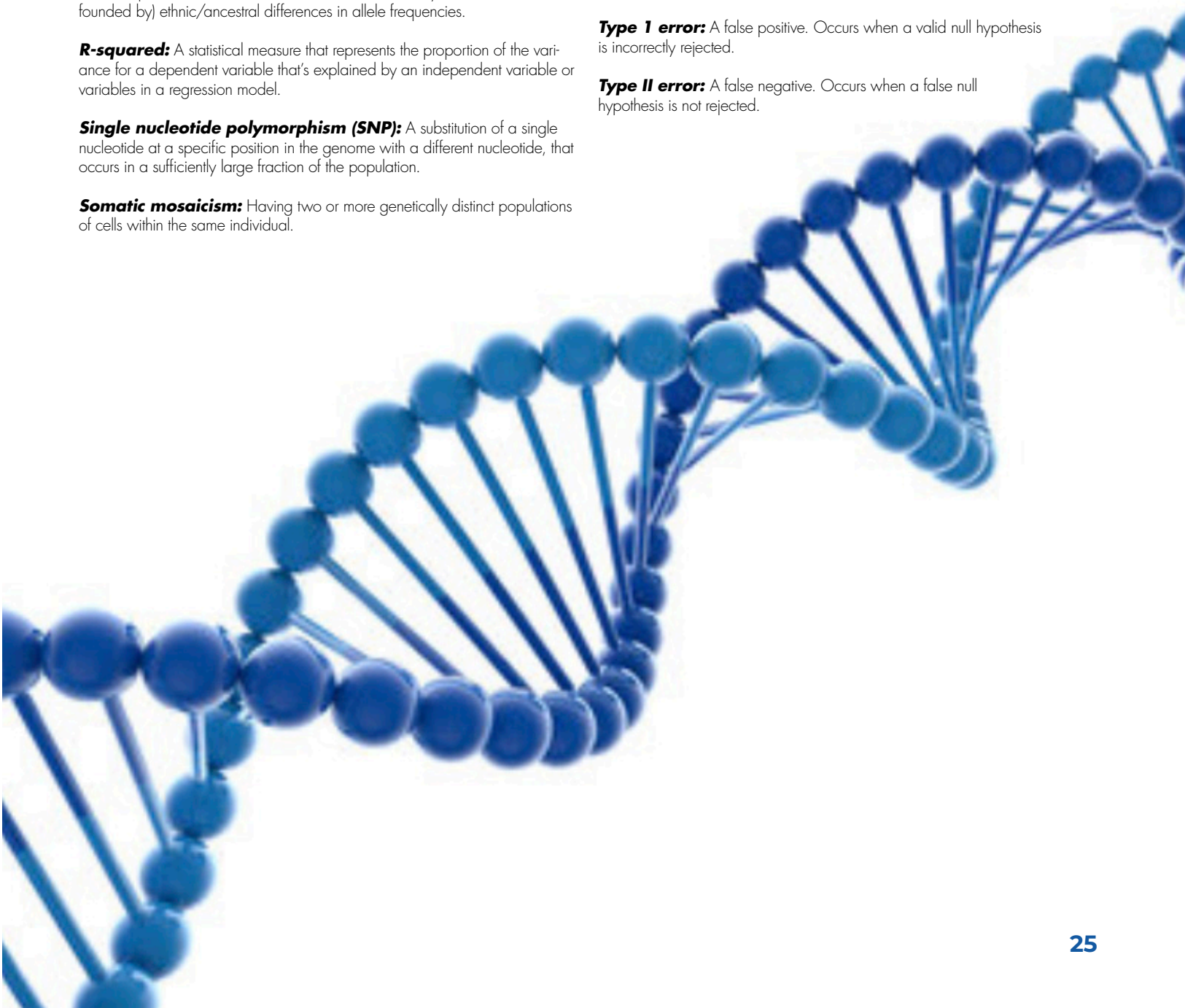
Structured population: A population that contains persons of different ancestral/ethnic backgrounds and different allele frequencies.

Training sample: A sample to which a polygenic score, derived from a discovery sample, is applied. In general, the training sample is used to tweak the polygenic score so as to achieve the highest R-squared possible.

Twin study: Used to derive heritability estimates on the basis of the presumed genetic similarity between monozygotic (MZ) v. dizygotic (DZ) twins.

Type 1 error: A false positive. Occurs when a valid null hypothesis is incorrectly rejected.

Type II error: A false negative. Occurs when a false null hypothesis is not rejected.



References

Alford, John R., Carolyn L. Funk, and John R. Hibbing. 2005. "Are Political Orientations Genetically Transmitted?" *American Political Science Review* 99 (2):153-67.

Altman, A., and M. Krzywinski. 2018. "The curse(s) of dimensionality." *Nat Methods* 15 (6):399-400.

Bauder, R. A., M. Herland, and T. M. Khoshgoftaar. 2019. "Evaluating Model Predictive Performance: A Medicare Fraud Detection Case Study." In *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*. Los Angeles, CA.

Beaver, K. M., and J. Belsky. 2012. "Gene-environment interaction and the intergenerational transmission of parenting: testing the differential-susceptibility hypothesis." *Psychiatr Q* 83 (1):29-40.

Beaver, K. M., M. Delisi, M. G. Vaughn, and J. P. Wright. 2010a. "Association between the A1 allele of the DRD2 gene and reduced verbal abilities in adolescence and early adulthood." *J Neural Transm (Vienna)* 117 (7):827-30.
2010b. "Association between the A1 allele of the DRD2 gene and reduced verbal abilities in adolescence and early adulthood." *J Neural Transm* 117 (7):827-30.

Beaver, K. M., J. P. Wright, M. DeLisi, A. Walsh, M. G. Vaughn, D. Boisvert, and J. Vaske. 2007a. "A gene x gene interaction between DRD2 and DRD4 is associated with conduct disorder and antisocial behavior in males." *Behav Brain Funct* 3:30.

Beaver, Kevin M., Christina Mancini, Matt DeLisi, and Michael G. Vaughn. 2010c. "Resiliency to Victimization: The Role of Genetic Factors." *Journal of Interpersonal Violence* 26 (5):874-98.

Beaver, Kevin M., John Paul Wright, Matt DeLisi, Leah E. Daigle, Marc L. Swatt, and Chris L. Gibson. 2007b. "Evidence of a Gene X Environment Interaction in the Creation of Victimization." *International Journal of Offender Therapy and Comparative Criminology* 51 (6):620-45.

Beaver, Kevin M., John Paul Wright, Matt DeLisi, Anthony Walsh, Michael G. Vaughn, Danielle Boisvert, and Jamie Vaske. 2007c. "A gene x gene interaction between DRD2 and DRD4 is associated with conduct disorder and antisocial behavior in males." *Behavioral and brain functions : BBF* 3:30-.

Berg, J. J., and G. Coop. 2014. "A population genetic signal of polygenic adaptation." *PLoS Genet* 10 (8):e1004412.

Berg, J. J., A. Harpak, N. Sinnott-Armstrong, A. M. Joergensen, H. Mostafavi, Y. Field, . . . G. Coop. 2019a. "Reduced signal for polygenic adaptation of height in UK Biobank." *Elife* 8.

Berg, Jeremy J., Xinjun Zhang, and Graham Coop. 2019b. "Polygenic Adaptation has Impacted Multiple Anthropometric Traits." *bioRxiv*:167551.

Bjelland, D. W., K. A. Weigel, N. Vukasinovic, and J. D. Nkrumah. 2013. "Evaluation of inbreeding depression in Holstein cattle using whole-genome SNP markers and alternative measures of genomic inbreeding." *Journal of Dairy Science* 96 (7):4697-706.

Border, R., E. C. Johnson, L. M. Evans, A. Smolen, N. Berley, P. F. Sullivan, and M. C. Keller. 2019. "No Support for Historical Candidate Gene or Candidate Gene-by-Interaction Hypotheses for Major Depression Across Multiple Large Samples." *Am J Psychiatry* 176 (5):376-87.

Border, R., and M. C. Keller. 2017. "Commentary: Fundamental problems with candidate gene-by-environment interaction studies - reflections on Moore and Thoemmes (2016)." *J Child Psychol Psychiatry* 58 (3):328-30.

Bradshaw, Matt, and Christopher G. Ellison. 2008. "Do Genetic Factors Influence Religious Life? Findings from a Behavior Genetic Analysis of Twin Siblings." *Journal for the Scientific Study of Religion* 47 (4):529-44.

Brooks-Crozier, Jennifer. 2011. "The Nature and Nurture of Violence: Early Intervention Services for the Families of MAOA-Low Children as a Means to Reduce Violent Crime and the Costs of Violent Crime " *Connecticut Law Review* 140:531-73.

Calus, M. P. 2010. "Genomic breeding value prediction: methods and procedures." *Animal* 4 (2):157-64.

Cardon, L. R., and L. J. Palmer. 2003. "Population stratification and spurious allelic association." *Lancet* 361 (9357):598-604.

Caspi, Avshalom, Joseph McClay, Terrie E. Moffitt, Jonathan Mill, Judy Martin, Ian W. Craig, . . . Richie Poulton. 2002. "Role of Genotype in the Cycle of Violence in Maltreated Children." *Science* 297 (5582):851-4.

Chabris, C. F., B. M. Hebert, D. J. Benjamin, J. Beauchamp, D. Cesarini, M. van der Loos, . . . D. Laibson. 2012. "Most reported genetic associations with general intelligence are probably false positives." *Psychol Sci* 23 (11):1314-23.

Chabris, C. F., J. J. Lee, D. Cesarini, D. J. Benjamin, and D. I. Laibson. 2015. "The Fourth Law of Behavior Genetics." *Curr Dir Psychol Sci* 24 (4):304-12.

Charmantier, A., D. Garant, and D. G. Kruuk, eds. 2014. *Quantitative genetics in the wild*. Oxford Oxford University Press

Charney, E. 2012. "Behavior genetics and postgenomics." *Behav Brain Sci* 35 (5):331-58.

Cheruiyot, E. K., T. T. T. Nguyen, M. Haile-Mariam, B. G. Cocks, M. Abdelsayed, and J. E. Pryce. 2020. "Genotype-by-environment (temperature-humidity) interaction of milk production traits in Australian Holstein cattle." *J Dairy Sci* 103 (3):2460-76.

Choi, S. W., T. S. Mak, and P. F. O'Reilly. 2020. "Tutorial: a guide to performing polygenic risk score analyses." *Nat Protoc* 15 (9):2759-72.

Coleman, J. R. I., J. Bryois, H. A. Gaspar, P. R. Jansen, J. E. Savage, N. Skene, . . . G. Breen. 2019. "Biological annotation of genetic loci associated with intelligence in a meta-analysis of 87,740 individuals." *Mol Psychiatry* 24 (2):182-97.

Cook, J. P., A. Mahajan, and A. P. Morris. 2020. "Fine-scale population structure in the UK Biobank: implications for genome-wide association studies." *Hum Mol Genet*.

Cox, S. L., C. B. Ruff, R. M. Maier, and I. Mathieson. 2019. "Genetic contributions to variation in human stature in prehistoric Europe." *Proc Natl Acad Sci U S A* 116 (43):21484-92.

D'Onofrio, B. M., L. J. Eaves, L. Murrelle, H. H. Maes, and B. Spilka. 1999. "Understanding biological and social influences on religious affiliation, attitudes, and behaviors: a behavior genetic perspective." *J Pers* 67 (6):953-84.

Davies, G., M. Lam, S. E. Harris, J. W. Trampush, M. Luciano, W. D. Hill, . . . I. J. Deary. 2018. "Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function." *Nat Commun* 9 (1):2098.

Davis, O. S., L. M. Butcher, S. J. Docherty, E. L. Meaburn, C. J. Curtis, M. A. Simpson, . . . R. Plomin. 2010. "A three-stage genome-wide association study of general cognitive ability: hunting the small effects." *Behav Genet* 40 (6):759-67.

Daw, J., and G. Guo. 2011. "The influence of three genes on whether adolescents use contraception, USA 1994-2002." *Popul Stud (Camb)* 65 (3):253-71.

Dawes, Christopher T., and James H. Fowler. 2009. "Partisanship, Voting, and the Dopamine D2 Receptor Gene." *The Journal of Politics* 71 (3):1157-71.
DeLisi, Matt, Kevin M. Beaver, Michael G. Vaughn, and John Paul Wright. 2009. "All in the Family: Gene × Environment Interaction Between DRD2 and Criminal Father Is Associated With Five Antisocial Phenotypes." *Criminal Justice and Behavior* 36 (11):1187-97.
Duncan, L. E., M. Ostacher, and J. Ballon. 2019. "How genome-wide association studies (GWAS) made traditional candidate gene studies obsolete." *Neuropsychopharmacology* 44 (9):1518-23.
Eaves, L, N. G. Martin, and A. C. Heath. 1990. "Religious Affiliation in Twins and Their Parents: Testing a Model of Cultural Inheritance." *Behavior Genetics* 20 (1):1-22.
Farrell, M. S., T. Werge, P. Sklar, M. J. Owen, R. A. Ophoff, M. C. O'Donovan, . . . P. F. Sullivan. 2015. "Evaluating historical candidate genes for schizophrenia." *Mol Psychiatry* 20 (5):555-62.
Fisher, Ronald Aylmer Sir. 1990 [1918]. *The genetical theory of natural selection*. Oxford: The Clarendon Press.
Flint, J., R. J. Greenspan, and K. S. Kendler. 2020. *How Genes Influence Behavior*. Oxford: Oxford Univeristy Press.
Fowler, J. H., J. E. Settle, and N. A. Christakis. 2011. "Correlated genotypes in friendship networks." *Proceedings of the National Academy of Sciences of the United States of America* 108 (5):1993-7.
Fowler, James H., and Christopher T. Dawes. 2008. "Two Genes Predict Voter Turnout." *The Journal of Politics* 70 (3):579-94.
Guo, G., and K. H. Tillman. 2009a. "Trajectories of depressive symptoms, dopamine D2 and D4 receptors, family socioeconomic status and social support in adolescence and young adulthood." *Psychiatr Genet* 19 (1):14-26.
Guo, G., and Y. Tong. 2006. "Age at first sexual intercourse, genes, and social context: evidence from twins and the dopamine D4 receptor gene." *Demography* 43 (4):747-69.
Guo, Guang, Michael Roettger, and Jean Shih. 2007a. "Contributions of the DAT1 and DRD2 genes to serious and violent delinquency among adolescents and young adults." *Human Genetics* 121 (1):125-36.
Guo, Guang, and Kathryn Harker Tillman. 2009b. "Trajectories of depressive symptoms, dopamine D2 and D4 receptors, family socioeconomic status and social support in adolescence and young adulthood." *PSYCHIATRIC GENETICS* 19 (1).
Guo, Guang, Kirk Wilhelmsen, and Nathan Hamilton. 2007b. "Gene-lifecourse interaction for alcohol consumption in adolescence and young adulthood: five monoamine genes." *American journal of medical genetics. Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics* 144B (4):417-23.
Guo, J., Y. Wu, Z. Zhu, Z. Zheng, M. Trzaskowski, J. Zeng, . . . J. Yang. 2018. "Global genetic differentiation of complex traits shaped by natural selection in humans." *Nat Commun* 9 (1):1865.
Haberstick, Brett C., Jeffrey M. Lessem, Christian J. Hopfer, Andrew Smolen, Marissa A. Ehringer, David Timberlake, and John K. Hewitt. 2005. "Monoamine oxidase A (MAOA) and antisocial behaviors in the presence of childhood and adolescent maltreatment." *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 135B (1):59-64.
Halpern, C. T., C. E. Kaestle, G. Guo, and D. D. Hallfors. 2007. "Gene-environment contributions to young adult sexual partnering." *Arch Sex Behav* 36 (4):543-54.
Harden, K. P. 2020. ""Reports of My Death Were Greatly Exaggerated": Behavior Genetics in the Postgenomic Era." *Annu Rev Psychol*.
Hastie, T., and R. Tibshirani. 2003. "Expression Arrays and the p *a* n Problem." *Techincal Report, Stanford University*.
Haworth, S., R. Mitchell, L. Corbin, K. H. Wade, T. Dudding, A. Budu-Aggrey, . . . J. Timpson N. 2019. "Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis." *Nat Commun* 10 (1):333.
Hewitt, J. K. 2012. "Editorial policy on candidate gene association and candidate gene-by-environment interaction studies of complex traits." *Behav Genet* 42 (1):1-2.
Hill, W. D., R. E. Marioni, O. Maghzian, S. J. Ritchie, S. P. Hagenaars, A. M. McIntosh, . . . I. J. Deary. 2019. "A combined analysis of genetically correlated traits identifies 187 loci and a role for neurogenesis and myelination in intelligence." *Mol Psychiatry* 24 (2):169-81.
Hu, L., X. Yao, H. Huang, Z. Guo, X. Cheng, Y. Xu, . . . D. Li. 2018. "Clinical significance of germline copy number variation in susceptibility of human diseases." *J Genet Genomics* 45 (1):3-12.
Huizinga, David, Brett C. Haberstick, Andrew Smolen, Scott Menard, Susan E. Young, Robin P. Corley, . . . John K. Hewitt. 2006. "Childhood Maltreatment, Subsequent Antisocial Behavior, and the Role of Monoamine Oxidase A Genotype." *Biological Psychiatry* 60 (7):677-83.
Inouye, M., G. Abraham, C. P. Nelson, A. M. Wood, M. J. Sweeting, F. Dudbridge, . . . N. J. Samani. 2018. "Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention." *J Am Coll Cardiol* 72 (16):1883-93.
Jarvis, J. P., A. P. Peter, and J. A. Shaman. 2019. "Consequences of CYP2D6 Copy-Number Variation for Pharmacogenomics in Psychiatry." *Front Psychiatry* 10:432.
Jenko Bizjan, B., T. Katsila, T. Tesovnik, R. Sket, M. Debeljak, M. T. Matsoukas, and J. Kovac. 2020. "Challenges in identifying large germline structural variants for clinical use by long read sequencing." *Comput Struct Biotechnol J* 18:83-92.
Johnston, Charlotte, Benjamin B. Lahey, and Walter Matthys. 2013. "Editorial Policy for Candidate Gene Studies." *Journal of Abnormal Child Psychology* 41 (4):511-4.
Joseph, Jay. 2014. *The Trouble with Twin Studies: A Reassessment of Twin Research in the Social and Behavioral Sciences*. Abingdon, UK: Routledge.
Kamin, L. J., and A. S. Goldberger. 2002. "Twin studies in behavioral research: a skeptical view." *Theor Popul Biol* 61 (1):83-95.
Kerr, Norbert L. 1998. "HARKing: Hypothesizing After the Results are Known." *Personality and Social Psychology Review* 2 (3):196-217.
Khera, A. V., M. Chaffin, K. G. Aragam, M. E. Haas, C. Roselli, S. H. Choi, . . . S. Kathiresan. 2018. "Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations." *Nat Genet* 50 (9):1219-24.
Kim-Cohen, J., A. Caspi, A. Taylor, B. Williams, R. Newcombe, I. W. Craig, and T. E. Moffitt. 2006. "MAOA, maltreatment, and gene-environment interaction predicting children's mental health: new evidence and a meta-analysis." *Mol Psychiatry* 11 (10):903-13.

Koch, E., M. Ristroph, and M. Kirkpatrick. 2013. "Long range linkage disequilibrium across the human genome." *PLoS ONE* 8 (12):e80754.

Lam, M., W. D. Hill, J. W. Trampush, J. Yu, E. Knowles, G. Davies, . . . T. Lencz. 2019. "Pleiotropic Meta-Analysis of Cognition, Education, and Schizophrenia Differentiates Roles of Early Neurodevelopmental and Adult Synaptic Pathways." *Am J Hum Genet* 105 (2):334-50.

Lappalainen, T., S. Koivumäki, E. Salmela, K. Huoponen, P. Sistonen, M. L. Savontaus, and P. Lahermo. 2006. "Regional differences among the Finns: a Y-chromosomal perspective." *Gene* 376 (2):207-15.

Lazaridis, I., N. Patterson, A. Mittnik, G. Renaud, S. Mallick, K. Kirsanow, . . . J. Krause. 2014. "Ancient human genomes suggest three ancestral populations for present-day Europeans." *Nature* 513 (7518):409-13.

Lee, J. J., R. Wedow, A. Okbay, E. Kong, O. Maghzian, M. Zacher, . . . D. Cesarini. 2018a. "Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals." *Nat Genet* 50 (8):1112-21. 2018b. "Supplementary Note for Gene discovery and polygenic prediction from a 1.1-million-person GWAS of educational attainment." *Nat Genet* 50 (8):1-205.

Leslie, S., B. Winney, G. Hellenthal, D. Davison, A. Boumertit, T. Day, . . . W. Bodmer. 2015. "The fine-scale genetic structure of the British population." *Nature* 519 (7543):309-14.

Liu, J., Y. Zhou, S. Liu, X. Song, X. Z. Yang, Y. Fan, . . . N. Wu. 2018. "The coexistence of copy number variations (CNVs) and single nucleotide polymorphisms (SNPs) at a locus can result in distorted calculations of the significance in associating SNPs to disease." *Hum Genet* 137 (6-7):553-67.

Locke, A. E., K. M. Steinberg, C. W. K. Chiang, S. K. Service, A. S. Havulinna, L. Stell, . . . N. B. Freimer. 2019. "Exome sequencing of Finnish isolates enhances rare-variant association power." *Nature* 572 (7769):323-8.

Loehlin, John C., and Robert C. Nichols. 1976. *Heredity, environment, & personality: a study of 850 sets of twins.* Austin: University of Texas Press, c1976.

Maier, R. M., P. M. Visscher, M. R. Robinson, and N. R. Wray. 2018. "Embracing polygenicity: a review of methods and tools for psychiatric genetics research." *Psychol Med* 48 (7):1055-67.

McClernon, F. J., B. F. Fuemmeler, S. H. Kollins, M. E. Kail, and A. E. Ashley-Koch. 2008. "Interactions between genotype and retrospective ADHD symptoms predict lifetime smoking risk in a sample of young adults." *Nicotine Tob Res* 10 (1):117-27.

McEvoy, B. P., and P. M. Visscher. 2009. "Genetics of human height." *Econ Hum Biol* 7 (3):294-306.

McSwiggan, S., B. Elger, and P. S. Appelbaum. 2017. "The forensic use of behavioral genetics in criminal proceedings: Case of the MAOA-L genotype." *Int J Law Psychiatry* 50:17-23.

Mostafavi, H., A. Harpak, I. Agarwal, D. Conley, J. K. Pritchard, and M. Przeworski. 2020. "Variable prediction accuracy of polygenic scores within an ancestry group." *Elife* 9.

Mulim, Henrique Alberto, Luis Fernando Batista Pinto, Aline Zampar, Gerson Barreto Mourão, Altair Antônio Valloto, and Victor Breno Pedrosa. 2020. "Assessment of Genotype by Environment Interaction Via Reaction Norms for Milk Yield in Holstein Cattle of Southern Brazil." *Annals of Animal Science* 20 (3):1101-12.

Muller-Spahn, F. 2008. "Individualized preventive psychiatry: syndrome and vulnerability diagnostics." *Eur Arch Psychiatry Clin Neurosci* 258 Suppl 5:92-7.

Naumova, Oksana Yu, Maria Lee, Sergei Yu Rychkov, Natalia V. Vlasova, and Elena L. Grigorenko. 2013. "Gene expression in the human brain: the current state of the study of specificity and spatiotemporal dynamics." *Child Development* 84 (1):76-88.

Nelson, D., J. Kelleher, A. P. Ragsdale, C. Moreau, G. McVean, and S. Gravel. 2020. "Accounting for long-range correlations in genome-wide simulations of large cohorts." *PLoS Genet* 16 (5):e1008619.

Okbay, A., J. P. Beauchamp, M. A. Fontana, J. J. Lee, T. H. Pers, C. A. Rietveld, . . . D. J. Benjamin. 2016. "Genome-wide association study identifies 74 loci associated with educational attainment." *Nature* 533 (7604):539-42.

Olson, James M., Philip A. Vernon, Julie Aitken Harris, and Kerry L. Jang. 2001. "The heritability of attitudes: A study of twins." *Journal of Personality and Social Psychology* 80 (6):845-60.

Pam, A., S. S. Kemker, C. A. Ross, and R. Golden. 1996. "The "equal environments assumption" in MZ-DZ twin comparisons: an untenable premise of psychiatric genetics?" *Acta Genet Med Gemellol (Roma)* 45 (3):349-60.

Park, L. 2019. "Population-specific long-range linkage disequilibrium in the human genome and its influence on identifying common disease variants." *Sci Rep* 9 (1):11380.

Phillips, C., J. Amigo, A. Carracedo, and M. V. Lareu. 2015. "Tetra-allelic SNPs: Informative forensic markers compiled from public whole-genome sequence data." *Forensic Sci Int Genet* 19:100-6.

Phillips, C., J. Amigo, A. O. Tillmar, M. A. Peck, M. de la Puente, J. Ruiz-Ramírez, . . . M. V. Lareu. 2020. "A compilation of tri-allelic SNPs from 1000 Genomes and use of the most polymorphic loci for a large-scale human identification panel." *Forensic Science International: Genetics* 46.

Plomin, R., and S. von Stumm. 2018. "The new genetics of intelligence." *Nat Rev Genet* 19 (3):148-59.

Popejoy, A. B., and S. M. Fullerton. 2016. "Genomics is failing on diversity." *Nature* 538 (7624):161-4.

Price, Alkes L., Michael E. Weale, Nick Patterson, Simon R. Myers, Anna C. Need, Kevin V. Shianna, . . . David Reich. 2008. "Long-range LD can confound genome scans in admixed populations." *American Journal of Human Genetics* 83 (1):132-9.

Prichard, Z., A. Mackinnon, A. F. Jorm, and S. Easteal. 2008. "No evidence for interaction between MAOA and childhood adversity for antisocial behavior." *Am J Med Genet B Neuropsychiatr Genet* 147B (2):228-32.

Racimo, F., J. J. Berg, and J. K. Pickrell. 2018. "Detecting Polygenic Adaptation in Admixture Graphs." *Genetics* 208 (4):1565-84.

Richardson, K., and S. Norgate. 2005. "The equal environments assumption of classical twin studies may not hold." *Br J Educ Psychol.* 75 (Pt 3):339-50.

Rietveld, C. A., S. E. Medland, J. Derringer, J. Yang, T. Esko, N. W. Martin, . . . P. D. Koellinger. 2013. "GWAS of 126,559 individuals identifies genetic variants associated with educational attainment." *Science* 340 (6139):1467-71.

Robinson, M. R., G. Hemani, C. Medina-Gomez, M. Mezzavilla, T. Esko, K. Shakhbazov, . . . P. M. Visscher. 2015. "Population genetic differentiation of height and body mass index across Europe." *Nat Genet* 47 (11):1357-62.

Sabol, S. Z., S. Hu , and D. Hamer. 1998. "A functional polymorphism in the monoamine oxidase A gene promoter." *Human Genetics* 103:273–9.

Sanna, S., A. U. Jackson, R. Nagaraja, C. J. Willer, W. M. Chen, L. L. Bonnycastle, . . . K. L. Mohlke. 2008. "Common variants in the GDF5-UQCC region are associated with variation in human height." *Nat Genet* 40 (2):198-203.

Santos, J. C., C. H. M. Malhado, J. A. Cobuci, M. P. G. de Rezende, and P. L. S. Carneiro. 2020. "Genotype-environment interaction for productive traits of Holstein cows in Brazil described by reaction norms." *Trop Anim Health Prod* 52 (5):2425-32.

Savage, J. E., P. R. Jansen, S. Stringer, K. Watanabe, J. Bryois, C. A. de Leeuw, . . . D. Posthuma. 2018. "Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence." *Nat Genet* 50 (7):912-9.

Savulescu, Julian. 2014. "The Nature of the Moral Obligation To Choose the Best Children." In *The Future of Bioethics,* ed. A. Akabayashi. Oxford, UK: Oxford Univeristy Press.

Schönemann, P. H. 1997. "On models and muddles of heritability." *Genetica* 99 (2-3):97-108.

Shanahan, M. J., S. Vaisey, L. D. Erickson, and A. Smolen. 2008. "Environmental contingencies and genetic propensities: social capital, educational continuation, and dopamine receptor gene DRD2." *Ajs* 114 Suppl:S260-86.

Shanahan, Michael J., Lance D. Erickson, Stephen Vaisey, and Andrew Smolen. 2007. "Helping Relationships and Genetic Propensities: A Combinatoric Study of DRD2, Mentoring, and Educational Continuation." *TWIN RESEARCH AND HUMAN GENETICS* 10 (2):285-98.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22 (11):1359-66.

Sniekers, S., S. Stringer, K. Watanabe, P. R. Jansen, J. R. I. Coleman, E. Krapohl, . . . D. Posthuma. 2017. "Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence." *Nat Genet* 49 (7): 1107-12.

Sohail, M., R. M. Maier, A. Ganna, A. Bloemendal, A. R. Martin, M. C. Turchin, . . . S. R. Sunyaev. 2019. "Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies." *Elife* 8.

Souza, Laaina de Andrade, Carlos Henrique Mendes Malhado, José Braccini Neto, Raimundo Martins Filho, and Paulo Luiz Souza Carneiro. 2016. "Genotype-Environment Interactions on the Weight of Tabapua Cattle In the Northeast Of Brazil." *Revista Caatinga* 29:206-15.

Sylvestre, John, Geranda Notten, Nick Kerman, Alexia Polillo, and Konrad Czechowki. 2018. "Poverty and Serious Mental Illness: Toward Action on a Seemingly Intractable Problem." *American Journal of Community Psychology* 61 (1-2):153-65.

Thomas, D. C., and J. S. Witte. 2002. "Point: Population stratification: A problem for case-control studies of candidate-gene associations?" *Cancer Epidemiology Biomarkers & Prevention* 11 (6):505-12.

Torres, Fátima, Fátima Lopes, and Patrícia Maciel. 2020. "Relevance of Copy Number Variation to Human Genetic Disease." *eLS, John Wiley & Sons, Ltd (Ed.).*

Turchin, M. C., C. W. Chiang, C. D. Palmer, S. Sankararaman, D. Reich, ANthropometric Traits Consortium Genetic Investigation of, and J. N. Hirschhorn. 2012. "Evidence of widespread selection on standing variation in Europe at height-associated SNPs." *Nat Genet* 44 (9):1015-9.

Turkheimer, E. 2000. "Three Laws of Behavior Genetics and What They Mean." *Current Directions in Psychological Science* 9 (5):160-64.

Turkheimer, Eric. 2018. "P-Hacking in Gwas." In *GHA Project: Turkheimer's Projects: Genetics and Human Agency,* ed. E. Turkheimer. Charlottesville, VA

Vaske, J., M. Makarios, D. Boisvert, K. M. Beaver, and J. P. Wright. 2009. "The interaction of DRD2 and violent victimization on depression: an analysis by gender and race." *J Affect Disord* 112 (1-3):120-5.

Vaughn, M. G., K. M. Beaver, M. DeLisi, B. E. Perron, and L. Schelbe. 2009. "Gene-environment interplay and the importance of self-control in predicting polydrug use and substance-related problems." *Addict Behav* 34 (1):112-6.

Visscher, P. M., and M. E. Goddard. 2019. "From R.A. Fisher's 1918 Paper to GWAS a Century Later." *Genetics* 211 (4):1125-30.

Wald, N. J., and R. Old. 2019. "The illusion of polygenic disease risk prediction." *Genet Med* 21 (8):1705-7.

Weedon, MN, G Lettre, RM Freathy, CM Lindgren, BF Voight, JR Perry, . . . TM Frayling. 2007. "A common variant of HMGA2 is associated with adult and childhood height in the general population." *Nat Genet* 39:1245 - 50.

Widom, C. S., and L. M. Brzustowicz. 2006. "MAOA and the "cycle of violence:" childhood abuse and neglect, MAOA genotype, and risk for violent and antisocial behavior." *Biol Psychiatry* 60 (7):684-9.

Wood, L. M. W. 2020. Twin Studies and the Equal Environments Assumption – An Evaluation of the Genetic Heritability Account of Behaviour, The University of Guelph, Ontario, Canada.

Wray, N. R., K. E. Kemper, B. J. Hayes, M. E. Goddard, and P. M. Visscher. 2019. "Complex Trait Prediction from Genome Data: Contrasting EBV in Livestock to PRS in Humans: Genomic Prediction." *Genetics* 211 (4):1131-41.

Wray, N. R., J. Yang, B. J. Hayes, A. L. Price, M. E. Goddard, and P. M. Visscher. 2013. "Pitfalls of predicting complex traits from SNPs." *Nat Rev Genet* 14 (7):507-15.

Yong, E. 2019. "A Waste of 1,000 Research Papers." *The Atlantic.*

Yue, X. P., C. Dechow, and W. S. Liu. 2015. "A limited number of Y chromosome lineages is present in North American Holsteins." *J Dairy Sci* 98 (4):2738-45.

Zaidi, Arslan A., and Iain Mathieson. 2020. "Demographic history impacts stratification in polygenic scores." *bioRxiv*:2020.07.20.212530.

Zoledziewska, M., C. Sidore, C. W. K. Chiang, S. Sanna, A. Mulas, M. Steri, . . . F. Cucca. 2015. "Height-reducing variants and selection for short stature in Sardinia." *Nat Genet* 47 (11): 1352-6.